

Emotional head motion predicting from prosodic and linguistic features

Minghao Yang¹ · Jinlin Jiang² · Jianhua Tao¹ · Kaihui Mu¹ · Hao Li¹

Received: 30 March 2015 / Revised: 1 November 2015 / Accepted: 26 February 2016 /
Published online: 3 March 2016
© Springer Science+Business Media New York 2016

Abstract Emotional head motion plays an important role in human-computer interaction (HCI), which is one of the important factors to improve users' experience in HCI. However, it is still not clear how head motions are influenced by speech features in different emotion states. In this study, we aim to construct a bimodal mapping model from speech to head motions, and try to discover what kinds of prosodic and linguistic features have the most significant influence on emotional head motions. A two-layer clustering schema is introduced to obtain reliable clusters from head motion parameters. With these clusters, an emotion related speech to head gesture mapping model is constructed by a Classification and Regression Tree (CART). Based on the statistic results of CART, a systematical statistic map of the relationship between speech features (including prosodic and linguistic features) and head gestures is presented. The map reveals the features which have the most significant influence on head motions in long or short utterances. We also make an analysis on how linguistic features contribute to different emotional expressions. The discussions in this work provide important references for realistic animation of speech driven talking-head or avatar.

✉ Minghao Yang
mhyang@nlpr.ia.ac.cn

Jinlin Jiang
jiangjinlin2014@163.com

Jianhua Tao
jhtao@nlpr.ia.ac.cn

Kaihui Mu
khmu@nlpr.ia.ac.cn

Hao Li
hli@nlpr.ia.ac.cn

¹ The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

² School of International Studies, University of International Business and Economics, Beijing, China

Keywords Visual prosody · Head gesture · Prosody clustering

1 Introduction

The topic of multimodal human-computer-interaction (HCI) has recently gained great insights by a growing number of researchers from speech, computer graphics and vision areas. Many interesting work about audio-visual fusion/mapping has been proposed for multimodal information processing. For instance, speech based facial animation [1, 4, 5, 18, 21, 29], speech based lips movement [5, 19, 45], and audio-visual based emotion recognition [14, 37, 49, 50]. There are also some work for head motions [7, 10, 12, 15] and body gestures [2, 13, 23, 25, 40, 44], however, most of them just focused on the gesture animation and recognition. Some researchers have tried to link the speech to head motions. For instance, Graf applied pitch accent and phrase boundary labeled by the ToBI prosody classification scheme to synthesize new head motions [18]; Zhang created head movements by using the expressivity of prosodic word measured by the pleasure-arousal-dominance (PAD) model [51]; Zhou used a dynamic unit selection method to synthesize head motion sequences from an audio-visual database [30]. Even with these work, we are still far from discovering the relationship between speech representation and emotional head motions. In human face-to-face conversation, someone talks in nod and smile only when he (she) is in good feelings. And there are some other persons, who always converse in good manner even they say “no” in bad feelings. The emotional head motions are influenced by different emotion states and reflect a person’s character. If a virtual agent owns his (her) character emotional head motions, it will improve users’ experience in human-computer interaction. In this work, we first make a detailed analysis on the relationship of speech representation and head gestures and try to find which speech features have the most significant influence on head movements in long or short utterances. Finally, we construct a bimodal mapping model from speech representation to head gesture for different emotion states.

We first construct an audio-visual database which covers six basic emotional states. In order to obtain reliable visual features, we propose a two-layer clustering scheme to classify elementary head gesture patterns. Head gesture patterns are the cluster results obtained from hierarchical clustering of head motions. With these clusters, a speech to head gesture mapping model is constructed by a Classification and Regression Tree (CART). After the training of CART, we also get a systematical statistic map based on the statistic results of CART. The map shows the relationship between the speech features (including prosodic and linguistic features) and head gestures, and reveals which features have the most significant influence on head movements. We also discuss what kinds of prosodic and linguistic features have greater influence on emotional head motions. And the parameters which have higher subjective evaluation scores are adopted in the animation model. The total framework of the proposed method is presented in Fig. 1.

The remaining parts of this work are organized as follows: the related work are presented in section 2; the selected prosodic and linguistic features used for bimodal mapping model are introduced in section 3; the proposed two-layered clustering method and the mapping model between speech and head gestures are discussed in section 4 and 5 respectively; experiments and how our predicting model is used to determine mapping parameter will be introduced in section 6, and the influence of prosodic and linguistic features on emotional head motions are also discussed in detail in this section; conclusions are drawn in section 7.

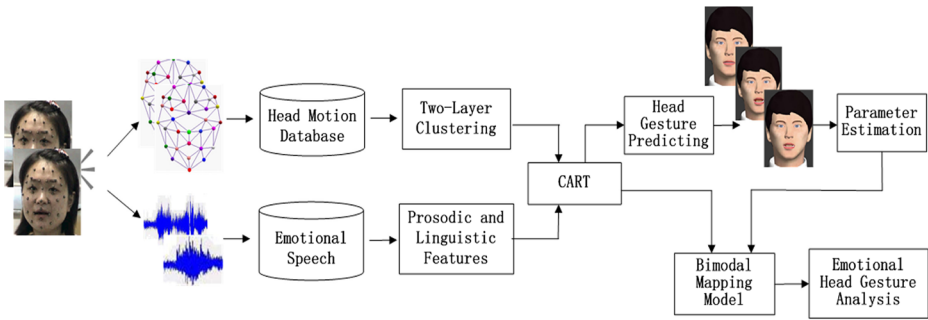


Fig. 1 Flow chart of the bimodal mapping model between speech and emotional head motions

2 Related work

Since the work of Parke in 1972 [31], remarkable progress has been achieved to generate realistic facial animation in multi-modal HCI communication. There are two traditional techniques in building the mapping model from speech prosody and linguistic features to gestures: physically based modeling [8, 11, 16, 38, 42] and statistics based mapping [5, 11, 12, 15, 19, 21, 32, 33, 44–46, 52].

Physically based modeling was used early in speech driven facial animation, which included lip, jaw, tongue movement and emotion expression. It presented the organ shape change and revealed the corresponding physical movement of organs with speech generation. As a pioneering work for real-time 3D facial animation, Keith Waters creatively proposed a muscle model for animating 3D facial expression [42]. His work produced many research work on 3D facial expression [8, 16] and talking agent [11, 28, 38]. A key frame animation character was built through a controller driven by physics model in [16]. In this method, every key frame was designed manually and facial animation was driven by a number of springs. In the last decade, more work focused on building the relationship between visual channels and speech signals. Finite element method was adopted in [22] to analyze how the tongue, lip and jaw physical movement parameters (speed, acceleration, force and power from muscle) influence the vowel generation. New lip muscle model was proposed to drive Chinese 3D talking-head [30] and generate realistic facial animation [43].

Statistics based mapping was often used to generate real-time facial and gesture animation. Cluster based mosaicking has been successfully adopted to synthesize emotion expression in multimodal communication system [1, 39]. Recently, machine learning methods, including hidden markov model (HMM), neural network (NN) [12], gaussian mixture model (GMM), and HMM based mosaicking schema have been proposed to build mapping model for multi-modal channels. Since Yamamoto adopted HMM to drive lip movement [45], HMM based statistics methods have achieved impressive progress in talking face [26, 27], emotional 3D talking-head [5, 21], articulatory movements predicting [52], etc. As an effective classified and mapping schema, neural network was adopted by Richard to generate speech-synchronized animation based on key-frame selection and deformation technique [33]. Deep neural network was also demonstrated an effective method for talking-head animation [12]. Boosting-Gaussian mixture model was proposed as a unified model for emotional speech conversion in [21].

Pose and gesture contributed to impressive expression and nature communication. As to pose animation generation, the early methods usually transferred random or predefined actions to virtual agent [4]. Since late 1990s, with the increase of resolution and price reduction of

motion capture devices, actors' actions could be effectively mapped to virtual human [9, 16, 38]. Recently, extensible markup language (XML) has been often used to describe the relationship between agent's motion and speech in HCI multimodal system [15, 24, 25, 29, 40]. SEMAINE project learned persons' real emotion from face-to-face chat video recorded by multiples high resolution cameras, and mapping the recorded emotion to virtual agent [6, 32]. Martin adopted XML technique to control a 'realistic' 3D agent "Greta", who could communicate complex information through the combination of verbal and nonverbal signals [15, 29]. "Max" adopted motion planning technique to drive the agent lips, gesture and body movement synchronously [25]. With these methods, body animation, head gestures were planned by rules or generated by transferring actors' actions to virtual agent. The inherent relationships between visual and audio channels were not discussed and how the speech features influence emotional expression still remains unclear.

Head gesture predicting was an important part of visual prosody. The relationship between head motion and speech is closer than that of hand and body gesture with speech features. More reliable clusters could be obtained from head motions than from facial animations and body movement. In this work, we aim to study the inherent relationship by building the mapping model between head motions and Chinese speech prosody and linguistic features.

3 Selected prosodic and linguistic features

Six kinds of prosodic and linguistic features are effective for natural style speech synthesis [3, 17, 34]. We adopt similar prosodic and linguistic features in constructing speech drive emotion head motion in this work. These features are the stress point (S), the boundary type (B) from the prosody, the position of a syllable in a sentence (PL), tone type in Mandarin Chinese (T), part-of-speech of word (POS) and the length of the prosodic word (LW) from linguistic features. These features could be automatically predicted from text stream [3] by a text analysis module of a Chinese text-to-speech (TTS) system [17].

All stress points are classified into three types: soft ("1"), normal ("2") and stress ("3") and each stress (S) point is marked with syllable. The boundary type describes the length of mute among prosody, which is marked by prosodic words ("."), prosodic phrase ("|") and intonational phrase ("\$") based on their mute length. Head gesture is usually fixed at the beginning point or ending point of a sentence and the length of all sentences are normalized as unit length. Tone type (T) is an important feature in Chinese pronunciation, which presents the stress in a syllable. Tone type (T) contains five types, including "level tone (1)", "rising tone (2)", "falling-rising tone (3)", "falling tone (4)" and "light tone (5)". These five stone types are marked by number 1–5 respectively. Part-of-speech (POS) sometimes presents accent in Chinese pronunciation, for example, adverbs in exclamatory sentence, interrogative words in interrogative sentences [20]. Based on the part-of-speech tagging presented by [35, 36], all the POS are marked by number 0–43 in our method (Table 1). Finally, position of syllables (PL) is determined by length of syllable.

Table 2 lists an example of all the text features in Chinese Pinyin transcript for the sentence 'He is friendly to people' ("ta dai ren ke hao le" in Chinese). Based on the definition for Chinese prosodic and linguistic features, all the features are organized in the syllable level and putted together as a vector. Note that we assume that the head motion is consistent in one syllable, which may smooth some sharp head motions. This assumption does not influence the total trend of head motions.

Table 1 Part-of-speech tagging for Chinese text adopted in our method

Id	Name	Id	Name	Id	Name	Id	Name
0	a adjective	11	g morpheme	22	ns address name	33	tg time word morpheme
1	ad Adjective as adverbial	12	h prefix	23	nt organization name	34	u auxiliary
2	ag adjective morpheme	13	i idiom	24	nx nominal character string	35	v verb
3	an adjective with nominal function	14	j abbreviation	25	nz other proper noun	36	vd verb as adverbial
4	b non-predicate adjective	15	k suffix	26	o onomatopoeia	37	vg verb morpheme
5	bg category morpheme	16	l fixed expressions	27	p preposition	38	vn verb with nominal function
6	c conjunction	17	m numeral	28	q classifier	39	w symbol and non-sentential punctuation
7	d adverb	18	mg numeric morpheme	29	r pronoun	40	x unclassified items
8	dg adverb morpheme	19	n common noun	30	rg pronoun morpheme	41	y modal particle
9	e interjection	20	ng noun morpheme	31	s space word	42	yg mood morpheme
10	f directional locality	21	nr personal name	31	t time word	43	z descriptive

4 Two-layered clustering of head motion features

4.1 Extraction of Euler angles and translations

In our bimodal database (how the database is built will be introduced in section 6.1), the trajectories of the 32 markers contains all of the head gesture information (Fig. 5a and b). In the work, the rigid head motion features, Euler angles and translations, are extracted from those trajectories in three axes. The translations of head movements are measured as the translations of the marker placed at nose. The rotation matrix is computed by the method proposed in [10]. The

Table 2 Samples of Chinese prosodic and linguistic features in the sentence “ta dai ren ke hao le” (He is friendly to people)

Text's features	Samples
Stress (S)	(ta)/2 (dai ren)/2 (ke hao le)/3
Boundary type (B)	(ta) (dai ren) (ke hao le)\$
Length of prosodic word (LW)	(ta)/1 (dai ren)/2 (ke hao le)/3
Tone type (T)	Ta1 dai4 ren2 ke2 hao3 le5
Part-of-speech (POS)	(ta)/r (dai ren)/v (ke hao)/d (le)/y
Position of syllables (PL)	ta/1/6 dai/2/6 ren/3/6 ke/4/6 hao/5/6 le/1

three Euler angles of rotation, indicated by e_x , e_y and e_z , are obtained from the rotation matrix around three axes. R_{ij} indicates the item in rotation matrix which is placed at row i and column j .

$$e_z = \tan^{-1} \left(\frac{R_{21}}{R_{11}} \right) \tag{1}$$

$$e_y = \frac{\tan^{-1}(-R_{31})}{(R_{11} \cos e_z + R_{21} \sin e_z)} \tag{2}$$

$$e_x = \frac{\tan^{-1}(R_{13} \sin e_z - R_{23} \cos e_z)}{(R_{22} \cos e_z - R_{12} \sin e_z)} \tag{3}$$

The Euler angles and translations combine to form a vector containing 6 static motion parameters for each frame. The trajectories of three Euler angles for the sentence ‘‘He is friendly to people’’ are shown in Fig. 2. In addition to the six static motion parameters, the temporal dynamics parameters obtained from the first and second derivatives are integrated together into 18 dimensions head gesture vector. All the 18 dimensions vectors cluster and shape into the head gesture vector. The boundaries of each syllable in a sentence are aligned with HTK [48] tools based on the acoustic speech recorded by the audio recording device. These boundaries are utilized to help find the available head gesture frames in the synchronized head motions.

4.2 K-means clustering of the head gesture

We use a two-layer clustering method to find the elementary head gesture patterns from the 18 dimensional rigid head motion features. At the first step of our two-layer clustering method, K-Means clustering is used to extract the basic head gesture patterns for different emotional states.

To determine how similar two samples are, two kinds of similarity metrics, Euclidean distance and Pearson correlation, are usually utilized to compute similarity scores. Their mathematical expressions are given as Eqs. (4) and (5): Here, x_i and y_i are the items from the 18 dimensional vectors we obtained in section 4.1. The denotation r_e and r_p is the Euclidean and Pearson similarity score for these two vectors. Higher value of r_e and r_p indicates two vectors being more similar. The Pearson correlation coefficient r_p is a measure of how well two sets of data fit on a straight line, which gives better results in the situations when the data are not well normalized.

$$r_e = 1 / \left(1 + \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \right) \tag{4}$$

$$r_p = \max \left(0.05, \left(\left(\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right) / \sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \sum_{i=1}^n x_i \right)^2 * \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \sum_{i=1}^n y_i \right)^2} \right) \right) \tag{5}$$

In our experiments, the number of K is set big enough and the initialized K vectors are randomly distributed in space, which makes all head gestures be covered. The

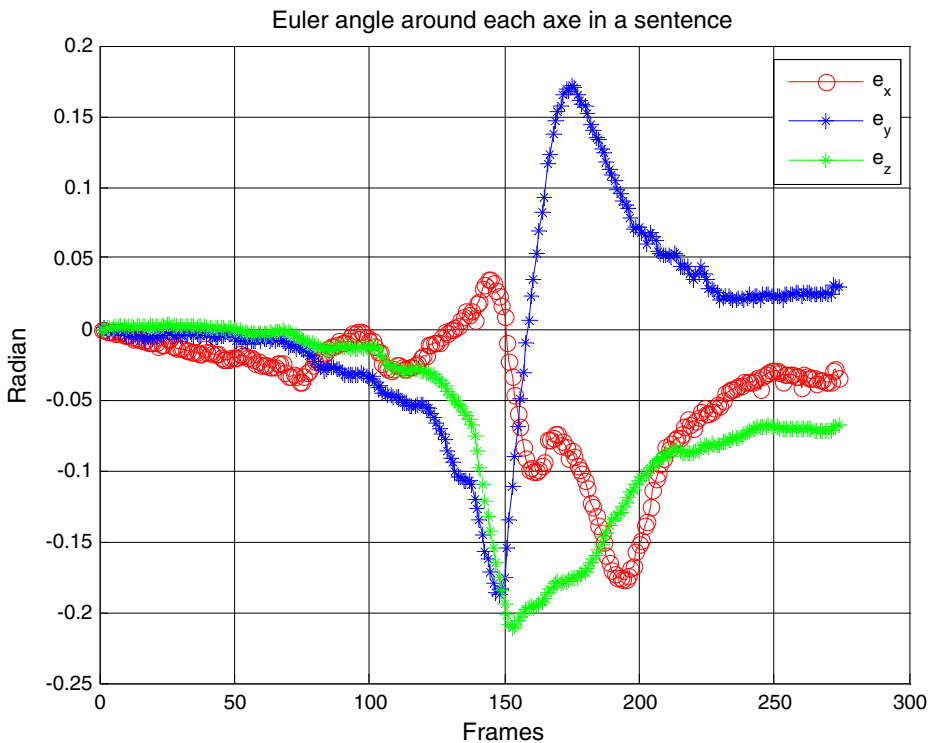


Fig. 2 Trajectories of three Euler angles around three axes for the sentence “He is friendly to people” in happiness state

Pearson correlation coefficient is adopted to determine vectors’ distance during cluster procedure. While the different initialing values of K lead to different clustering results and the variable positions of initialing vectors result in different clusters boundaries, K -means cluster makes it difficulty to obtain reliable elementary patterns at this step. We further adopt a hierarchical clustering method to obtain more reliable cluster boundaries in our two-layer clustering method.

4.3 Hierarchical cluster of the head gesture

A hierarchical clustering is used to determine what the initializing K -means clustering could be merged. Figure 3 presents the head gesture hierarchical clustering results, where the threshold α is valued as α_1 and α_2 ($\alpha_1 < \alpha_2$) respectively in happiness state. In Fig. 3, the numbers in circles t are the initiating cluster index obtained from K -means cluster. The nodes with the same color belong to the same clustering. Figure 3a and b presents the clustering results for hierarchical clusters when $\alpha = 0.1$ and $\alpha = 0.2$ respectively. If the value of α is bigger, then some clusters merge into a new cluster. For example, when the value of threshold is α_1 , the nodes (1) and (12, 19) belong to two different clustering. And when α is α_2 , these three nodes are merged into one bigger clustering (1, 12, 19). Hierarchical cluster makes some adjacent groups keep separately or merge into a new clustering by different thresholds for initiating K -means clustering.

Since the variable values of α lead to different head gesture clustering, it is necessary to find an automatic method to calculate the value of α or to determine the range of α for different emotional states. We further propose a distance histogram schema to calculate α automatically from the initiating K-means clustering.

For each pair of the K clusters, supposing that D (D is calculated as $1/r_p$) presents the distance between two cluster centers, we obtain $K*(K-1)/2$ distances. Let $D_{max}=\max(D_{ij})$, $D_{min}=\min(D_{ij})$, where $i,j \in (0,1,2,\dots,K)$ and $(i \neq j)$, then each pair falls in a Bin, which index is r. We denote it as $Bin(r)$ ($r \in (0,1,2,\dots,K)$) and every Bin's range could be calculated as Eq.(6).

$$D_{Bin} = (D_{max}-D_{min})/K \tag{6}$$

Distance histogram presents an intuitionistic view for data distribution, which is further used to determine the value of α (7), where function $H(Bin(l))$ presents the hit numbers that how many pairs are located in $Bin(l)$ according to their distance. λ is a coefficient related to different emotional states (how the value of λ is estimated will be introduced in section 6.2).

$$\alpha = D_{min} + \lambda \cdot I_{loc} \cdot D_{Bin}, \text{ where } I_{loc} = \frac{K}{l=1} \arg(\max(H(Bin(l)))) \tag{7}$$

Figure 4 presents the initiating K-means clustering distance histogram for the head gesture in happiness emotional state. The max value of ratio (total hit numbers of each Bin) is normalized as 1.0. In happiness emotional state, we can see that the max ratio is located at $Bin(2)$, and $D_{max}=2.0$, $D_{min}=0.100353$ ($D_{Bin}=0.038$), finally we obtain: $\alpha=0.100353 + \lambda * 2 * 0.038$. And α is used to determine what the initializing K-means clustering could be merged automatically.

5 Mapping model from prosodic and linguistic features to head gestures

After the reliable head gesture patterns are obtained by the proposed two-layered cluster method, the mapping model from prosodic and linguistic features to head gestures

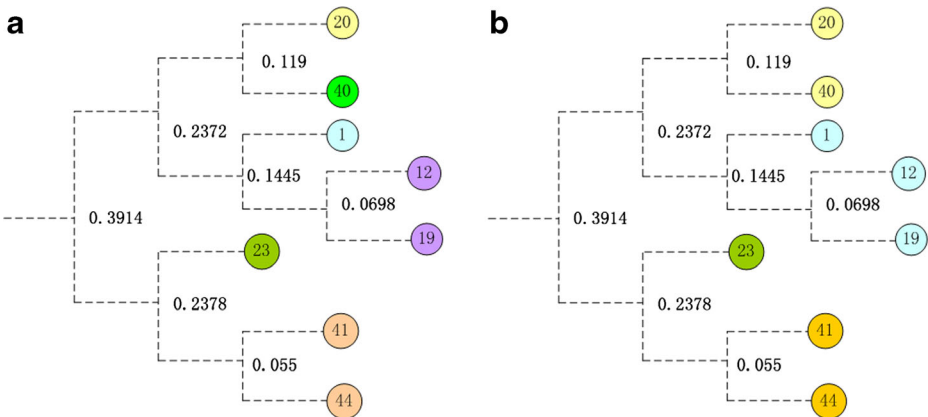


Fig. 3 Hierarchical clustering results of the head gesture for α_1 (=0.1) (a) and α_2 (=0.2) (b) respectively (In happiness emotional state), where α is the upper distance boundary for clustering and the circles in same color could be merged into one cluster

could be easily constructed. Traditional text or speech driven head motion systemization methods mainly focus on facial animation from text or speech [18, 30, 51]. With these methods, the relationships between prosody, linguistic features and head gesture were not discussed.

CART is able to select the head gesture pattern for a input textual vector in time cost $O(n)$. In the training model of CART, the textual features, which have outstanding influence on head gesture classification, are located at the nodes closer to the root of CART. In this way, CART has the ability to explain and interpret which linguistic or prosodic features have greater influence on head gesture generation.

In our method, the available frames of head gesture features in each sentence are assigned to the nearest group, which are used as the training parameters as the synchronized prosodic and linguistic features for CART modeling (<http://salford-systems.com/cart.php>). Gini Impurity (8) and entropy (9) are used in our method to determine a prosodic and linguistic feature which should be move to a lower or higher layer of the CART tree [41].

$$I_G = \sum_{i=1}^n \left(p(x_i) \cdot \sum_{j=1, j \neq i}^n p(x_j) \right) \tag{8}$$

$$I_E = - \sum_{i=1}^n (p(x_i) \log_2 p(x_i)) \tag{9}$$

Given a classification for certain linguistic or prosodic feature, all elements could be divided into two categories: A and A^c . In Eqs. (8) and (9), $p(x_i)$ is the probability that a vector x_i belongs to A and $\sum_{j=1, j \neq i}^n p(x_j)$ is the sum of probability that all the other vectors $x_j (j \neq i)$ belong to A. The bigger I_G or I_E is, the more ambiguous this classification is. Then current linguistic or prosodic feature has less influence on head gesture. It should be moved to deeper layer of the tree in the CART model.

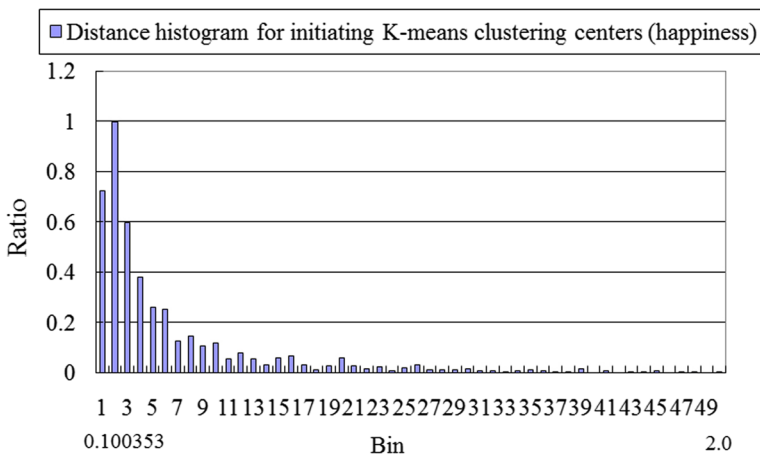


Fig. 4 The initiating K-means clustering distance histogram for the head gesture in happiness emotional state

6 Experiments and analysis

6.1 Multimodal data collection and processing

The multimodal database used in this work is recorded by an actress (Fig 5a), with 32 markers on her face (Fig 5b). The colloquial, sentences of different length are selected from interrogative and exclamatory sentences. Totally 489 sentences are selected overall the database. Each sentence is spoken in neutral and other five emotional states, including anger, fear, happiness, sadness and surprise, which are selected according to the meaning of the sentences and spoken mimicking that situation (details please refer to Table 3). The actress was asked to perform the same as usual. Having only one performer in our bimodal database, it is easy to maintain the style and personality in each emotional state. It is also convenient to cluster the head gesture patterns in each emotional state.

The motion analysis system is utilized to capture the movements of a typical marker set placed on the actress's face. An audio recording device is applied to record the speech spoken by the actress. In the recording, a particular voice sentinel is located to synchronize two acoustic signals, one recorded by the audio recording device and the other from the AVI file captured by the motion analysis system. Since the AVI file is synchronized with the motion of the marker set, the acoustic speech recording by the audio recording device also synchronizes with the movements of the marker set. Using the methods provided by [47], the syllable features could be obtained from acoustic signals in HTS, and we use the syllable to align the text features and speech features automatically. In this way, speech features could be accordingly coordinated with corresponding visual features.

6.2 Hierarchical clustering

Two-layer cluster method is adopted in our method to obtain reliable clusters for head gesture. At the first layer, K clusterings are obtained by K-means cluster method. In this step, the value of K is set big enough so that every clustering could be organized compactly. In our experiments, the value of K is set to 50 for each emotional state.

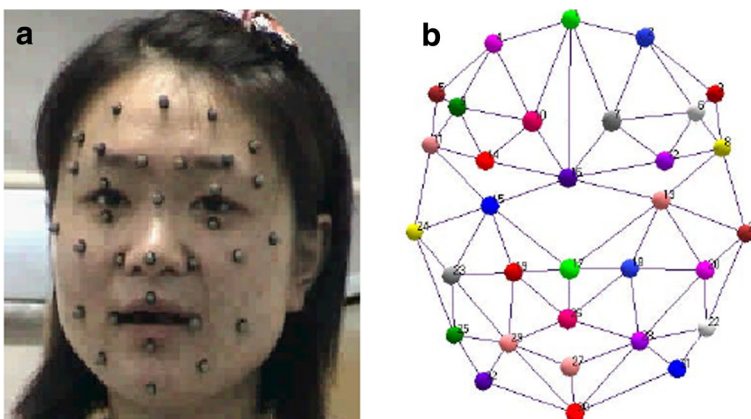


Fig. 5 **a** All 32 markers placed on actress' face; and **b** markers' corresponding position in 3D space

Table 3 The number of sentence recorded for each emotional state in our bimodal database

Emotional states	The number of sentences recorded
Neutral	489
Anger	245
Fear	157
Happiness	165
Sadness	57
Surprise	103

At the second layer, distance histogram is used to determine the threshold that some initiating clusters should gather together. Figure 6 presents the distance histogram for all emotional states in our experiments according to the subsection 4.3 (as the dominant Bins are located at the previous sections of distance histogram, we only list the former 15 Bins in Fig. 6). It could be seen from Fig. 6 that most ratio values are located at Bin(2) in most emotional states, except in sadness emotional state. Based on the discussion in subsection 4.3, the values of r_{loc} and D_{Bin} could be automatically estimated according to the distance histogram.

The value of threshold α is determined by the values of l_{loc} , D_{Bin} and λ . λ is an empirical parameter, which is hardly estimated from distance histogram automatically. To overcome this problem, we empirically set the value of λ as the following four values: $\lambda_1=0.5$, $\lambda_2=1.0$, $\lambda_3=2.0$ and $\lambda_4=3.0$. Then we obtain four mapping models $M_j(j \in (1, 2, 3, 4))$ for each emotional state respectively. Subjective evaluation (MOS) is used to score the animation sequences generated by our head motion synthesis system (section 6.3.1) for these four models. Then the final value of λ is obtained according to the highest value of models (Eq. (10)). Function $MOS(\cdot)$ in Eq. (10) means the subjective evaluation score for the model M_j .

$$j = \max \left(\arg_{j=1}^4 (MOS(M_j)) \right) \quad (10)$$

According to the scores from MOS (Fig. 8), in most cases, when the value of λ is 1.0 or 2.0, the head motion synthesis animations are more nature than the situations when λ is valued 0.5 or 3.0. Then in our bimodal mapping model, we set the value of λ 1.0 and 2.0 respectively, and further analyze the relationship between acoustic signals and visual signals. Table 4 lists the final cluster results by our two-layer cluster method. We can see that the bigger value of λ leads to less clusters. In the following sections, we will further discuss the inherent relationship between prosody linguistic features and head gesture patterns for different λ .

6.3 Head motion synthesis and parameter evaluation by subjective evaluation

6.3.1 Head motion synthesis

As we have discussed in section 4.3, the value of λ is determined by the highest subjective evaluation for head motion synthesis. Four animations sequences are synthesized with different values of λ (or M_j) for the same text and the same emotional states. For each sentence, the

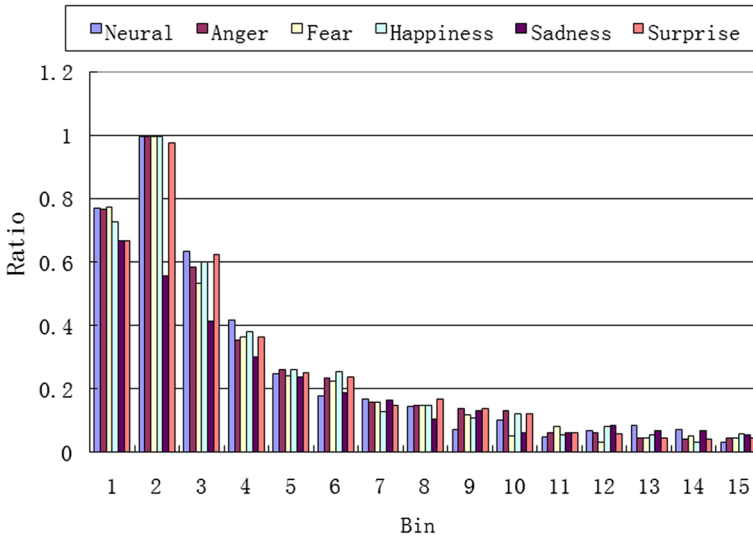


Fig. 6 The K-means initiating clustering distance histogram for the head gesture in different emotional states

subjects were asked to score the expressivity of visual speech on a five level mean opinion score (MOS) scale. The values of λ which correspond to the most realistic and nature head motion animation, are adopted for hierarchical clustering of the head gesture (section 4.3). In this section, we first simply introduce head motion synthesis method. How the values of λ are determined will be introduced in the following subsections.

In our experiments, the talking head could be driven either by texts (without corresponding speech coups) or speech (with corresponding text coups). For the cases with both speech and text, the emotional head motions could be generated from speech and text simultaneously by the model obtained in training procedure. While for the cases with only texts, the emotional head motions could be generated by texts and the speech features from TTS system. The corresponding CART model for that emotional state is selected to synthesize four new head motion sequences. The emotional state is assigned by users. Using these features as input, the CART model outputs new head motion sequences in syllable level. Then, spline interpolation is applied to generate the rest head motion sequence based on the head motion sequences coming from CART model. The size of frames in the sequence is determined by the duration predicted by text analysis

Table 4 The number of head gesture patterns for different emotional states under different threshold

Emotional states	λ	
	1.0	2.0
Neutral	27	17
Anger	25	15
Fear	29	15
Happiness	29	18
Sadness	27	16
Surprise	26	16

and frame rate used in animating the 3D talking-head. The following shows some of the synthesized curves smoothed by spline interpolation:

Figure 7 shows the synthesized curves of translations and Euler angles in happiness state on a sample sentence “The scenery is very beautiful” with $\lambda=2.0$. The scope of translation values is between -10 and 10 and the translation along with z-axis changes greatly displaying a forward and backward tilting in head motion. Euler angles are all within the scope of -0.05 to 0.05 . The synthesized head movements are rendered with a Chinese three-dimensional avatar using the head motion features smoothed by spline interpolation. In this work, only the Euler angles around three axes are utilized to animate the avatar which uses the FAP defined in MPEG-4 to drive the feature points in face model.

6.3.2 Parameter evaluation from subjective evaluation

20 subjects were asked to score the 20 sentences selected from six emotional states. Figure 8 presents the average scores of MOS for head gesture synthesis animation sequence. In Fig. 8, four sessions of head motions for these sentences are scored according to four scale values of λ . The number 1–4 presents the MOS score for the values of λ : 0.5, 1.0, 2.0 and 3.0 respectively.

The average MOS scores of four sessions are 3.3 (Session (1), $\lambda=0.5$), 3.7 (session (2), $\lambda=1.0$), 3.8 (session (3), $\lambda=2.0$) and 3.2 (Session (4), $\lambda=3.0$) respectively. When λ is valued by 1.0 and 2.0, the MOS scores are obviously higher than those of other two scales: 0.5 and 3.0. Most subjects consider that there are perceptible dithering problem in the sequences for $\lambda=0.5$. And when the value of λ is 3.0, the animation sequences are sensed a little stiff for some frames especially in surprise and anger emotional states. Most subjections are satisfactory with the results when $\lambda=1.0$ and $\lambda=2.0$.

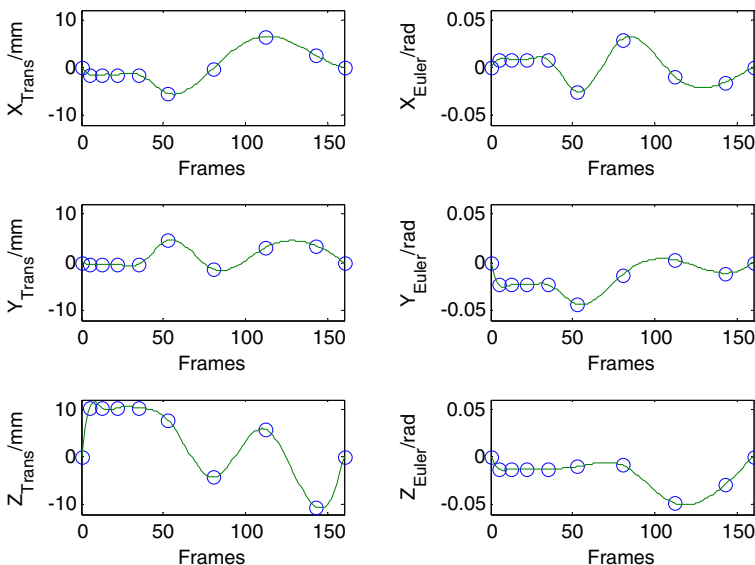
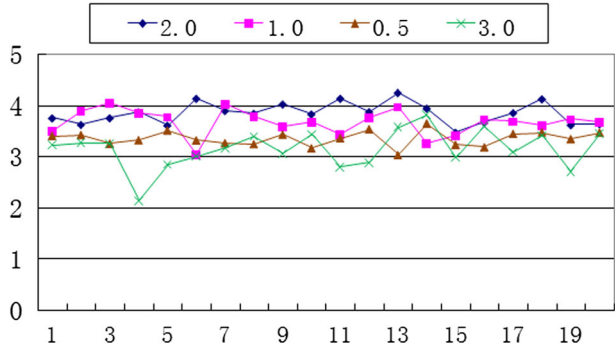


Fig. 7 Synthesized curves smoothed by spline interpolation in the sentence “Zhe Li De Jing Se Hao Mei A” (The scenery is very beautiful) for $\lambda=2.0$ (Happiness)

Fig. 8 Average MOS scores of 20 sentences for different λ (2.0, 1.0, 0.5 and 3.0)



We further compare the animation sequences under different emotional states for the situations that the values of λ are 1.0 and 2.0 respectively (Fig. 9). As is shown in Fig. 9, when the value of λ is 1.0, the animation sequences are more satisfactory in neural and happiness emotional states. On the contrary, the synthesis results are more nature in the negative emotional states (surprise, fear and anger) when $\lambda = 2.0$.

According to the comparison and analysis for animation sequences in different emotional states with different λ , the values of λ are set 1.0 and 2.0 in our two-layer cluster method finally.

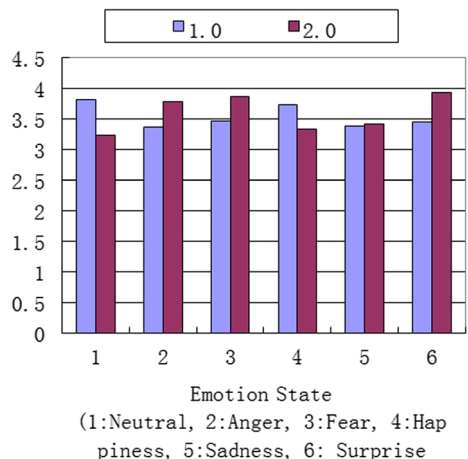
6.4 Evaluation of bimodal mapping model

6.4.1 Prediction accuracy evaluation

After the bimodal model has been constructed by CART, we use 10-fold cross validation to evaluate the prediction accuracy of our mapping model. Table 5 lists the average prediction accuracy for six emotional states when the values of λ are 1.0 and 2.0 respectively.

In Table 5, the predict accuracies are different in different emotional states. Particularly, the predict accuracy of neural emotional state is obviously lower than those

Fig. 9 Average MOS scores for the animation sequences under different emotional states with $\lambda = 1.0$ and $\lambda = 2.0$ respectively



of other states. It reflects that the head gesture patterns are more similar to each other in neural emotional state than those of other states. And the elementary head gestures are more difficult to classify in neural emotional state. When the value of λ is 2.0, the predict accuracies of fear and surprise states are above 70 %, which are obviously higher than those of other emotional states. A potential reason is that the actress has played exaggeratingly in fear and surprise states or her expression styles are naturally more abundant and impressive in these two states than those in other emotional states. In general, most emotional head gestures could be correctly predicted from prosodic and linguistic features by our proposed model.

We also can see from Table 5 that the prediction accuracy with $\lambda=2.0$ is higher than the situation with $\lambda=1.0$. The reason behind this phenomena is that when λ is bigger, some head gestures which are close to each other merge into a new cluster. The merger of these similar head gestures leads to the decrease of the number of head patterns. On the other hand, it contributes to the increase of the prediction accuracy to some extent. This also proves that the proposed two-layer cluster method could obtain reliable gesture patterns which is of benefit to obtain satisfactory head motion animation results (Fig. 8 in section 6.3.2), and to find the inherent relationship between speech prosody, linguistic features and head gestures.

6.4.2 Which prosodic and linguistic features have greater influence on head gesture

In the CART training model, the most important factors will be moved to the top layers of the tree. This means that a CART is useful not only for classification, but also for interpretation.

Tables 6 and 7 show the importance of each variable for every emotional state in long and short sentences respectively. In Table 6, the most important variable has the score of 100. The values of other variables drop down according to their different importance. The values in Table 7 indicate the importance of each variable in short utterances.

Based on the statistics results from CART, we further analyze which linguistic or prosodic features have greater influence on head gesture. The following conclusions are obtained from Tables 6 and 7:

- We can see from Table 6 that the length of prosodic word(LW) contributes mostly to head gesture classification. It reflects the physiological phenomena that the speaker's head

Table 5 Average prediction accuracy for 6 emotional states

Emotional state	Prediction accuracy for different λ	
	1.0	2.0
Neutral	57.9 %	63.2 %
Anger	62.0 %	68.2 %
Fear	64.4 %	71.2 %
Happiness	63.0 %	66.5 %
Sadness	64.1 %	68.9 %
Surprise	65.7 %	70.8 %

Table 6 Importance of each variable in long sentences for each emotional state

Emotional state	λ	Variable importance(%)					
		LW	POS	B	T	PL	S
Neutral	1.0	100	55.7	50.5	60.2	46.5	17.7
	2.0	100	55.8	46.7	63.8	42.5	30.9
Anger	1.0	100	72.8	52.6	71.4	46.5	27.8
	2.0	100	80.7	42.0	55.2	38.4	34.8
Fear	1.0	100	73.0	50.0	65.0	35.1	24.9
	2.0	100	63.4	46.5	56.8	39.9	20.4
Happiness	1.0	100	73.9	47.4	72.2	43.4	37.0
	2.0	100	86.0	51.7	71.6	40.0	17.4
Sadness	1.0	100	84.5	54.8	71.4	44.8	31.3
	2.0	100	77.0	63.4	75.4	50.4	47.0
Surprise	1.0	100	76.1	51.4	74.8	42.1	43.4
	2.0	100	88.4	41.8	65.7	31.6	52.3

gestures usually change widely from beginning to ending in a long sentence. A person’s head movement patterns usually has fixed model in certain emotional state for a long talking period.

- Apart from the influence caused by length of prosodic word(LW), the second important variable is the part-of-speech (POS). The reason is that there are totally 44 kinds of POS parameters, the number of which is obviously greater than that of other prosodic and linguistic features. Similarly, tone type(T) makes high ratio for generation head gestures, especially for anger, happiness and surprise state ($\lambda = 1.0$) and sadness states($\lambda = 2.0$). It presents that “POS” and “T” play important roles in emotion expression in Chinese communication.
- In Table 7, stress(S) brings notable fluency to head motions in fear and sadness states followed by position of syllable(PL). It presents that stress(S) often has greater influence on head gestures, especially in negative emotional states in short sentence. While in neural and happiness states, stress has less influence on head motions even in short sentences. It reveals that when speaker is in active or positive emotional states (neural and happiness), stress(S) is not easily observed from her or his

Table 7 Importance of each variable in short sentences for each emotional state

Emotional state	λ					
	1.0			2.0		
Neutral	LW	T	POS	LW	T	POS
Anger	LW	POS	S	LW	POS	T
Fear	LW	S	PL	LW	S	T
Happiness	LW	POS	PL	LW	POS	PL
Sadness	LW	S	PL	LW	S	PL
Surprise	PL	LW	S	LW	B	PL

pronunciation, especially in long sentences. While in Table 6, the least important variable is always stress. The reason is that there are only three kinds of parameters in stress feature. Therefore, in a long sentence, the influence caused by stress is relatively weaker than other prosodic and linguistic features.

- As shown in Table 7, position of syllable(PL) and boundary type(B) greatly influence the head gesture in surprise state. The reason is that head motions are often tightly related with interjection in Chinese pronunciation, especially in strong and sort exclamatory sentences. However, in Table 6, boundary type(B) does not contribute to head motion generation dramatically in surprise state, since the head movements caused by interjection are eliminated by other features such as POS and LW in long sentences.
- The parameter number in boundary type(B) is equal to that of stress(S). While boundary type exerts greater influence than that of stress in long sentences (Table 6). We can see from this comparison that pronunciation pause caused by boundary type(B) is also an important feature for head movement in long sentences.
- To make a conclusion, as to the relationship between prosody, linguistic features and the head motion sequences, the linguistic features (“LW”, “POS”, “PL” and “T”) often have greater influence than prosodic features (“S” and “B”) in long sentences. While in negative states (surprise, sadness and fear), prosodic features (“B”) are obvious more influential to head gestures in short sentences.

7 Conclusions

Studies on the relationship between speech prosody, linguistic features and head gesture contribute to the understanding of complicated multimodal communication. In this work, we built an emotional head motion predicting model. In order to find what kinds of prosodic and linguistic features have the most significant influence on head gesture generation, we proposed a two-layer cluster method and adopted CART model to construct the mapping model between speech representation and head gestures. In the proposed two-layer cluster method, the hierarchical parameters were determined by subjective evaluation on different parameter-related talking-head animation sequences, and the two-layer clusters schema was able to obtain more reliable clustering results from textual features. CART was further adopted to model the relationship between prosody, linguistic features and elementary head gesture patterns, which is of benefit to classify and interpret which features have greater influence on visual gesture generation. The analysis about how linguistic features contribute to different emotional expressions was presented in detail. The discussions in this work provide important reference for realistic animation of speech driven talking-head or avatar.

In our work, the bimodal database involves only one actress’ data. The bimodal mapping model is endowed with personal characteristics of the actress. The experiments show that the proposed method could effectively reveal how this actress’ emotional head expression is influenced by prosodic and linguistic features. One of our further work is to consider more players’ audio-visual data in our model.

Acknowledgments This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61332017, No.61375027, No.61203258, No.61273288, No.61233009, No.61425017).

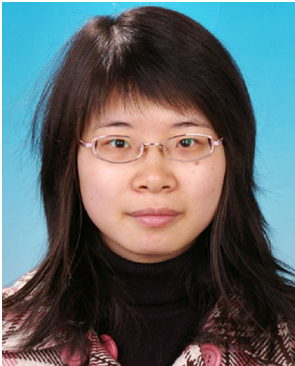
References

1. Alberto B, Piero C, Giuseppe RL, Giulio P (2014) LuciaWebGL a new WebGL-based talking head, 15th Annual Conference of the International Speech Communication Association, Singapore (InterSpeech 2014 Show & Tell Contribution)
2. Aleksandra C, Tomislav P, Pandzic IS (2009) RealActor: character animation and multimodal behavior realization system. *IVA*: 486–487
3. Ananthakrishnan S, Narayanan S (2008) Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Trans Audio Speech Lang Process* 16(1):216–228
4. Badler N, Steedman M, Achorn B, Bechet T, Douville B, Prevost S, Cassell J, Pelachaud C, Stone M (1994) Animated conversation: rule-based generation of facial expression gesture and spoken intonation for multiple conversation agents. *Proceedings of SIGGRAPH*, 73–80
5. Ben-Youssef A, Shimodaira H, Braude DA (2014) Speech driven talking head from estimated articulatory features, The 40th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), Florence, Italy
6. Bevacqua E, Hyniewska SJ, Pelachaud C (2010) Evaluation of a virtual listener smiling behavior. *Proceedings of the 23rd International Conference on Computer Animation and Social Agents*, Saint-Malo, France
7. Bo X, Georgiou Panayiotis G, Brian Baucom, Shrikanth S (2014) Narayanan, power-spectral analysis of head motion signal for behavioral modeling in human interaction, 2014 I.E. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), Florence, Italy
8. Bodenheimer B, Rose C, Rosenthal S, Pella J (1997) The process of motion capture: dealing with the data. In: Thalmann (ed) *Computer animation and simulation*. Springer NY 318 Eurographics Animation Workshop
9. Boulic R, Becheiraz P, Emering L, Thalmann D (1997) Integration of motion control techniques for virtual human and avatar real-time animation. In: *Proc. of Virtual Reality Software and Technology*, Switzerland: 111–118
10. Busso C, Deng Z, Neumann U, Narayanan S (2005) Natural head motion synthesis driven by acoustic prosodic features. *Comput Anim Virtual Worlds* 16(3–4):283–290
11. Cassell J, Vilhjalmsson HH, Bickmore TW (2001) Beat: the behavior expression animation toolkit. In: *Proceedings of SIGGRAPH*, 477–486
12. Chuang D, Pengcheng Z, Lei X, Dongmei J, ZhongHua Fu (2014) Northwestern, speech-driven head motion synthesis using neural networks, 15th Annual Conference of the International Speech Communication Association, Singapore (InterSpeech 2014)
13. Cohn JF, Schmidt KL (2004) The timing of facial motion in posed and spontaneous smiles. *Int J Wavelets Multiresolution Inf Process* 2:1–12
14. Cowie R, Douglas-Cowie E (2001) Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*. pp. 33–80
15. de Rosis F, Pelachaud C, Poggi I, Carofiglio V, De Carolis N (2003) From Greta's mind to her face: modeling the dynamics of affective states in a conversational embodied agent, special issue on applications of affective computing in human-computer interaction. *Int J Hum Comput Stud* 59(1–2):81–118
16. Faloutsos P, van de Panne M, Terzopoulos D (2001) Composable controllers for physics-based character animation. In: *SIGGRAPH '01: proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM Press, New York, p 251–260
17. Fangzhou L, Huibin J, Jianhua T (2008) A maximum entropy based hierarchical model for automatic prosodic boundary labeling in Mandarin. In: *Proceedings of 6th International Symposium on Chinese Spoken Language Processing*
18. Graf HP, Cosatto E, Strom V, Huang F (2002) Visual prosody: facial movements accompanying speech. In: *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*. Washinton D.C., USA
19. 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
20. Huibin J, Jianhua T, Wang X (2008) Prosody variation: application to automatic prosody evaluation of mandarin speech. In: *Speech prosody*, Brail
21. Jia J, Shen Z, Fanbo M, Yongxin W, Lianhong C (2011) Emotional audio-visual speech synthesis based on PAD. *IEEE Trans Audio Speech Lang Process* 19(3):570–582
22. Jianwu D, Kiyoshi H (Feb 2004) Construction and control of a physiological articulatory model. *J Acoust Soc Am* 115(2):853–870
23. Kipp M, Heloir A, Gebhard P, Schroeder M (2010) Realizing multimodal behavior: closing the gap between behavior planning and embodied agent presentation. In: *Proceedings of the 10th International Conference on Intelligent Virtual Agents*. Springer

24. Kopp S, Jung B, Lebmann N, Wachsmuth (2003) I: Max - a multimodal assistant in virtual reality construction. *KI -Kunstliche Intelligenz* 4/03 117
25. Kopp S, Wachsmuth I (2004) Synthesizing multimodal utterances for conversational agents. *Comput Anim Virtual Worlds* 15(1):39–52
26. Lei X, Zhiqiang L (2007) A coupled HMM approach for video-realistic speech animation. *Pattern Recogn* 40(10):2325–2340
27. Lei X, Zhiqiang L (2007) Realistic mouth-synching for speech-driven talking face using articulatory modelling. *IEEE Trans Multimedia* 9(3):500–510
28. Lijuan W, Xiaojun Q, Wei H, Frank KS (2010) Synthesizing photo-real talking head via trajectory-guided sample selection. *INTERSPEECH* 2010
29. Martin JC, Niewiadomski R, Devillers L, Buisine S, Pelachaud C (2006) Multimodal complex emotions: gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics, special issue Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids*, 3(3): 269–292
30. Meng Z, Kaihui M, Jianhua T (2008) An expressive TTVS system based on dynamic unit selection. *J Syst Simul* 20(z1):420–422
31. Parke F (1972) Computer generated animation of faces. *Proceedings of the ACM National Conference*
32. Pelachaud (2009) Modelling multimodal expression of emotion in a virtual agent. *Philos Trans R Soc B Biol Sci* 364:3539–3548
33. Scott AK, Parent RE (2005) Creating speech-synchronized animation. *IEEE Trans Vis Comput Graph* 11(3):341–352
34. Shao Y, Han J, Zhao Y, Liu T (2007) Study on automatic prediction of sentential stress for Chinese Putonghua Text-to-Speech system with natural style. *Chin J Acoust* 26(1):49–92
35. Shiwen Y, Xuefeng Z, Huiming D (2000) The guideline for segmentation and part-of-speech tagging on very large scale corpus of contemporary Chinese. *J Chin Inf Process* 6:58–64
36. Shiwen Y, Xuefeng Z, Huiming D (2002) The basic processing of contemporary Chinese corpus at Peking University SPECIFICATION 16(6)
37. Song M, Bu J, Chen C, Li N (2004) Audio-visual based emotion recognition- a new approach. In: *Proc. of the 2004 I.E. Computer Society Conference on Computer Vision and Pattern Recognition*. pp.1020–1025
38. Stone M, DeCarlo D, Oh I, Rodriguez C, Stere A, Lees A, Bregler C (2004) Speaking with hands: creating animated conversational characters from recordings of human performance. *ACM Trans Graph (SIGGRAPH'04)* 23(3):506–51
39. Tony E, Poggio T (2000) Visual speech synthesis by morphing visemes. *Int J Comput Vis* 38:45–57
40. Wachsmuth (2008) 'I, Max' - communicating with an artificial agent. In: Wachsmuth I, Knoblich G (eds) *Modeling communication with robots and virtual humans*. Springer, Berlin, pp 279–295
41. Wang QR, Suen CY (1984) Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition. *IEEE Trans Pattern Anal Mach Intell, PAMI* 6: 406–417
42. Waters K (1987) A musele model for animating three dimensional facial ExPRession. *Computer Graphics (SIGGRAPH,87)* 22(4): 7–24
43. Wei Z, Zengfu W (2009) Speech rate related facial animation synthesis and evaluation. *J Image Graph* 14(7):1399–1405
44. Welbergen HV, Reidsma D, Ruttkay ZM, Zwiers EJ (2010) A BML realizer for continuous, multimodal interaction with a virtual human. *J Multimodal User Interf* 3(4):271–284, **ISSN 1783-7677**
45. Yamamoto E, Nakamura S, Shikano K (1998) Lip movement synthesis from speech based on Hidden Markov Models. *Speech Comm* 26(1–2):105–115
46. Yamamoto SNE, Shikano K (1997) Speech to lip movement synthesis by HMM. In: *Proc.AVSP'97*. Rhodes, Greece
47. Young S, Evermann G, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P (2002) *The HTK book (for HTK version 3.2)*. Cambridge University Engineering Department
48. Young S, Jansen J, Odell J, Ollason D, Woodland P (1990) *The HTK book*. Entropic Labs and Cambridge University, 2.1
49. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans Pattern Anal Mach Intell* 31(1):39–58
50. Zeng Z, Tu J, Liu M, Huang TS, Pianfetti B, Roth D, Levinson S (2007) Audiovisual affect recognition. *IEEE Trans Multimedia* 9(2):424–428
51. Zhang S, Wu Z, Meng MLH, Cai L (2007) Head movement synthesis based on semantic and prosodic features for a Chinese expressive avatar. In: *IEEE Conference on International Conference on Acoustics, Speech and Signal Processing*
52. Zhenhua L, Richmond K, Yamagishi J (2010) An analysis of HMM-based prediction of articulatory movements. *Speech Comm* 52(10):834–846



Minghao Yang received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences. His current research interests include Speech Generation Visualization, Multimodal Human-Interaction and Emotion Recognition. He participated two China National 863 Program projects and one project of The Major Research Plan of National Natural Science Foundation of China. Now he hosts a General Program of National Natural Science Foundation of China<<Visualization for speech organs based on voice and medical video analysis(No: 61273288)>>, and host a sub project of Key Program of National Natural Science Foundation of China<<The neural control mechanism and modeling of speech(No:61233009)>>. He published over 20 papers on ACM Multimedia, TIP, MTAP, JMUI, ICASSP and other important international journals or conferences (SCI or EI index). He serves as the chair or program committee member for several major conferences, including ACII, NCMMS, MLSP etc. He obtained the best paper award of HMMME 2013, and the best paper award nominated by Audio/Visual Emotion Challenge Workshop (2014).



Jinlin Jiang is with the School of International Studies, University of International Business and Economics. She received her Ph.D. degree of applied linguistics from National Research Center for Foreign Language Education, Beijing Foreign Studies University. Her current research interests include automated scoring of human translation, quantification of human translation quality, and corpus-based study of English texts. She hosted a Program of National Social Science Foundation of China “Construction of Computer-assisted Scoring System for Student Translation of Special-Purposed Texts” (No. 11CYY007). She now hosts a Program of Social Science Foundation of Beijing “The Image of Beijing in Foreign Press: A perspective of Corpus-based Critical Discourse Analysis”. She published over 20 papers in journals of applied linguistics, computational linguistics and education technology (SCI, EI or CSSCI index)



Jianhua Tao received his PhD from Tsinghua University in 2001, and got his Ms from Nanjing University in 1996. He is currently a Professor in NLPR, Institute of Automation, Chinese Academy of Sciences. His current research interests include speech synthesis and coding methods, human computer interaction, multimedia information processing and pattern recognition. He has published more than eighty papers on major journals and proceedings including IEEE Trans. on ASLP, and got several awards from the important conferences, such as Eurospeech, NCMMSC, etc. He serves as the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, NCMMSC etc. He also serves as the steering committee member for IEEE Transactions on Affective Computing, associate editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions, Deputy Editor-in-chief for Chinese Journal of Phonetics.



Kaihui Mu received his Ph.D. degree from pattern recognition and intelligent systems at the Institute of Automation of Chinese Academy of Sciences, Beijing, China in 2011. His research interests include pattern recognition, machine learning and especially talking head, visual speech synthesis and multimodal interaction. He is now an assistant research fellow of Communication University of China.



Hao Li received his BS in automation from Xidian University, China, in 2010. Currently, he is working toward the PHD degree in pattern recognition and intelligent systems at the Institute of Automation of Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, machine learning and especially talking head, visual speech synthesis and multimodal interaction.