Chapter 6
Facial Expression Synthesis Based on Emotion Dimensions for Affective Talking Avatar

Shen Zhang[1], Zhiyong Wu[1,2], Helen M. Meng[3], and Lianhong Cai[1]

[1] Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
[2] Tsinghua-CUHK Joint Research Center for Media Sciences,
Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen
[3] Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, HKSAR, China
zhangshen05@mails.tsinghua.edu.cn, john.zy.wu@gmail.com,
hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

Abstract. Facial expression is one of the most expressive ways for human beings to deliver their emotion, intention, and other nonverbal messages in face to face communications. In this chapter, a layered parametric framework is proposed to synthesize the emotional facial expressions for an MPEG4 compliant talking avatar based on the three dimensional PAD model, including pleasure-displeasure, arousal-nonarousal and dominance-submissiveness. The PAD dimensions are used to capture the high-level emotional state of talking avatar with specific facial expression. A set of partial expression parameter (PEP) is designed to depict the expressive facial motion patterns in local face areas, and reduce the complexity of directly manipulation of low-level MPEG4 facial animation parameters (FAP). The relationship among the emotion (PAD), expression (PEP) and animation (FAP) parameter is analyzed on a virtual facial expression database. Two levels of parameter mapping are implemented, namely the emotion-expression mapping from PAD to PEP, and the linear interpolation from PEP to FAP. The synthetic emotional facial expression is combined with the talking avatar speech animation in a text to audio visual speech system. Perceptual evaluation shows that our approach can generate appropriate facial expressions for subtle and complex emotions defined by PAD and thus enhance the emotional expressivity of talking avatar.

Keywords: facial expression, emotion, talking avatar, PAD.

1 Introduction

In face-to-face interactions, people expressed themselves through a number of different modalities, such as speech, facial expression, body gestures etc. Speaking is not only moving the lips, but also includes the eye-contact, eyebrow-raise, and

T. Nishida et al. (Eds.): Modeling Machine Emotions for Realizing Intelligence, SIST 1, pp. 109–132.
springerlink.com

other facial movements. Facial expression is one of the most effective ways for human to express their intentions, attitude, inner emotional states and other nonverbal messages in speech communications [13]. Early efforts in psychological research have established that the affective information in human communications is delivered at different ratio by different modalities: the verbal expression (e.g., words, spoken text) only accounts for 7% of the affective meaning of speakers' feeling and attitude. The vocal expression (e.g., prosody, stress) conveys 38% of the affective message. By marked contrast, the facial expression accounts for 55% of the affective information [32]. Especially when inconsistent messages are expressed by different modality, the facial expression tends to be more reliable for human perception of emotion [35]. Facial expression is also considered as the one of the most important aspect in analyzing and modeling natural human emotions [40]. The study of facial expression has attracted broad attentions from psychologists [39, 45], biologists [10], computer engineers [53], and artists [15]. The reason why facial expression receives so many interests is not only because it plays important roles in human communications, but also its complexity and difficulty to analyze and synthesize.

Talking avatar, an animated speaking virtual character with vivid human-like appearance and natural synthetic speech, has gradually shown their potential application in human computer intelligent interactions, because of its communication abilities to deliver rich verbal and nonverbal information by voice, tones, eye-contact, body gestures and facial expressions, etc. Possible applications of talking avatar include a virtual storyteller for children [21, 51], a virtual guider or presenter for personal or commercial website [36, 56], a representative of user in computer game [29] and a funny puppetry for computer mediated human communications [42]. It is clearly promising that the talking avatar will become an expressive multimodal interface in human computer interaction. However, current research on human emotion has not yet advanced to the stage where it is possible to synthesize an affective intelligent talking avatar who can express their feelings and emotions through vocal and facial behaviors as natural as human beings. To this end, in our research on expressive talking avatar synthesis, it is necessary to generate appropriate facial expression according to the emotional content in speech in order to enhance the affective expressivity of talking avatar. Our long-term goal is to implement an expressive talking avatar with lip-articulation and emotional facial expressions for text-to-audio-visual-speech (TTAVS) system.

In this chapter, a parametric layered framework is proposed to synthesize emotional facial expression for an MPEG-4 compliant talking avatar based on psychological emotion dimensions. The following three research questions will be addressed in our work:

1) How to describe various emotional states in a quantitative and unified way?
2) How to model the facial expression based on animation control parameters?
3) What is the interrelation between emotional state and facial expression?

Quick answers to these questions are that we adopt the PAD psychological model to describe arbitrary emotional state by three dimensions, namely pleasure-displeasure (P), arousal-nonarousal (A) and dominance-submissiveness (D) [33].

A set of Partial Expression Parameters (PEP) is proposed to depict the typical facial expressive movement in face regions based on the correlation among MPEG-4 facial animation parameters (FAP). The PAD and PEP mapping function and the PEP-FAP translation function are trained on a relatively small facial expression database. A parametric layered framework is thus established with the PAD as high-level emotion descriptor, the PEP as mid-level facial expression configuration and the FAP as low-level controller of face model deformation on three-dimensional avatar.

The chapter is organized as follows. Section 2 presents the literature review on the emotion description and facial expression synthesis, and introduces the PAD emotional model. The layered framework for parametric facial expression synthesis based on emotional dimensions (i.e. pleasure-arousal-dominance) is described in Section 3. Details of PEP definition and its translation to FAP are illustrated in Section 4. In Section 5, the establishment of pseudo facial expression database on talking avatar is reported. The interrelation between emotional state description (PAD) and facial expression configuration (PEP) of talking avatar is investigated in Section 6. The integration of the proposed facial expression module into text-to-audio-visual-speech system is described in Section 7. The resulting emotional facial expression synthesized by the proposed PAD-PEP-FAP framework and a series of perceptual evaluation are illustrated in Section 8. Finally, we conclude our work and discuss the future direction.

2 Related Work

Facial expression and its relationship with emotion are topics having been actively studied. Early academic study on facial expression may date back to the contribution made by Darwin [10]. Since then emotion and facial expression are studied mainly in psychological area. The development of facial action coding system (FACS) has accelerated the research of automatic facial expression analysis in the filed of computer vision since 1978 [27]. After Picard proposed the theory of affective computing in 1997 [40], facial expression is regarded as one of the most important visual channels to perceive and express human emotion. It is necessary for affective computer to understand and perform appropriate facial expression in human computer communication. From the aspect of talking avatar synthesis, in this section, we review the previous work on emotion description by psychologists, and facial expression synthesis by computer scientists. In addition, for parametric facial animation, we introduce two important coding systems, namely the FACS proposed by Ekman [14] and facial animation framework in MPEG-4 standards [34].

2.1 Categorical Emotion Description

How to describe human emotion and facial expression appropriately and effectively? This question has been discussed by scientists (mainly psychologists) for a long time. Various existing emotion descriptive models can be summarized into two aspects: the categorical and dimensional.

In his research on human expression of emotions, Darwin worked with a set of "state of mind", which includes emotion (e.g. anger, jealousy, terror, love), motivational, behavioral or personality traits (e.g. determination, defiance, ambition), sensations (e.g. bodily pain, hunger) and cognitive processes (e.g. abstraction, meditation). In addition, Darwin proposed that the expressions of "high spirits" and "low spirits" are best recognized [45]. The "state of mind" and "high/low spirits" can be regarded as a mixture of categorical and dimensional approach to describe emotion.

The "basic emotion" theory proposed by Ekman et al. has great influences on psychological research on facial expression. This theory is centered on a small set of discrete categories, which includes happiness, sadness, surprise, fear, anger, and disgust, namely the "big six" emotions. These basic emotions are considered to be universally perceived with respect to facial expression, regardless of culture [12]. The "big six" emotions is perhaps the most popular categorical approach to describe emotional state. The advantage of categorical approach lies in that it is the most straightforward and simplest way to describe facial expression in daily life. The existence of large amount of emotion-denoting words in natural languages makes the categorical approach powerful, intuitive and consistent with human's experiences [57]. However, the vagueness and ambiguity of natural language make it hard to clearly recognize the subtle variation of spontaneous occurring emotions. It is also difficult to describe the continuum or non-extreme emotional state by discrete list of categories. Although the basic emotion categories are widely adopted in facial expression analysis and recognition [27], when it comes to synthesis, the basic emotions only cover a relatively small set of prototypical facial expressions. For expressive talking avatar who aims to speak, move and behave as naturally as human, the expression of basic emotions or discrete emotion categories are not enough, while a unified and flexible approach is needed to model the complex and subtle facial expressions in human life.

2.2 Dimensional Emotion Description

An alternative approach is the dimensional description of emotion. For practical and theoretical purpose, early efforts attempted to describe emotion based on a hybrid of both basic-to-complex categories and emotional primitives. The Plutchik's wheel (2D) and "cone-shaped model" (3D) are proposed to describe emotions by exploring their relations in terms of similarity and intensity [43]. The 3D emotion cone's vertical dimension represents intensity, and the circle represents degrees of similarity among different emotions. There are eight primary emotions and other emotions occur as combinations, mixtures, or components of the primary emotions. The difference among emotions is measured by their degree of similarity to one another. Emotions can vary in degrees of intensity or levels of arousal. The Plutchik's theory tried to uncover the relations among different emotions by dimensional analysis, but is still limited in categorical concept and thus not a completely dimensional model.

For dimensional emotion description, ideally, arbitrary emotional state can be quantitatively measured in terms of a small number of basic dimensions in a

multi-dimensional space. Many researchers try to propose different emotion dimensions to describe the essential property of various emotional state, and three basic dimensions, with different names, can be summarized from previous research [7, 37, 44]. In this study, we choose the following terms pleasure (i.e. valence, evaluation), arousal (i.e. activation, activity), dominance (i.e. power, potency). The pleasure dimension distinguishes the positive-negative quality of emotional state, the arousal dimension measures the degree of physical activity and mental alertness of emotional state, and the dominance dimension is defined in terms of control versus lack of control, mainly focus on the social aspect of emotion. Based on these basic dimensions, different emotional models are proposed and explored in multi-disciplinary area. Cowie et al. proposed 2D disk-shaped activation-evaluation emotion space [8]. Mehrabin et.al developed the PAD emotional model by which all emotions can be located in 3D emotion space where the pleasure, arousal and dominance dimensions are orthogonal and independent to each other, as shown in Figure 1 [28, 33].

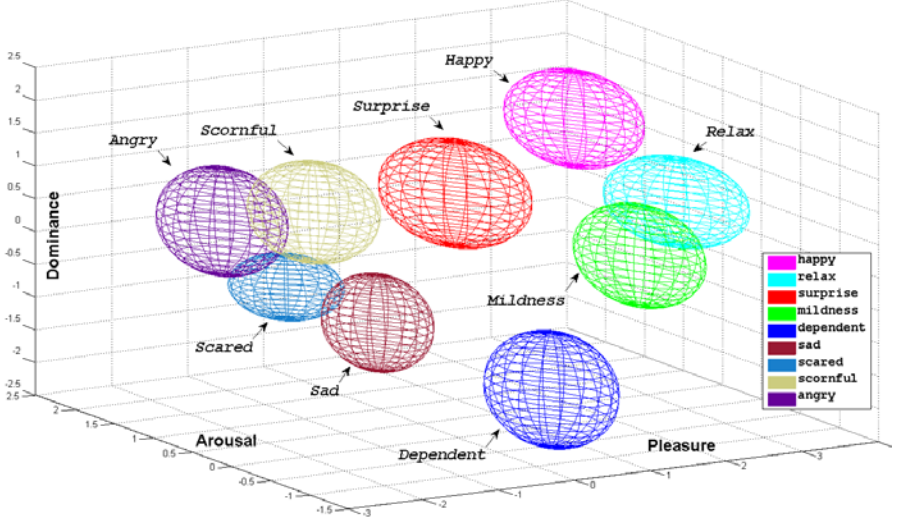


Fig. 1. The distribution of 9 emotional states/words in PAD emotion space. The center of each ellipsoid is the mean and the radius is standard deviation of each emotional state. The PAD data is provided by Institute of Psychology, Chinese Academy of Sciences [28].

The emotion dimensions can measure the continuous variation of emotional states, but they are not so intuitive as emotion labels in natural language. To this end, different rating methods and tools are developed to annotate the value of emotional dimension, such as the self assessment manikins (SAM) tool [25], the Feeltrace system [8], and the abbreviated PAD emotion scales [28]. The use of dimensional emotion description has been explored in multimodal affect recognition and synthesis [3, 61], emotional speech synthesis [48], facial expression recognition [7]. However, research on facial expression synthesis

based on emotion dimensions is rare. In this chapter, we explore the use of PAD emotional model to synthesize continuum of varied facial expression for talking avatar. According to Mehrabian's theory [33], human emotion is not limited to isolated categories but can be described along three nearly orthogonal dimensions: pleasure-displeasure (P), arousal-nonarousal (A) and dominance-submissiveness (D), namely the PAD emotional model which is proven to be appropriate in describing emotions. The PAD emotional model has been successfully applied in acoustic correlates analysis of emotional speech [9] and expressive speech synthesis [59, 60].

2.3 Emotional Facial Expression Synthesis

Previous research on facial expression synthesis adopted the basic emotion theory [11, 30, 58], namely the "big six". The reason why basic emotion is widely used is not only because it is simple and intuitive, but also because most of the data used in early research are series of deliberate facial expression with extreme intensity rather than spontaneous occurring facial expression in natural conditions. For facial expression synthesis, most research focus on realistic-looking facial animation to enhance the naturalness while neglect delivering the inner emotional state or communication message by emotional facial expression. Ruttkay et al.[46] proposed the "Emotion Disc" and "Emotion Square" to generate continuum of emotional facial expression by two-dimensional navigation. Albrecht et al. [2] extended the work of MPEG-4 based facial expression synthesis [53], and designed a method to generate mixed expressions where the two-dimensional "Activation-Evaluation" emotion space is explored. Kshirsagar et al. [24] have proposed a multi-layered framework for facial expression synthesis which consists of high-level description of dynamic emotion intensity, mid-level configuration of static facial expressions, and low-level FAPs definition. Exploration on the paralinguistic meaning of facial movements is undertaken for isolated facial area, such as eyebrow and eye-lid [19]. However, there are few research results on the correlation between emotional dimensions and facial movements for synthesis purpose, at least not as much as the study of correlation between emotional dimensions and acoustic features for emotional speech synthesis [47]. To this end, the quantitative study on facial correlates with emotional dimensions is conducted in this chapter to realize the parametric facial expression synthesis for affective talking avatar.

2.4 FACS and FAP

The description and measurement of facial feature movement is an essential aspect to facial expression synthesis. The emotional sate is usually expressed by subtle change of facial movement in local area, while the prototypical facial expression with extreme intensity occurs relatively infrequently. To capture the local and subtle facial feature movements, Ekman et al. proposed facial action coding system (FACS), which is a human-observer-based system designed to capture the subtle movement of isolated facial features. The basic facial motion pattern is called "action unit" (AU) in FACS[14]. There are 44 AUs defined in FACS, among which 30 AUs are anatomically related to contraction of a specific set of

facial muscles, and the other 14 AUs are referred as miscellaneous actions. Trained experts can use these AUs individually or by their combinations to describe facial expression with 5-point ordinal scale measuring the intensity of muscle contraction. Another set of emotion-specified facial action coding system are proposed as EMFACS [17]. The FACS is originally designed to be human-readable not machine-readable, therefore there are many attempts to translate facial features movements to AUs for automatic facial expression analysis [6, 16, 52]. From the perspective of computer animation, the action units define the motion patterns of facial features, but not provide the quantitative temporal and spatial information required by face animation [20]. Therefore, the FACS system is widely adopted in filed of facial expression recognition, but it is not suitable for synthesis and animation purpose.

On the contrary to FACS, the facial animation framework developed under MPEG4 standards is designed completely from the perceptive of computer animation [34]. In accordance to action units, the facial animation parameters (FAPs) are designed based on the study of minimal perceptible actions and are closely related to sets of muscle actions [22, 38, 50, 55]. 84 facial feature points are specified to provide spatial references for defining 68 FAPs as quantitative movement of specific facial feature points. The FAPs represent a complete set of basic facial actions including head motion, tongue, eye, and mouth control, and thus it can be used to control synthetic face model to perform various kinds of facial expressions with varied intensity, even the exaggerated or distorted facial expression that only occurs on cartoon face.

From the view of facial expression synthesis, FAPs are much flexible and suitable to animate face models displaying various facial movements. However it is a tedious work to manipulate FAPs directly for generating emotional facial expression, since the FAPs control the facial feature points movement for low-level mesh deformation, rather than meaningful facial motion patterns, not mention the emotional facial expression. To this end, in this chapter, we propose the partial expression parameter (PEP) to depict the expressive movement of facial features (e.g. eyebrow, eye, mouth etc.), which combines the advantage of both FACS in human perceptible description of facial movement and FAP in low-level deformation of face model.

3 Layered Framework for Facial Expression Synthesis

To synthesize emotional facial expression for affective talking avatar, we designed a parametric layered framework based on emotion dimensions as Figure 2 shows. In this layered framework, the input is the high-level emotional parameters indicating the intended emotional state of the talking avatar, namely the pleasure-arousal-dominance (PAD) triple. The output is the expressive facial movement of virtual avatar. For geometric facial deformation, we adopt the facial animation parameters (FAPs), which is designed to control the face model points to generate various facial movements. However, it is too complicated for facial expression synthesizer to manipulate FAPs directly since they only define the motions (e.g. translations and rotations) of single facial feature points. To this end, a set of

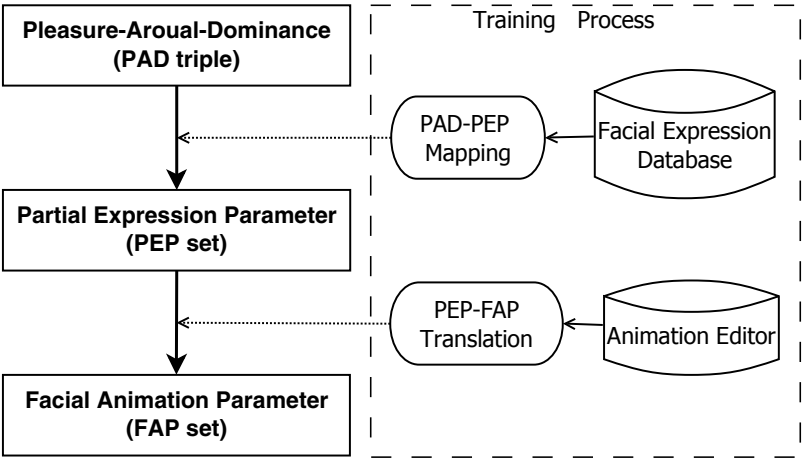


Fig. 2. Layered framework for PAD-driven facial expression synthesis

partial expression parameters (PEPs) are proposed to depict local emotional facial expressions, such as eyebrow-raise, mouth-bent and eye-open etc. The PEPs model the correlation among FAPs within local face region, and thus reduces the complexity of FAP-driven facial expression synthesis. To synthesize emotional facial expression, the PAD triple is first mapped to PEPs, after that the PEPs are translated to FAPs, finally the FAPs are used to animate the talking avatar to perform the facial expression which is expected to match the emotional state described by the input PAD triple.

With PAD as high-level description of emotional state, PEP as mid-level configuration of facial expression and FAP as low-level animation controller of face model, a layered framework for facial expression synthesis is implemented, where the PAD-PEP mapping model is trained on a facial expression database on talking avatar, and the PEP-FAP translation model is defined experimentally using a home-grown expression editor based on the study of FAP correlations [26] as well as the FACS manual [14]. In the following sections, we will introduce the approaches to implement the above layered framework and its integration to our emotional text to audio visual speech system [58].

4 PEP-FAP Translation Model

The partial expression parameter (PEP) is inspired by the study of FAP correlation and the observation of emotional facial expression within local face region. The PEP is proposed to depict the common emotional facial movement within specific facial regions. For low-level facial expression animation, a translation between PEP and FAP is needed. In this section, we describe the implementation of a linear interpolation to translate the PEP to its corresponding FAP subset.

4.1 Partial Expression Parameter (PEP)

As stated in section 2.4, the FAP is not very suitable as the configuration of emotional facial expression, since they only define the motions (e.g. translations and rotations) of single facial feature points. On the other side, there exist high correlations among different FAPs within the same facial regions as reported in the study of FAP interpolation algorithm for transmitting FAP in low bandwidth [26]. This motivates us to propose a new parameter to capture the FAP correlations as well as depict the common emotional facial movement in local facial regions, such as mouth-bent, eye-open and eyebrow-raise etc.

The 68 FAPs are categorized into 10 groups corresponding to different facial features such as eye, eyebrow, mouth, jaw and cheek etc. By referring the FAP group definition, the AU definition in FACS and observation on real facial expression image [23, 30], we proposed a set of PEP parameters defined in Table 1. The PEP aims to simulate the representative and meaningful emotional facial expression that commonly occurs in our daily life under the hypothesis that there exist fixed facial motion patterns to express specific emotional state.

Table 1. PEP definition with partial expression description

Face Region	PEP code (Left/Right)	Partial Expression Description	
		[0,-1]	[0,1]
Eye-brow	1.1(L/R)	Eyebrow lower down	Eyebrow raise up
	1.2(L/R)	Relax Eyebrow	Squeeze Eyebrow
	1.3(L/R)	In the shape of "\ /"	In the shape of "/ \"
Eye	2.1(L/R)	Close eye-lid	Open eye-lid
	2.2(L/R)	(Eyeball) look right	(Eyeball) look left
	2.3(L/R)	(Eyeball) look up	(Eyeball) look down
Mouth	3.1	Close mouth	Open mouth
	3.2	Mouth-corner bent down	Mouth-corner bent up
	3.3	Mouth sipped	Mouth protruded (pout)
	3.4	Mouth stretched in	Mouth stretched out
Jaw	4.1	Jaw move up	Jaw lower down
	4.2	Jaw move right	Jaw move left

For each PEP parameter, we defined its corresponding FAP subgroup, its direction of motion in positive value, its motion unit, and its translation mapping function to FAPs. The FAP subgroup is selected based on the FAP distribution within local regions. Most of PEP parameters are bidirectional which indicates the referred motion pattern can occur in two opposite directions to display different facial expression. For example, the mouth-corner-bent-up usually indicates happiness or joyful while mouth-corner-bent-down occurs together with sadness, anger etc. The PEP parameter, with a scalar value ranging from -1 to +1, simulates the motion intensity or degree of muscle contraction, and thus describes the continuous or subtle change of local facial expression. For better understanding, the movement of mouth-bent (PEP3.2) is illustrated in Figure 3 with its value ranging from -1 to +1. The motion unit of PEP

parameter is designed based on the motion unit of FAP (FAPU). In the following section, we introduce the PEP-FAP translation template which convert PEP parameter to corresponding FAPs.

(a) PEP3.2 (Mouth-bent) ranges from 0 to 1

(b) PEP3.2 (Mouth-bent) ranges from 0 to -1

Fig. 3. Partial expression movement of mouth-bent ranging in [-1, 1

4.2 PEP-FAP Translation Template

For each PEP parameter, a subgroup of FAP is defined for animating face model to perform corresponding local facial expression described by the PEP. Previous study on FAPs correlation indicates that there exist high correlations among different FAPs, which has been successfully exploited to interpolate the unknown FAP value from a set of decided FAPs [26]. This motivates us to implement a PEP-FAP translation template to convert the PEP to its corresponding FAPs.

According to Lava's study on FAP correlation [26], three sets of FAPs are analyzed which correspond to three main facial features, namely eyebrow, eye, and mouth, all of them are the most active and expressive facial features to display facial expression. The selection of correlated FAPs is based on a priori knowledge about the anatomy of human faces. Previous research defined key FAP as the FAP that has the highest sum of correlation coefficients with respect to other FAPs, and the other FAPs are interpolated by the key FAP [26].

Table 2. Related FAP group and key FAP for each PEP parameter

Face Region	PEP Code	FAP Number	Key FAP	Related FAP Group
Eye-brow	1.1	6	F33	[F33,F31,F35,F34,F32,F36]
	1.2	2	F37	[F37,F38]
	1.3	6	F31	[F31,F35, F33, F32, F36, F34]
Eye	2.1	4	F19	[F19,F21,F20,F22]
	2.2	2	F23	[F23,F24]
	2.3	2	F25	[F25,F26]
Mouth	3.1	12	F5	[F5,F10,F11,F52,F57,F58, F4,F8,F9,F51,F55,F56]
	3.2	4	F12	[F12,F13,F59,F60]
	3.3	2	F16	[F16,F17]
	3.4	4	F6	[F6,F7,F53,F54]
Jaw	4.1	1	F3	[F3]
	4.2	1	F15	[F15]

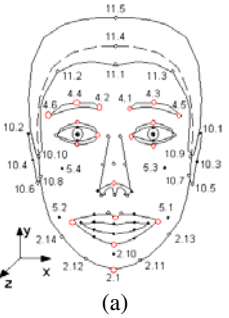
For each PEP parameter, we define its associated FAP group and key FAP as shown in Table 2. The selection of key-FAP is in accordance with the result in [26], and the other non-key FAP are linear interpolated by key-FAP. The mathematical form of the linear PEP-FAP translation template is defined in Equation 1. For the i-th PEP parameter in region R (P_i^R), the value of key-FAP (F_k) is determined by P_i^R directly, and the value of non-key FAP (F_j) is linear interpolated by the F_k with the coefficient α_k^j The F_k^{max} is the maximum value of F_k. F_k^{max} and α_k^j are experimentally determined using a home-grown facial expression editor and also referring the result in previous study [41].

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

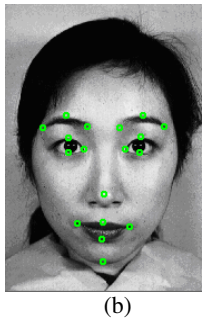
5 Pseudo Facial Expression Database

In order to model the relationship between PAD emotional dimensions and PEP parameters, a pseudo facial expression database is created. Here, the "*pseudo*" means that the database is not the real human facial expression but the synthetic facial expression on 3D talking avatar as shown in Figure 4.

(a)



(b)



(c)

Fig. 4. Avatar front view **Fig. 5.** Annotated facial points (b), FDPs (a) and FAPU (c)

We choose the Japanese Female Facial Expression (JAFFE) database as reference to create our pseudo database [30]. First the PEP parameter is extracted by measuring the facial features movement in real human image, and then is used to animate talking avatar to reproduce the similar facial expression. The pseudo facial expression database can be viewed online [62]. The establishment of pseudo facial expression database validates the effectiveness of PEP on synthesizing expressive facial movement on one hand. On the other hand, the individual (such as different age, face shape, and neutral state face) and appearance (such as the skin texture) effects on facial expression perception are eliminated by reproducing facial expression on a single talking avatar. Currently in this chapter, we focus on the geometric change of facial features such as the motion of eyebrow and lip etc.,

while the appearance change is not considered, such as the wrinkle, furrows and laugher line etc.

5.1 JAFFE Expression Database

The JAFFE expression database contains 213 expression images with 10 Japanese females posing three or four examples for each of seven basic expressions, namely *Neutral*, *Happy*, *Sad*, *Surprise*, *Angry*, *Disgust* and *Fear*. For each image 18 facial feature points are annotated manually in accordance to MPEG-4 facial definition points (FDPs) [34] as shown in Figure 5(a) and 5(b). The PEP parameters are then extracted based on these annotations to create pseudo facial expression database.

5.2 PEP Extraction for Pseudo Facial Expression Database

Based on the above annotation, 12 PEP parameters are extracted by measuring the movement of facial feature points defined in Table 3. The other 6 PEPs are not extracted by lack of depth information (PEP 2.2 and 2.3 for eye-ball movement, PEP 3.3 for mouth movement in z-direction), or the movement is not obvious (PEP 4.2 for horizontal jaw movement) in the database. The FAPU shown in Figure 5(c) is utilized to normalize PEP to eliminate the individual difference in face shape. By applying the extracted PEP in talking avatar animation, we create the pseudo facial expression database with similar synthetic expressions as real human.

Table 3. PEP measurement by facial point movement

Face Region	PEP code	PEP Measurement	Units (FAPU)
Eye-brow	1.1(L/R)	4.3.y (L)	ENS
		4.4.y (R)	
	1.2(L/R)	4.1.x (L)	ES
		4.2.x (R)	
	1.3(L/R)	$\arctan\left(\frac{4.1.y-4.5.y}{4.1.x-4.5.x}\right)$ (L)	AU
		$\arctan\left(\frac{4.2.y-4.6.y}{4.2.x-4.6.x}\right)$ (R)	
Eye	2.1(L/R)	3.9.y – 3.13.y (L)	IRISD
		3.10.y – 3.14.y (R)	
	2.2(L/R)	not-extracted	AU
	2.3(L/R)	not-extracted	AU
Mouth	3.1	8.1.y – 8.2.y	MNS
	3.2	$\frac{8.3.y+8.4.y}{2} - \frac{8.1.y+8.2.y}{2}$	MNS
	3.3	not-extracted	MNS
	3.4	8.3.x – 8.4.x	MW
Jaw	4.1	2.1.y	MNS
	4.2	2.1.x (not-extracted)	MW

A *k-means clustering* experiment is conducted on the pseudo database to validate the reproduced facial expressions. The clustering result is presented in Table 4, which states the number of seven intended expression in each PEP clustering group. The synthetic expression of each clustering center is illustrated in Figure 6. According to the clustering result, the extracted PEP can distinguish the basic emotions to some extent in that the PEP for the same expression is clustered into the same group with the exception of *Angry* and *Disgust*. This may be reasonable because that some similar facial movements are shared by *Angry* and *Disgust* as reported in [58].

Table 4. Sample number of 7 expressions (labels in the first row) in each PEP clustering center (C_i in the first column)

	HAP	SUP	ANG	NEU	SAD	FEAR	DIS	Total
C1	**25**	1	2	0	1	0	0	29
C2	4	**20**	1	0	0	3	1	29
C3	2	0	**18**	0	9	4	***12***	45
C4	0	6	1	**30**	0	0	0	37
C5	0	0	0	0	**15**	9	2	26
C6	0	3	0	0	4	**11**	2	20
C7	0	0	8	0	2	5	**12**	27
Total	31	30	30	30	31	32	29	213

Fig. 6. Synthetic expressions for PEP value of clustering center in Table 4

6 PAD-PEP Mapping Model

The PAD emotional model proposed by Mehrabin is not only a tool to describe the human emotional state, but also enables researcher to build quantitative and clear

connections between high-level human emotion perception and low-level acoustic/visual signals [4, 9, 60]. As the layered framework of PAD-driven facial expression synthesis stated in section 3, the PAD-PEP mapping model is learned from pseudo facial expression database. The database established in section 5 already contains the PEP configurations, and thus the PAD values for each synthetic expression need to be annotated for training the PAD-PEP mapping model.

6.1 PAD Annotation for Pseudo Expression Database

Based on the PAD scales proposed by Mehrabian [33], a Chinese-English bilingual version of abbreviated PAD emotion scales is provided [28], where the *P*, *A* and *D* values can be evaluated by the 12-item questionnaire shown in Table 5. The abbreviated PAD scale was proved as a versatile psychological measuring instrument that is capable of adapting to a variety of applications including emotion annotation.

Table 5. 12-item PAD questionnaire for expression annotation and evaluation

Emotion		-4	-3	-2	-1	0	1	2	3	4	Emotion	
Angry	愤怒的										Activated	有活力的
Wide-awake	清醒的										Sleepy	困倦的
Controlled	被控的										Controlling	主控的
Friendly	友好的										Scornful	轻蔑的
Calm	平静的										Excited	激动的
Dominant	支配的										Submissive	顺从的
Cruel	残忍的										Joyful	高兴的
Interested	感兴趣的										Relaxed	放松的
Guided	被引导的										Autonomou	自主的
Excited	兴奋的										Enraged	激怒的
Relaxed	放松的										Hopeful	充满希望的
Influential	有影响力										Influenced	被影响的

For each synthetic expression in the database, we annotate the *P*, *A* and *D* value using the 12-item questionnaire. During annotation, the annotator is required to describe the synthetic expression using 12 pairs of emotional words as given in Table5. For each pair of the emotional words, which are just like two ends of a scale, the annotator is asked choose the one that better describes the synthetic expression with a 9 level score varying from -4 to +4. The *P*, *A* and *D* values are then calculated from this questionnaire using the method described in [33]. Before annotating, the annotator is trained by an expert to use the PAD questionnaire.

The annotation result is summarized in Table 6, with the corresponding distribution in PAD emotional space shown in Figure 7. We can see that the PAD annotations for each basic emotion category are distributed in nearly different

Table 6. Average PAD annotation for the pseudo expression database

Intent Emotion	MEAN		
	P	A	D
ANG	-0.59	0.08	0.47
DIS	-0.59	-0.01	0.40
FEAR	-0.08	0.18	-0.39
HAP	0.63	0.40	0.29
NEU	0.03	-0.04	-0.07
SAD	-0.28	-0.12	-0.37
SUP	0.41	0.55	0.19

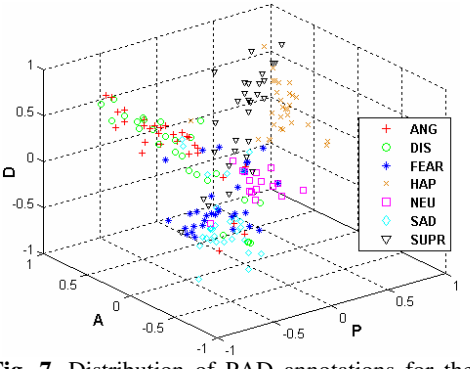


Fig. 7. Distribution of PAD annotations for the pseudo expression database

areas with the exception of *Angry* and *Disgust*. This result is consistent with the clustering result of PEP as stated in section 5.2.

6.2 PAD-PEP Mapping Model

Different from the previous work on realistic-looking facial animation, our work focuses on exploring the relationship between high-level emotion description and mid-level facial expression configuration, in other words, our goal is to implement a mapping between PAD and PEP. The pseudo expression database, which contains both the PAD annotation and PEP configuration, is used to train the PAD-PEP mapping model.

In order to find a proper emotion-expression mapping between PAD and PEP, we have tried the polynomial function with the first and second order as well as non-linear exponential function. As the experimental result reveals, it is the second order polynomial function that achieves the best fitting result, and its mathematic form is shown in Equation 2.

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

Where ***PEP*** is the PEP vector of expression configuration, and ***E*** is the PAD vector [*P, A, D*], $\boldsymbol{E}^2$ is a vector with each element the square of its counterpart in ***E*** respectively, i.e. [P^2, A^2, D^2], $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding coefficient matrix, $\boldsymbol{\delta}$ is the constant offset vector. Each dimension of ***PEP*** vector is estimated respectively with the same mathematic form as equation 3 shown. The PEP_i is the *i*-th dimension in ***PEP*** vector, and all the corresponding coefficients are indicated with the subscript *i* and *P*,*A* and *D* labels.

$$PEP_i = [\alpha_{Pi}, \alpha_{Ai}, \alpha_{Di}] \begin{bmatrix} P^2 \\ A^2 \\ D^2 \end{bmatrix} + [\beta_{Pi}, \beta_{Ai}, \beta_{Di}] \begin{bmatrix} P \\ A \\ D \end{bmatrix} + \delta_i \tag{3}$$

7 Integration to Emotional Text to Audio Visual Speech System

The emotional facial expression synthesis for talking avatar is a critical aspect for implementing the emotional text-to-audio-visual-speech system. Previous solutions on synthesizing emotional facial expression accompanying speech are achieved by combination of a set of basic facial expressions corresponding to different full-blown emotions [1, 5, 31, 49]. However we aims to build an emotional text-to-audio-visual-speech (ETTAVS) system based on emotion dimensions, namely the dimensions of pleasure, arousal and dominance, which has been successfully exploited in audio visual speech synthesis [9, 59, 60, 63]. This motivates us to integrate the PAD based emotional facial expression synthesis into the ETTAVS system. The overview of the system is shown in Figure 8. The input text message in first semi-automatically annotated with PAD values. The PAD value is then used to generate the emotional facial expression (i.e. PEP configuration), and to convert the synthetic speech generated by TTS engine to emotional speech [59, 60]. The emotional facial expression is then dynamically interpolated according to the speech prosody (see details in Section 8). After that, the synthetic head and facial movement are combined with the synthetic viseme generated by viseme synthesizer [54] to animate the 3D talking avatar, the audio-visual synchronization module controls the timing of audio speech play and visual speech animation [58].

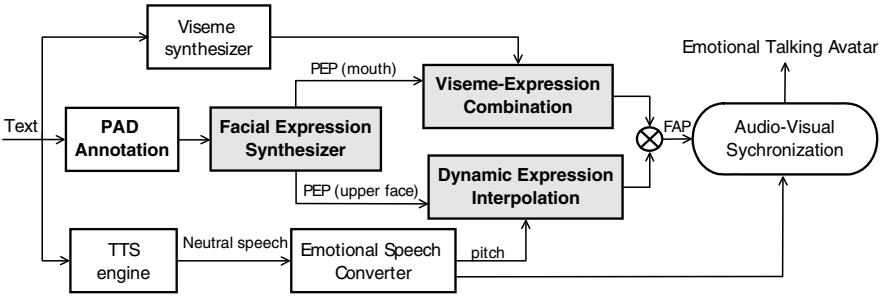


Fig. 8. Overview of the Emotional Text to Audio Visual Speech System

8 Experiments and Results

8.1 Experiments on PAD-PEP Mapping Function

The pseudo expression database consists of 213 expression samples, which are extracted from the JAFFE database consisting of 10 subjects and 7 expressions. In order to reduce the limitation caused by the small database, we adopt the K-fold cross-validation method to train the PAD-PEP mapping function. For the test set, we select 10*2=20 expression image samples by randomly selecting 2 different expressions for all 10 subjects. The training set, which consists of the rest 193 expression image samples, is then divided into 10 subsets with each subset covering all 10 subjects and each subject with 1 or 2 expression samples. By such

division scheme, we are able to capture the common facial expression movement shared by different people as much as possible.

The coefficients of the PAD-PEP mapping function is estimated using the least square errors method, and each dimension of PEP (e.g. PEP_i) is estimated separately with the same function as shown in equation 3, which means that we finally have 12 sets of coefficients corresponding to 12 PEP parameters. In the K-fold cross-validation training process (k=10), there are 10 iterations corresponding to 10 validating subset. In each iteration, we calculate the correlation coefficient between real and estimate data on the validating subset. The correlation is taken as criteria to evaluate the fitting performance of the trained function. The minimum, maximum and average value of correlation coefficient is summarized in Table 7. The trained function with the average fitting performance among all the 10 iterations is chosen as the final result and used to evaluate the performance on test set.

Table 7. Correlation coefficients for evaluating PAD-PEP mapping function

PEP code	Validation set (k=10)			Test set
	Min	Max	Avg	
1.1L	0.56	0.88	0.77	0.74
1.1R	0.52	0.86	0.77	0.77
1.2L	0.13	0.82	0.43	0.51
1.2R	0.13	0.64	0.44	0.62
1.3L	0.63	0.89	0.79	0.80
1.3R	0.56	0.91	0.77	0.82
2.1L	0.53	0.84	0.75	0.60
2.1R	0.51	0.85	0.76	0.62
3.1	0.36	0.89	0.65	0.52
3.2	0.24	0.93	0.71	0.75
3.4	0.06	0.77	0.52	0.78
4.1	0.36	0.83	0.57	0.66
Average	0.38	0.84	0.66	0.69

From the result, we can see that the performance of mapping function on test set is acceptable with correlation coefficient be around 0.70. There is still some space for performance to be improved. This may be explained as that the pseudo expression database consist only 10 subjects with 7 basic expressions, which is insufficient for training the mapping function to capture common expression patterns among different people. The possible solution is either to increase the number of the subjects or to collect more expression images for a specific subject.

8.2 Synthetic Expression and Perceptual Evaluation

A series of perceptual evaluation is conducted to evaluate the performance of the layered framework for expression synthesis. The synthetic expressions for the 6 basic emotion words (*happy*, *surprise*, *sad*, *scared*, *angry*, and *disgusting*) and some other candidate emotion words with corresponding PAD value are shown in Figure 9. 14 subjects are then invited in the evaluation to finish the 12-item PAD

Fig. 9. PAD-driven synthetic expression for selected emotions (Emotion [P,A,D])

questionnaire as shown in Table 5 for each synthetic facial expression image. Another emotion labeling experiment is also conducted where the subjects are required to select one word that best describe the synthetic expression from 14 candidate emotion words, which are *Happy*, *Optimism*, *Relax*, *Surprise*, *Mildness*, *Dependent*, *Bored*, *Sad*, *Scared*, *Anxious*, *Scornful*, *Disgusting*, *Angry*, and *Hostile*. The PAD values for each basic emotion word and the PAD perception values for synthetic expression are summarized in Table 8.

Table 8. Result of PAD evaluation and emotion labeling. (The PAD value of *Hostile* is [-0.42, 0.20, 0.22])

PAD-driven synthetic expression	Original PAD of emotion words			Evaluation PAD of Synthetic image			Expression Label (with voting percent)	
	P	A	D	P	A	D		
Happy	0.55	0.24	0.28	0.42	0.12	0.10	Happy	(67%)
Surprise	0.34	0.34	0.04	0.36	0.45	-0.05	Surprise	(100%)
Sad	-0.18	0.03	-0.14	-0.01	-0.26	-0.27	Sad	(42%)
Scared	-0.19	0.26	-0.13	0.01	-0.04	-0.25	Sad	(50%)
Angry	-0.40	0.22	0.12	-0.17	0.02	-0.08	Hostile	(58%)
Disgust	-0.36	0.08	0.13	-0.56	0.15	0.44	Disgusting	(50%)

A voting method is applied to determine the emotion category for each synthetic facial expression image based on the result of expression labeling. The PAD correlation coefficients between emotion words and synthetic expression are 0.89(P), 0.68(A) and 0.70(D) respectively, this result is consistent with the reliability and validity of Chinese version of abbreviated PAD emotion scales reported in [28]. Confusion between *Sad* and *Scared* as well as *Angry* and *Hostile* is found in expression labeling experiment, it is because that the input PAD value for each of the two confused words are very close, which originates from the

confusion in human understanding between these emotion words and thus leads to the similarity between synthetic expressions.

The experimental results indicates that the PAD emotional parameters can be used in describing the emotion state as well as facial expression; and the proposed layered framework is effective for PAD-driven facial expression synthesis.

8.3 Dynamic Facial Expression Accompanying Speech

As stated in Section 7, the emotional facial expression is integrated into our ETTAVS system. To enhance the emotional expressivity of facial expression accompanying speech, the synthetic emotional facial expression by PAD-PEP-FAP approach is dynamically interpolated in accordance with the prosodic change in emotional speech on sentence level. Based on the observation of the video of real speaker's dynamic facial expression in a speech, we proposed the "*peak-intensify*" rule that the facial expression will be intensified when the speech prosody get aroused (i.e. the peak time). The speech pitch (F0) is utilized as important cues to locate the *peak* time, and the PEP value for facial expression at peak time is calculated as Equation (4), while the dynamics of facial expression over whole sentence is obtained by interpolating the PEP from zero to the value at peak time using the Piecewise Cubic Hermite Interpolating method [18]. Figure 10 illustrates the interpolation result of eyebrow squeeze (PEP1.2) for a sentence with disgust emotion. Figure 11 presents the synthetic facial expression series and speech signal.

$$\boldsymbol{PEP}^{peak} = \alpha^{C} \bullet \boldsymbol{PEP}^{main} \quad \left(\alpha = \left|\boldsymbol{F}_0^{peak} / \boldsymbol{F}_0^{average}\right|\right) \tag{4}$$

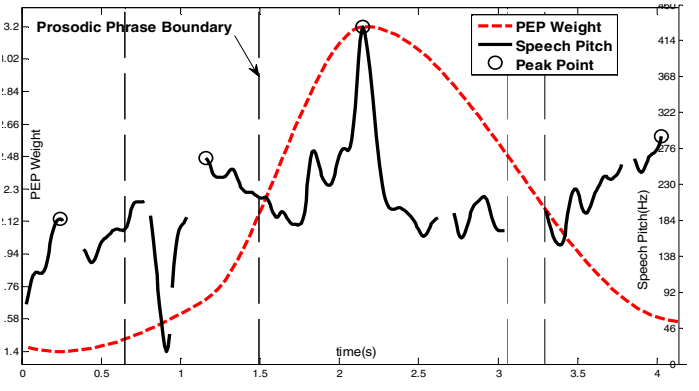


Fig. 10. PEP Interpolation based on Speech Pitch(F0) for a sentence

The *peak-intensify* rule is only applied to the upper face area, while the mouth area is not affected by this rule since the shape of mouth must be consistent with the viseme configuration. For the mouth animation in emotional speech, we take a combination of viseme and facial expression. A linear weighted function is

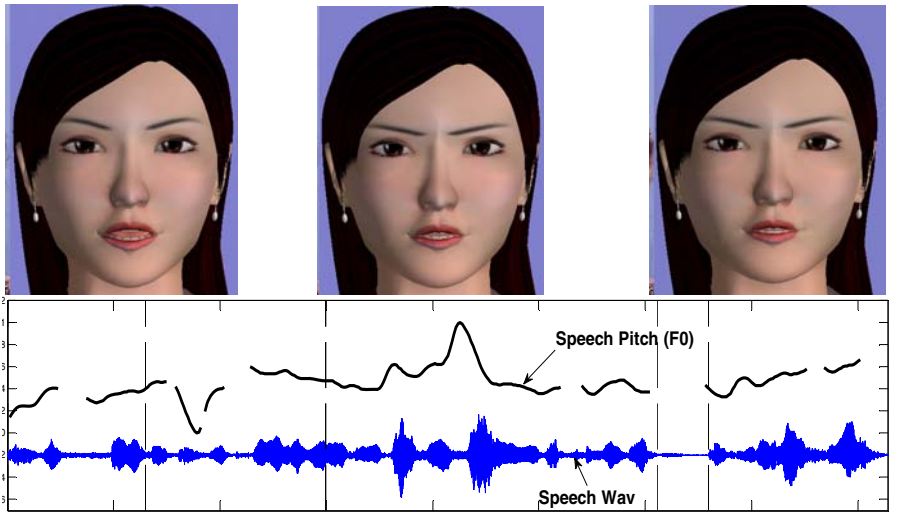


Fig. 11. Dynamics of emotional facial expression and speech prosody in a sentence

applied to merge the animation parameter (FAP) from viseme and facial expression, as shown in Equation 5. It should be noticed that, the merge function is only applied to the FAPs that control the width and height of outer lip corner, and the selection of these FAPs is experimentally manually decided to achieve best animation result. The coefficient α is manually set to determine whether the viseme dominates the lip movement or the facial expression. In this work, we take α=0.6 to obtain the final animation parameter for mouth.

$$\boldsymbol{FAP}_{\text{final}}^{\text{mouth}} = \alpha \boldsymbol{FAP}_{\text{viseme}}^{\text{mouth}} + (1-\alpha)\boldsymbol{FAP}_{\text{expr}}^{\text{mouth}} \tag{5}$$

9 Conclusion and Further work

This chapter introduces our work on synthesizing emotional facial expression on expressive talking avatar in emotional text to audio visual speech system. A layered framework of parametric facial expression synthesis based on emotion dimensions (i.e. pleasure-arousal-dominant) is proposed. The facial expression is synthesized according to the high-level PAD emotional parameters, which indicates the intended or desired emotional state of the talking avatar. The mid-level partial expression parameter (PEP) is designed to depict the facial expression movement within a specific face region (i.e. eyebrow, eye and mouth etc.). The MPEG-4 facial animation parameter (FAP) is adopted as low-level parameter for direct manipulation of facial points on the avatar. The PAD-PEP mapping model and PEP-FAP translation template are then implemented to translate the PAD parameters to PEP parameters and then to FAP parameters for facial expression animation.

The proposed approach is effective for facial expression synthesis, and enhances the emotional expressivity of talking avatar. Together with our previous work on expressive head movement synthesis [63], the emotional facial animation is integrated into the emotional text to audio visual speech system (ETTAVS). A dynamic interpolation rule and a linear combination strategy are proposed to achieve more natural facial animation accompanying speech.

Further work will focus on the semantic facial expression of talking avatar that helps to deliver the communicative message in human conversation. Our next goal is to implement a conversational engaging talking avatar for spoken dialog system.

Acknowledgments. This work is supported by National Natural Science Foundation of China (60805008,90820304) ,the National Basic Research Program of China (973 Program) (No. 2006CB303101) and the National HighTechnology Research and Development Program("863"Program) of China (No. 2007AA01Z198). This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies. We also thank Sirui Wang from Institute of Psychology, Chinese Academy of Science for providing the 14 typical emotion words and corresponding PAD values in the perceptual evaluation.

References

[1] Albrecht, I., Haber, J., Seidel, H.P.: Automatic generation of non-verbal facial expressions from speech. In: Proc. Computer Graphics International 2002, pp. 283–293 (2002)

[2] Albrecht, I., Schröder, M., Haber, J., Seidel, H.P.: Mixed feelings: expression of non-basic emotions in a muscle-based talking head. Virtual Reality 8(4), 201–212 (2005)

[3] Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: analysis and synthesis. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1075–1086 (2007)

[4] Cao, J., Wang, H., Hu, P., Miao, J.: PAD model based facial expression analysis. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Porikli, F., Peters, J., Klosowski, J., Arns, L., Chun, Y.K., Rhyne, T.-M., Monroe, L. (eds.) ISVC 2008, Part II. LNCS, vol. 5359, pp. 450–459. Springer, Heidelberg (2008)

[5] Cao, Y., Tien, W.C., Faloutsos, P., Pighin, F.: Expressive speech-driven facial animation. ACM Trans. on Graph 24 (2005)

[6] Cohn, J., Zlochower, A., Lien, J., Kanade, T.: Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. Psychophysiology 36, 35–43 (1999)

[7] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction, vol. 18(1), pp. 32–80 (2001)

[8] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: FEELTRACE: an instrument for recording perceived emotion in real time. In: Proceedings of the ISCA workshop on speech and emotion, Northern Ireland, pp. 19–24 (2000)

[9] Cui, D., Meng, F., Cai, L., Sun, L.: Affect related acoustic features of speech and their modification. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 776–777. Springer, Heidelberg (2007)

[10] Darwin, C.: The expression of the emotions in man and animals. University of Chicago Press, Chicago (1965)
[11] Du, Y., Lin, X.: Emotional facial expression model building. Pattern Recognition Letters 24(16), 2923–2934 (2003)
[12] Ekman, P.: Universals and cultural differences in facial expressions of emotion. In: Cole, J. (ed.) Proc. Nebraska Symposium on Motivation, vol. 19, pp. 207–283 (1971)
[13] Ekman, P.: About brows: emotional and conversational signals. In: Human ethology: claims and limits of a new discipline: contributions to the Colloquium, pp. 169–248. Cambridge University Press, England (1979)
[14] Ekman, P., Friesen, W.: Facial action coding system: A technique for the measurement of facial movement. Tech. rep. Consulting Psychologists Press (1978)
[15] Faigin, G.: The Artist's Complete Guide to Facial Expression. Watson-Guptill (2008)
[16] Fasel, B., Luttin, J.: Recognition of asymmetric facial action unit activities and intensities. In: Proceedings of International Conference of Pattern Recognition (2000)
[17] Friesen, W., Ekman, P.: Emfacs-7: emotional facial action coding system, Unpublished manuscript, University of California at San Francisco (1983)
[18] Fritsch, F.N., Carlson, R.E.: Monotone piecewise cubic interpolation. SIAM Journal on Numerical Analysis 17, 238–246 (1980)
[19] Granstrom, B., House, D.: Audiovisual representation of prosody in expressive speech communication. Speech Communication 46(3-4), 473–484 (2005)
[20] Hong, P., Wen, Z., Huang, T.S.: Real-time speech-driven face animation with expressions using neural networks, vol. 13(4), pp. 916–927 (2002)
[21] Ibanez, J., Aylett, R., Ruiz-Rodarte, R.: Storytelling in virtual environments from a virtual guide perspective. Virtual Reality 7, 30–42 (2003)
[22] Kalra, P., Mangili, A., Magnenat-Thalmann, N., Thalmann, D.: Simulation of facial muscle actions based on rational free form deformations. In: Proc. Eurographics 1992, pp. 59–69 (1992)
[23] Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
[24] Kshirsagar, S., Escher, M., Sannier, G., Magnenat-Thalmann, N.: Multimodal animation system based on the mpeg-4 standard. In: Proceedings Multimedia Modelling 1999, pp. 21–25 (1999)
[25] Lang, P.J.: Behavioral treatment and bio-behavioral assessment: computer applications. In: Technology in mental health care delivery systems, pp. 119–137. Ablex, Norwood (1980)
[26] Lavagetto, F., Pockaj, R.: An efficient use of mpeg-4 fap interpolation for facial animation at 70 bits/frame, vol. 11(10), pp. 1085–1097 (2001)
[27] Li, S.Z., Jain, A.K.: Handbook of Facial Recognition. Springer, New York (2005)
[28] Li, X., Zhou, H., Song, S., Ran, T., Fu, X.: The reliability and validity of the chinese version of abbreviated PAD emotion scales. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 513–518. Springer, Heidelberg (2005)
[29] Linden Research, Inc.: Second life: Online 3D virtual world, `http://secondlife.com/`
[30] Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998)
[31] Mana, N., Pianesi, F.: Hmm-based synthesis of emotional facial expressions during speech in synthetic talking heads. In: Proceeding of 8th International Conference on Multimodal Interfaces (ICMI 2006), Banff, AB, Canada, pp. 380–387 (2006)
[32] Mehrabian, A.: Communication without words. Psychology Today 2, 53–56 (1968)

[33] Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology: Developmental, Learning, Personality, Social 14(4), 261–292 (1996)
[34] Motion Pictures Expert Group: ISO/IEC 14496-2.: International standard, information technology-coding of audio-visual objects. part 2: Visual; amendment 1: Visual extensions (1999/Amd. 1: 2000(E))
[35] Mower, E., Mataric, M.J., Narayanan, S.: Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. IEEE Transaction on Multimedia 11(5), 843–855 (2009)
[36] Oddcast Inc.: Personalized speaking avatars service, http://www.voki.com/
[37] Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The Measurement of Meaning. University of Illinois Press (1957)
[38] Parke, F.I.: Parameterized models for facial animation, vol. 2(9), pp. 61–68 (1982)
[39] Ekman, P., Rosenberg, E.L.: What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS). Oxford University Press, US (2005)
[40] Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
[41] Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S.: Parameterized facial expression synthesis based on MPEG-4. EURASIP Journal on Applied Signal Processing 2002, 1021–1038 (2002)
[42] Reallusion, Inc.: Crazytalk for skype, http://www.reallusion.com/crazytalk4skype/
[43] Plutchik, R.: A general psychoevolutionary theory of emotion. In: Emotion: Theory, research, and experience. Theories of emotion, vol. 1, pp. 3–33. Academic, New York (1980)
[44] Russell, J., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality 11, 273–294 (1977)
[45] Russell, J.A., Fernez-Dols, J.M. (eds.): The Psychology of Facial Expression. Cambridge University Press, Cambridge (1997)
[46] Ruttkay, Z., Noot, H., Hagen, P.: Emotion disc and emotion squares: Tools to explore the facial expression space. Computer Graphics Forum 22(1), 49–53 (2003)
[47] Schröder, M.: Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 209–220. Springer, Heidelberg (2004)
[48] Schröder, M.: Expressing degree of activation in synthetic speech. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1128–1136 (2006)
[49] Tang, H., Huang, T.S.: MPEG4 performance-driven avatar via robust facial motion tracking. In: International Conference on Image Processing, ICIP, San Diego, CA, United state, pp. 249–252 (2008)
[50] Terzopolous, D., Waters, K.: Physically-based facial modeling, analysis and animation. Journal of Visualization and Computer Animation 1, 73–80 (1990)
[51] Theune, M., Meijs, K., Heylen, D., Ordelman, R.: Generating expressive speech for storytelling applications. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1137–1144 (2006)
[52] Tian, Y.I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis, vol. 23(2), pp. 97–115 (2001)
[53] Tsapatsoulis, N., Raousaiou, A., Kollias, S., Cowie, R., Douglas-Cowie, E.: Emotion recognition and synthesis based on MPEG-4 FAPs. In: MPEG-4 facial animation-the standard implementations applications, pp. 141–167. Wiley, Hillsdale (2002)
[54] Wang, Z., Cai, L., AI, H.: A dynamic viseme model for personalizing a talking head. In: Sixth International Conference on Signal Processing (ICSP 2002), pp. 26–30 (2002)

[55] Waters, K.: A muscle model of animating three dimensional facial expression. Computer Graphics 22(4), 17–24 (1987)
[56] Welbergen, H., Nijholt, A., Reidsma, D., Zwiers, J.: Presenting in virtual worlds: Towards an architecture for a 3D presenter explaining 2D-presented information. In: Maybury, M., Stock, O., Wahlster, W. (eds.) INTETAIN 2005. LNCS (LNAI), vol. 3814, pp. 203–212. Springer, Heidelberg (2005)
[57] Whissell, C.: The Dictionary of Affect in Language Emotion: Theory, Research and Experience. In: The Measurement of Emotions, vol. 4, pp. 113–131. Academic Press, London (1989)
[58] Wu, Z., Zhang, S., Cai, L., Meng, H.M.: 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, vol. 4, pp. 1802–1805 (2006)
[59] Yang, H., Meng, H.M., Cai, L.: Modeling the acoustic correlates of expressive elements in text genres for expressive text-to-speech synthesis. In: INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, vol. 4, pp. 1806–1809 (2006)
[60] Yang, H., Meng, H.M., Wu, Z., Cai, L.: Modelling the global acoustic correlates of expressivity for Chinese text-to-speech synthesis. In: Proc. IEEE Spoken Language Technology Workshop, pp. 138–141 (2006)
[61] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transaction on Multimedia 31(1), 39–58 (2009)
[62] Zhang, S.: Pseudo facial expression database, http://hcsi.cs.tsinghua.edu.cn/Demo/jaffe/emot/index.php
[63] Zhang, S., Wu, Z., Meng, H.M., Cai, L.: Head movement synthesis based on semantic and prosodic features for a Chinese expressive avatar. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007, vol. 4, pp. 837–840 (2007)