# Emotional Talking Agent: System and Evaluation

Shen Zhang, Jia Jia, Yingjin Xu, Lianhong Cai
Key Laboratory of Pervasive Computing, Ministry of Education,
State Key Laboratory on Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

***Abstract*—In this paper, we introduce a system that synthesizes the emotional audio-visual speech for a 3-D talking agent by adopting the PAD (Pleasure-Arousal-Dominance) emotional model. A GMM-based method is introduced to predict variation of acoustic features for emotional speech by PAD values, and a parametric framework of PAD-driven emotional facial expression synthesis is built. As the focus of this paper, we performed a series of perceptual evaluations to understand the reinforcement effect of vocal and facial expression of emotion, and investigated the usefulness and effectiveness of the emotional talking agent in human computer speech communications. Three questions are addressed: 1) To what extent do different interfaces affect human's comprehension of emotion? 2) How accurate the emotional information is conveyed by the talking agent? 3) Is the multimodal (audio-visual) interface helpful to human's emotion comprehension? An evaluation involving 19 participants was conducted to compare the effect of different interfaces (speech, mute agent and talking agent) on improving human's comprehension of emotion. The experimental results unveil the significant mutually reinforcing relationship between audio and video modality in emotion perception, and show that the users have a strong preference to multimodal interface for better comprehension of emotion. The results also prove the effectiveness of our PAD based emotional talking agent synthesis system.**

***Keywords- PAD, talking agent, multimodal reinforcement, emotion perception***

## I. INTRODUCTION

Comparing with face-to-face human communication, it is a big challenge for computer to express emotions as natural as human beings. Previous studies on emotion synthesis have focused on basic emotion categories, and the facial and vocal expression of emotion is dealt with respectively [1, 2]. However, in human communications, people usually express emotions not belonging to a single category but being rich and varied, and occurred in the form of multi-modality (speech, face, gesture etc.).

As the accelerated development of speech synthesis and facial animation, talking agent, a virtual character with natural speech, vivid facial expression and body gestures, has become a common interface in human computer communications. However, there are still many problems to be solved before we can realize an emotionally intelligent talking agent, such as how to express emotion in abundant variety rather than limited categories, and how to ensure the consistency of emotion expression in both auditory and visual modality?

To this end, we introduce an expressive Chinese talking agent system [3] which provides synthetic speech and synchronized facial expression as well as head motion. The PAD model [4], which enables us to measure emotional state in three dimensions (Pleasure, Arousal and Dominance), is adopted to synthesize speech and facial expressions in variety of emotional states. With the talking agent synthesized based on the PAD model, we aim to provide an expressive and intelligent multimodal interface which will facilitate the human computer communications.

Although clearly promising, the value of talking agent as a multimodal interface in improving human's comprehension of emotion needs to be validated. The joint effect of audio-visual modality on emotion recognition has been exploited and validated [5]. Some reports on the emotional McGurk effect shows the visual modality was in general more reliable than audio at conveying emotions in the condition of conflicting auditory and visual expression of emotion [6]. However the audio-visual reinforcement effect on emotion synthesis has not been studied to the same extent as recognition. For the application of talking agent, the vocal and facial expression of emotion is expected to be mutually reinforced to deliver more reliable information than unimodal condition.

In this paper, we first introduce the emotional talking agent synthesis system based on PAD emotion models. To validate the audio-visual interface (e.g. talking agent) in enhancing human's comprehension of emotional information, we conduct perceptual evaluations on our talking agent system. 19 participants were invited to a series of perceptual evaluations which are designed to study the effect of different interfaces (speech, mute agent and talking agent) on human's comprehension of emotion, in which three research questions are addressed:

(1) To what extent (how) do different interfaces (modality) affect human's comprehension of emotion?

(2) How accurate the emotional information is conveyed by the talking agent?

(3) Is the multimodal (audio-visual) interface helpful in improving user's comprehension of emotion?

The experiment results unveil the mutually reinforcing relationship between audio and video modality in emotion perception, and show that users have a strong preference to multimodal interface in human computer speech communication which will help their comprehension of the

This work is supported by NSFC (90820304, 60805008), the "863"Program of China（2009AA011905）, and the "973" Program (2006CB303101).

affective meanings. The results also prove the effectiveness of the PAD based emotional talking agent synthesis system.

## II. EMOTIONAL AUDIO-VISUAL SPEECH SYNTHESIS ON 3D TALKING AGENT

### A. PAD Emotion Space

The PAD emotion space is not only a scale for describing human emotional state, but can help us to build a clear connection between high-level human perception and low-level acoustic/visual signals. According to PAD theory, emotions are not limited to isolated categories but can be described along three nearly independent continuous dimensions: Pleasure-Displeasure (*P*), Arousal-Nonarousal (*A*), and Dominance-Submissiveness (*D*). Emotional states are not described by a set of categories or descriptive words, but denoted as points in three-dimensional PAD emotion space. In this way, different emotions can be distinguished quantitatively along the *P*, *A* and *D* dimensions respectively. A subjective scales [7] for evaluating PAD values by a 12-item questionnaire is adopted to obtain PAD annotation for affective speech and facial expressions. As shown in Table I, there are 12 pairs of descriptive word and for each pair of the words (Emotion-A and Emotion-B), which are just like two ends of a scale, the annotator is required to choose one of them that better describes the affective speech or facial 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, and are normalized to [-1, +1]. The 12-item PAD scale has been proved as a versatile psychological measuring instrument which is capable of adapting to a variety of applications including emotion annotation.

The overview of our work on affective audio-visual speech analysis and synthesis is shown in Figure1. We aim to model the correlation between PAD emotion space and acoustic/visual feature vector space, and then try to predict the affective acoustic features and visual features for emotional talking agent.

TABLE I. 12-ITEM QUESTIONNAIRE FOR PAD EVALUATION

| Emotion-A | Emotion-B | Emotion-A | Emotion-B |
|---|---|---|---|
| Angry | Activated | Cruel | Joyful |
| Wide-awake | Sleepy | Interested | Relaxed |
| Controlled | Controlling | Guided | Autonomous |
| Friendly | Scornful | Excited | Enraged |
| Calm | Excited | Relaxed | Hopeful |
| Dominant | Submissive | Influential | Influenced |

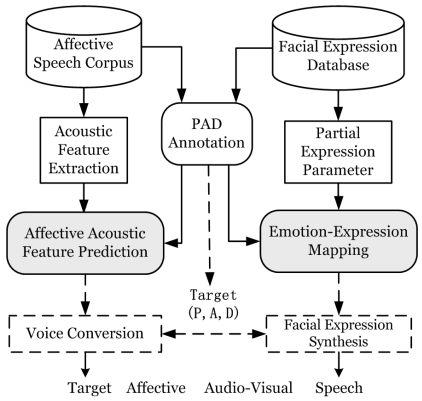



Figure 1. Overall framework of affective audio-visual speech synthesis

### B. Emotional Audio-Visual Speech Synthesis

In this section, we introduce the overall architecture of our emotional talking agent system. As in typical TTS synthesis applications, we use text as input of the approach as well as the target PAD. The emotional audio-visual speeches are synthesized on a 3D agent [3] by the following steps:

- Neutral speech synthesis and feature extraction

We first synthesize the neutral speech by TTS engine [8]. The TTS engine can also provide the information of Pin Yin, pitch sequence, syllable duration, and the prosodic word/phrase boundaries. Then we extract the acoustic features from neutral speech, including maximum pitch, pitch range, duration and energy.

- GMM-based emotional acoustic feature prediction

We use the GMM algorithm proposed by Kain [9] to build the prediction model of emotional acoustic features. The outputs of GMM are the acoustic feature differences between emotional speeches and neutral speeches, while the inputs are the annotated PAD values. The prediction process of GMM is shown in (1)-(3).

$$F(\boldsymbol{x}) = \sum_{i=1}^{M} h_i(\boldsymbol{x}) \left[ \boldsymbol{\mu}_i^{y} + \boldsymbol{\Sigma}_i^{yx} \boldsymbol{\Sigma}_i^{xx} \left( \boldsymbol{x} - \boldsymbol{\mu}_i^{y} \right) \right]^T \tag{1}$$

$$h_i(\boldsymbol{x}) = \alpha_i N\left(\boldsymbol{x}; \boldsymbol{\mu}_i^{x}, \boldsymbol{\Sigma}_i^{xx}\right) \Big/ \sum_{i=1}^{M} \alpha_i N\left(\boldsymbol{x}; \boldsymbol{\mu}_i^{x}, \boldsymbol{\Sigma}_i^{xx}\right) \tag{2}$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix} \quad \boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{x} \\ \boldsymbol{\mu}_i^{y} \end{bmatrix} \tag{3}$$

where $F(x)$ represents the output vector, $x$ represents the input vector. ($\alpha$, $\mu$, $\Sigma$) is estimated using the EM (Expectation Maximum) algorithm in the training process. The emotional speech is obtained by modifying the acoustic features of the neutral speech using TD-PSOLA algorithm.

- PAD-driven facial expression synthesis

According to the MPEG-4 FAP (Facial Animation Parameter) framework, we synthesize facial expression using based on the PAD-PEP-FAP mapping model. The PEP is proposed as Partial Expression Parameter to describe the expressive facial movement in specific regions based on the correlations between FAPs [10]. The mapping model is formulated as (4) and (5).

$$\boldsymbol{PEP} = \alpha \boldsymbol{E}^2 + \beta \boldsymbol{E} + \delta \tag{4}$$

where ***PEP*** is the PEP configuration of static facial expression, ***E*** is the corresponding PAD values, and $\boldsymbol{E}^2$ is a vector in which each element is the square value of its counterpart in ***E***, i.e., [$P^2$, $A^2$, $D^2$]. $\alpha$ and $\beta$ are the corresponding coefficient matrix, $\delta$ is a constant offset vector. The PEP is translated to FAP by a linear interpolation function. For the $i$th PEP in region $R$ ($P_i^R$), we defined a key-FAP ($F_k$) which has the highest correlations with other FAPs in regions $R$ as reported in [10]. The value of $F_k$ is linearly determined by $P_i^R$ directly with a bound of $F_k^{\max}$. The value of non-key FAP ($F_j$) is linearly interpolated by the key-FAP ($F_k$) with a coefficient $\alpha_k^j$. The $F_k^{\max}$ and $\alpha_k^j$ are both experimentally determined.

$$\begin{cases} F_k = P_i^R \bullet F_k^{\max} & (P_i^R \in [-1,+1]) \\ F_j = \alpha_k^j \bullet F_k & (\alpha_k^j \in [-1,+1],\ k \neq j) \end{cases} \quad (5)$$

- Viseme Generation

Generate the viseme based on the previous work on Chinese dynamic viseme by our lab [11]. There are totally 20 Chinese static viseme categories. A weight blending dynamic viseme model is adopted to describe the viseme dynamics in co-articulation context, and speed or pause duration change in spontaneous speaking. Based on MPEG-4 facial animation framework, we extract the FAPs of lip movement from videos for each of the Chinese phoneme. The static viseme set is determined by constructing the visual confusion tree based on a measure of normalized visual distances of each phoneme, and the number of viseme classes is determined by this confusion tree [11].

- Audio Visual Synchronization

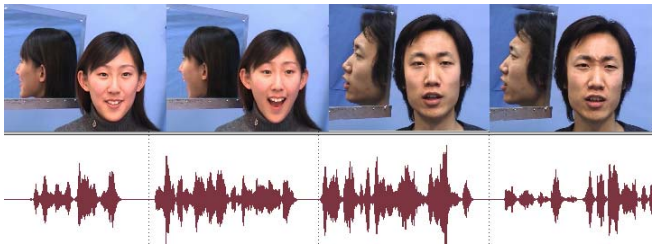
We use the Dynamic Bayes Network (DBN) to solve the synchrony problem between audio and visual modalities. We proposed a DBN based Audio-Visual Correlative Model (AVCM) [12], where the loose timing synchronicity between audio and video streams is restricted by word boundaries. The model inputs are the predicted emotional acoustic features and the FAPs related to the viseme. The previously-trained DBN based AVCM is applied to calculate the probability score, which is used as a measure of synthesis error. Then we use the downhill simplex method [12] to adjust the FAPs until we got the smallest synthesis error. After the synchronization between emotional speech and its viseme, the FAPs related to viseme and facial expression in mouse region are linearly combined, shown as (6). We take $\alpha$=0.8 to obtain the final animation parameters for mouth.

$$\boldsymbol{FAP}_{\text{final}}^{\text{mouth}} = \alpha \boldsymbol{FAP}_{\text{viseme}}^{\text{mouth}} + (1-\alpha)\boldsymbol{FAP}_{\text{expr}}^{\text{mouth}} \quad (6)$$

## III. EVALUATION ON EMOTIOANL TALKING AGENT

### A. Dataset Setup and Quality Evaluation

We have produced an emotional speech database which has 132 segments of human emotional audio-visual speech. The database was recorded by six subjects (four females and two males). There are two semantically neutral Chinese sentences, and the subjects were asked to speak the sentence under different emotion scenarios, including *Neutral*, *Relaxed*, *Submissive*, *Surprised*, *Happy*, *Disgusted*, *Scornful*, *Fearful*, *Sad*, *Anxiety* and *Angry*. Figure 2 presents some examples.



(a) A female speaker with "Happy" (b) A male speaker with "Disgusted"

Figure 2. The data sample of emotional audio.visual speech

Three labelers are invited to annotate the PAD value for each emotional speech, using a 12-item questionnaire [8]. Since the PAD values are provided by three human labelers, we first evaluate the quality of these annotations before analysis. The consistency among the three labelers is measured by computing the standard deviation of PAD values for each sentence in one modality. The PAD values have a small standard deviation (average is 0.1), which indicates the consistency among the three labelers.

### B. Statistical Analysis on emotional perception

We first synthesize 121 audio-visual speeches on 3-D talking agent using 121 PAD values of 11 emotions from the corpus. We invite four participants to annotate each synthetic speech with PAD again but with three different interfaces:

A – the communication channel with audio only

V – the communication channel with video only

AV – the communication channel with both audio and video

Based on the analysis of the annotated PAD values, we try to find some disciplines of human emotion perception on talking agent by different communication modalities. For each communication modality, we computed the mean value and the standard deviation of PAD for each emotion scenario. It seems that the PAD values of some emotion scenarios have large standard deviations for either audio or video only (the largest value is 0.67). We think this is because people may have perception confusion about some emotions only by listening to the audio or watching the video. For example, it is hard to distinguish fear and anxiety only by audio, so the labelers may give PAD values to these two kinds of emotional speeches in a large range. But for audio-visual emotional speeches, the three labelers give more similar PAD values for an emotion scenario (the largest standard deviation is just 0.19). In order to analyze the reinforcing relationship between audio and video in emotion perception visually, we draw an ellipsoid to describe the annotated PAD values in the PAD 3-D space for each emotion scenario, in A/V/AV modality respectively. The average PAD values were used as the center while the standard deviations were used as the radius. The overlap ratios of the ellipsoids of every two emotions were examined, which indicates the perception confusion between two emotions.

- *Audio* v.s. *Audio-Visual*: we first examined the PAD values of every two emotions annotated through A and AV channels. The results show that sometimes it is hard to tell the accurate emotion only by listening to speeches, such as surprise and happiness, fear and sorrow. But with the help of proper facial expressions, it is much easier to differentiate the emotions.

- *Video* v.s. *Audio-Visual*: the same situation happened between V and AV channels. The results show that people may have perception confusion only by only watching the video, such as neutral and relax; submissiveness and fear. But with the help of emotional speech, people can understand the real attitude or intention of the speaker.

- *Audio* v.s. *Video* v.s *Audio-Visual*: the situation between A and V modalities is more complex. The results show that A or V modality can achieve a better perception accuracy than the other for different emotion. For example, facial expression is more reliable for happiness and surprise;

while for submissiveness and fear, speech is more helpful. In most cases, the overlap ratio of the ellipsoids of every two emotions in modality AV is much smaller than either modality A or modality V. That means the emotional audio-visual speech can help people understand the speaker's real emotion.

To further confirm the reinforcing relationship between audio and video modality in emotion perception, we also try to quantitatively describe the confusion of emotions perceived by different communication channels. For each sample point in PAD space, let $d_1$ be the Euclidean distance between it and the centroid of the emotion scenario it belongs to; Let $d_2$ be the Euclidean distance between it and the centroid of another emotion scenario. If $d_1 \geq d_2$, this PAD sample is defined as a confused sample of the two emotion scenarios.

TABLE II. THE MEAN VALUE AND STANDARD DEVIATION OF THE PERPLEXITY OF EMOTION SCENARIOS BY A, V AND AV ON D.

| Modality | Mean Value | Standard Deviation |
|---|---|---|
| Audio | 15.25% | 14.41% |
| Video | 17.23% | 12.11% |
| Audio + Video | 10.86% | 11.25% |

The perplexity of two emotion scenarios is calculated as the number of confused samples divided by the total sample number of the two emotions. The results are shown in Table II. It indicates that AV modality achieves a much lower perplexity between emotion scenarios. On the other hand, A and V modalities have almost the same ability on emotion perception.

### C. Subjective evaluation on emotion perception

Since the analysis in the above sub-section unveils the mutually reinforcing relationship between audio and video modality in emotion perception, it is interesting to investigate how accurate the emotional information is conveyed by our talking agent. Do the '*listeners*' feel the multimodal interface helpful in improving their comprehension of emotion? In this study, we designed a one-way computer-mediated communication scenario, where the 3-D talking agent tried to express different emotions through speech and facial display, and the "*listeners*" tried to understand the real emotion of the talking agent by listening to the speech (via A channel), watching the mute agent (via V channel) or watching the audio-visual speech (via AV channel). The texts of the speeches are meaningless, without any emotion tendency.

- *Materials and Participants*

We synthesized 50 audio-visual speeches on 3-D talking agent using the PAD values of 11 emotions with our proposed system. And 4 or 5 PAD values were selected from each emotion scenario. We invited 15 participants as our listeners, 6 females and 9 males. The average age of them is 25.2 (standard deviation is 2.1). All of them are graduate students in Tsinghua University.

- *Measurements*

*Performance* was measured as comprehension accuracy, that is, how many emotional speeches were recognized correctly (same as the original scenario), in percentage.

*Confidence* measured the level of confidence in the correctness of their answers. The participants, after submitting each of their answers, were posed the following question: "to what extent are you confident the talking agent express the emotional state as you chose?" A MOS (Mean Opinion Score) evaluation is performed with scores from 1 to 5: 5 - very sure; 4 - sure; 3 - not quite; 2 - not; 1 - I don't know at all;

*User experience* of the emotional audio-visual interface (i.e. talking agent) was assessed by examining how the participants responded to the following questions: "Do you think the audio-visual interface is helpful to your comprehension of emotional scenario?" The helpfulness is scored by a 5-point scale: 5 - very helpful, 4 - helpful, 3 - fair, 2 - little help, 1 - confused help.

- *Results*

*Performance and Confidence*: the performance and confidence of emotion perception by audio, visual, and audio-visual interfaces are shown in Figure 3. A repeated measures ANOVA was used to analyze the data. The audio-visual interface obtains the highest score (70.0% and 3.88) than audio or visual both in performance and confidence ($F[2,297]=32.46$ $p=1.134e\text{-}13<0.05$). Especially, the audio-visual interface has a much higher perception performance than audio, which indicates that the facial expression has a great reinforcement effect on speech for emotion perception.

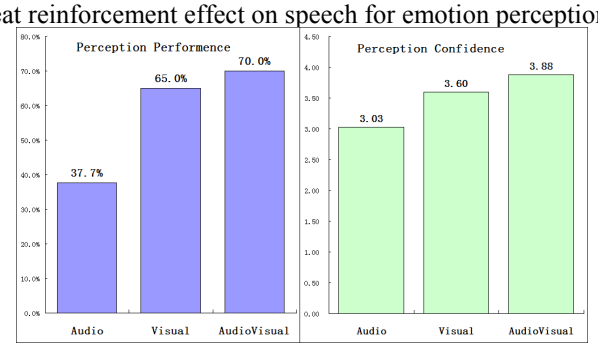



Figure 3. Perception performance and the confidence in different conditions.

Besides the main effects, we also noticed that video plays an important role in this experiment on talking agent. We think the reason is that the synthetic emotional speech can not convey as much information (emotion, attitude or intention) as natural speech, due to the limitation of nowadays speech synthesis techniques. Therefore, for talking agent, the emotion expression benefits from the facial movement much more. This also further makes clear the necessity of the multimodal interface in human computer speech communications.

- *User Experience*

In the experiment, participants also reported how they felt about the helpfulness of our audio-visual interface in emotion perception, besides the objective perception performance. The average user experience score is 4.39, and the standard deviation is 0.55. These results show the satisfaction of the participants for their perceiving and understanding the emotion information through AV channel.

## IV. CONCLUSIONS AND FUTURE WORK

This paper introduced an emotional audio-visual speech synthesis method on 3D talking agent. A unified model for

emotional speech conversion using GMM is proposed and the facial expression is synthesized simultaneously. The emotional audio-visual speech is synthesized in continuous emotion space. We conducted a series experiments to give a full-scale evaluation on the proposed emotional talking agent. The analysis of the experimental results reveals that human emotion perception on virtual talking agent has the discipline in term of modality. People may have serious confusion or misunderstanding about the affective meanings only by listening to the speech or watching the facial movements. But in audio-visual condition, there exists the reinforcing relationship between audio and video in emotion perception. The experimental results strongly demonstrate that the audio-visual speech can achieve better emotion perception accuracy than only either speech or video. Besides the improved performance and confidence, the participants also reported their satisfaction with the multimodal interface. That means multimodal interface can help people understand the speaker's real meanings with a pleasant user experience.

Another interesting finding is that video is the more critical channel on emotion perception on virtual talking agent. The reason could be that nowadays the speech synthesis and conversion techniques have their own limitation to convey affective meanings as exactly as natural speech. This result is consistent with the previous conclusion which has indicated the necessity of the multimodal interface in human computer speech communications.

The study also demonstrates the effectiveness of our PAD based emotional talking agent synthesis system. The performance and confidence of emotion perception by audio-visual interface are 70.0% and 3.88 (in 5 point scale), which show the accuracy of the emotional information conveyed by the talking agent.

Based on the observation of the average performance of our multimodal speech communication interface, our future work will focus on the improvement of the emotional speech synthesis/conversion algorithm, especially the investigation on speech spectrum features that could describe the affective meanings besides emotion more properly.

## REFERENCES


[1] Schröder, M., Emotional Speech Synthesis:A Review, In *Proc.Eurospeech* 2001, Aalborg, 561-564.

[2] Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., et al. Parameterized facial expression synthesis based on MPEG-4. *EURASIP Journal on Applied Signal Processing*. 10(2002), 1021–1038.

[3] JingJing, http://sepc495.se.cuhk.edu.hk/crystal/

[4] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology: Developmental, Learning, Personality, Social*, vol. 14, no. 4, 261–292, Dec. 1996.

[5] Zeng, Z.H, Pantic,M., Roisman, G.I., et al. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. Pattern Analysis Mach. Intell*. 31(2009), 39-58.

[6] Abelin, A. Seeing Glee but Hearing Fear? Emotional McGurk Effect in Swedish. In *Proc*. *4th Conf. on Speech Prosody*, 2008, 713-716.

[7] Li, X.M., Zhou, H.T., Song, S.Z., et al. The Reliability and Validity of the Chinese Version of Abbreviated PAD Emotion Scales. In *Proc. Int. Conf. Affective Computing Intelligent Interaction*, 2005, 513-518.

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

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

[10] Zhang, S., Wu, Z. Y., Meng, H. M., Cai, L. H., "Facial Expression Synthesis using PAD Emotional Parameters for a Chinese Expressive Avatar", in *Proc. Int. Conf. Affective Computing Intelligent Interaction*, 2007, 24–35.

[11] Z. Wang, L. Cai, and H. AI, "A dynamic viseme model for personalizing a talking head," in *Proc. 6th Int. Conf. Signal Process.*, Beijing, China, Aug. 2002, pp. 26–30.

[12] Y. Wang, Z. Wu, L. Cai, and H. M. Meng, "Modeling the synchrony between audio and visual modalities for speaker identification," in *Proc. 8th China Phonetic Conf. & Int. Symp. Phonetic Frontiers*, Beijing, China, 2008.