J Multimodal User Interfaces
DOI 10.1007/s12193-011-0073-5

ORIGINAL PAPER

A multimodal approach of generating 3D human-like talking agent

Minghao Yang · Jianhua Tao · Kaihui Mu · Ya Li · Jianfeng Che

Received: 24 April 2011 / Accepted: 18 October 2011

Abstract This paper introduces a multimodal framework of generating a 3D human-like talking agent which can communicate with user through speech, lip movement, head motion, facial expression and body animation. In this framework, lip movements are obtained by searching and matching acoustic features which are represented by Mel-frequency cepstral coefficients (MFCC) in audio-visual bimodal database. Head motion is synthesized by visual prosody which maps textual prosodic features into rotational and translational parameters. Facial expression and body animation are generated by transferring motion data to skeleton. A simplified high level Multimodal Marker Language (MML), in which only a few fields are used to coordinate the agent channels, is introduced to drive the agent. The experiments validate the effectiveness of the proposed multimodal framework.

Keywords Multimodal 3D talking agent · Lip movement · Head motion · MFCC · Facial expression · Gesture animation

1 Introduction

Natural human-like talking agent, as an exciting modality for human-computer interactions (HCI), has been studied intensively for the last ten years. Since the pioneering work of web woman newscaster *Ananova* [1], remarkable progress has been achieved in human-like talking agent [2–6, 9].

However, it's still a big challenge to make virtual agents conversation or response like a real person in human-computer interactions, because of human's great sensitivity over the subtleties of communications. Researchers and engineers focus on two issues to improve agent performance: (1) improving the animation skills for virtual bodies, making them life-like and able to express emotions; (2) building the controller for interactive channels, including speech, lip, head, body animation and facial expression, making them responsive to events in an interactive setting. Since speech is the dominant channel in dialog based human-computer interaction, how to fuse other animation channels with speech signals is crucial and challenging for natural life-style agents.

Most channels, such as facial expression, gesture and body, are not tightly connected with speech, whereas lip and head movement are closely related to speech signals. Traditional methods for animation usually focus on motion generation and rendering [10, 11], emotion expresses [12, 13]

This work is supported in part by National Science Foundation of China (No. 60873160 and No.90820303) and China-Singapore Institute of Digital Media (CSIDM).

Electronic supplementary material The online version of this article (doi:10.1007/s12193-011-0073-5) contains supplementary material, which is available to authorized users.

M. Yang (✉) · J. Tao · K. Mu · Y. Li
National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, 95♯ Zhongguancun East Road, Beijing 100190, P.R. China
e-mail: mhyang@nlpr.ia.ac.cn

J. Tao
e-mail: jhtao@nlpr.ia.ac.cn

K. Mu
e-mail: khMu@nlpr.ia.ac.cn

Y. Li
e-mail: yli@nlpr.ia.ac.cn

J. Che
Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: Jianfeng.Che@ia.ac.cn

Published online: 10 November 2011

or art expression [14]. However in the HCI fields, more attentions are paid on how to make all channels work synchronously with speech. A good controller would greatly help the agent to act like a human or to be able to interact flexibly with interactive setting [15, 16]. In view of that, Behavior Marked Language(BML) [6–8] and realization planning [17] are two classical techniques to coordinate different channels together. These methods are able to control each channel accurately and obtain good results. However, the potential disadvantage of these methods is that each channel must be considered elaborately, which makes these methods rather complicated, for artists and engineers, to design.

To simplify the design of animation controller, we introduce a multimodal framework for generating a 3D human-like talking agent in this paper. In our method, we mainly consider how all the channels and their emotional state work with speech signals. Some channels, closely related to speech, are directly driven by acoustic signals in low level; the other channels, which are non-closely related to speech, are realized by skeleton animation, which could be easily embedded to speech in high level. In this way, a simplified high level Multimodal Marker Language, in which only a few fields are adopted to control speech generation, lip, head, face and body movements, is proposed to drive the agent. The experiments show that the proposed multimodal framework can generate vivid and natural talking agent.

The rest parts of paper are organized as follows. Related work of agent's animation and controller techniques are introduced in the second section. In section three, we introduce the total framework of our multimodal approach. The details of different channels' animation techniques and their relationships with acoustic signals will be presented in section four, five, six. In section seven, we present the 3D talking agent driven by the proposed simple Multimodal Marker Language. A short summary is given in the last section.

2 Related work

2.1 Agent animation

Speech is the most important channel in dialog based human computer interaction. Many multimodal techniques, focusing on building relationship between others channels and acoustic signals, have been developed greatly, such as [2, 3]. However, some have not been well modeled by now. In these interactive channels, the most important ones are speech-driven lip synchronization, head movement, pose, and gesture.

Speech-driven lip synchronization make lip move with the acoustic synchronously. Two types of driven sources are usually utilized as input to a facial animation system: text and audio [18]. Text-driven face animation is often applied to an automatic and intelligent dialogue system, in which the speech is usually rendered by a Text-to-Speech (TTS) system. Since human speakers can create more expressive speech due to their capability to adapt intonation, animating a face model from real speech is greatly used in audio driven lip synchronization applications.

Natural head motion is an indispensable part of talking head. Numerous animation techniques are presented to synthesize head motions and these techniques could be categorized into two types. The first one is parameter-driven (rule-based) head motion synthesis [5]. In this method, the velocity and the global pose of the head motion are generated by predefined rules. The second one is data-driven head movement synthesis [19, 20]. In this method, patterns of head and facial movements are strongly correlated with the prosodic structure of the text. As for a freely given sentence, this method could generate more nature talking head than that of parameter-driven head motion.

Pose and gesture, including facial expression and gaze, contribute to understanding mood and character in communication. There are three methods for generating gaze, body and face animation for virtual agent. The early method is mapping random or predefined actions to virtual agent. The second one is building character animation through commercial platform and physics-based controller [14]. Every key frame must be designed by artist carefully in this method, and physics-based animation calculation is very time-consuming. The third kind of method is recording and mapping actor actions through motion capture devices, which does not highly depend on the design skills of artist [21–23].

In our method, pose, gesture animation are generated by transferring motion capture date to agent's skeleton. Face and gaze animation, which cover seven kinds of emotional state, are direct mapped from actor's emotions to agent. With the help of textual analysis technique, lip movements and head motion are obtained by matching features in training set. As channels have different relationship with speech, we combine pose, gesture animation with speech in high level, and realize lip-synchronization and head movement with acoustic signals in low level. This combination schema contributes to a simplified Multimodal Marker Language (MML) in our framework.

2.2 Agent controller

When animations are constructed, how to coordinate these channels, and make the agent *smart* enough and flexibly responsive to events in an interactive setting, is a hot topic in recent years. [24] proposed a hierarchically connected animation controllers based on SmartBody, an open source engine for animating agent. In [25], authors use baseline, a set of fixed parameters, to present the personalized agent

behavior, and adopt dynamicline, a set of flexible parameters, to drive agents' communicative goals and emotional states. As a highly modular and extensible script, Behavior Markup Language were proposed to provide a mix between the precise temporal and spatial control offered by procedural animation and the physical realism of physical simulation [6, 7]. Furthermore, BML is combined with a behavior lexicon and a kind of lower level animation script to integrate reactive emotional behaviors for agents in [8]. In recent years, emotion presentation and social perception become more and more important issues in virtual and mixed reality settings, where audio/visual coherently work is key technique for agent animation [9].

BML is an elegant technique for agent control. However, BML is more like a script standards rather than a set of techniques. When interaction scenes become complex and social perception is highly demanded, BML's design seems to be tedious work for artists or engineers. In this paper, we proposed a simple Marked Language to drive agent in a synchronous ways on basis of our multimodal framework, and experiences show that this method gives satisfied results.

3 Total framework

System framework is presented as Fig. 1, where text, pose (gesture) and expression (gaze) are three input modules. Speech, lip-sync, head motion, body animation, facial expression and gaze movement are six output modules. In output modules, speech (acoustic signals) is the dominant channel, which is generated by Text-to-Speech (TTS) system. At the same time, lip movement is generated from speech signals and head motion is synthesized based on the visual prosody of text. These two channels are closely related

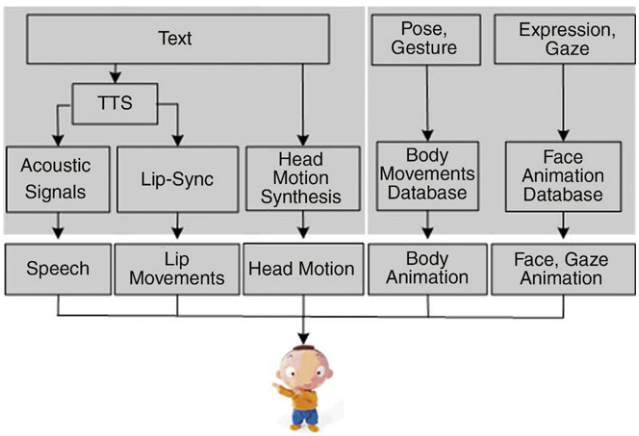

Fig. 1 Total framework of the multimodal approach. There are only three input types in our system. However, total six output modalities are generated from the three input types. These six modalities are speech, lip movements, head motion, facial expression, skeleton-driven body animation (pose and gesture), facial expression and gaze movement

to acoustic signals and are not marked in the Multimodal Marker Language (MML) and we will introduce MML in the sixth section.

Body movement, together with facial expression and gaze, presents the character or emotional state in conversations. A gesture or expression could be appended to several different sentences. On the other hand, a sentence could be presented with different gestures and expression. As body and gaze movement are non-closely related to text and speech, these channels needed to be marked obviously in Multimodal Marker Language (MML).

4 Real-time speech-driven lip synchronization

The goal of lip synchronization is to render lip movements and make it synchronized with the acoustic signal. Many algorithms have been developed to synthesize lip movement from speech, including linear prediction analysis [26], artificial neural networks (ANNs) [27], and hidden Markov models (HMMs) [28]. As linear prediction analysis easily lead to discontinuous mapping result and ANNs suffers from that the number of hidden layers and the number of nodes per layer may be experimentally determined [29], HMMs are the most common method to transform audio features into a stream of animation parameters for a facial model due to its capability of handling voice contextual information and modeling coarticulation. However, HMMs based approaches have relative long time delay.

We adopt a data-driven approach based on k-nearest neighbors (kNN), which is called collaborative filtering [30], to realize speech-driven lip synchronization in our system. The method is divided into following phases, first an audio-visual database, which uses frames as units, is built. Then, a visual sequence is constructed by concatenating the appropriate weighted visual frames from the database selected by collaborative filtering when acoustic signals are input into this system. This approach creates lip synchronization in real-time and we introduce it in details.

4.1 Data collection and FAPs extraction

The data collection device is presented in Fig. 2(a), where 8 cameras is adopted to track 25 markers on the face of an actress (Fig. 2(c)). In lip-sync data collection, 694 sentences are uttered by the performer in different emotional state during a series of motion capture sessions. We pay more attention to the markers placed on the corner of mouth, eyes, eyebrows of the actress, since these corner points provide more facial expression information.

The MPEG-4 Facial Animation specification provides a set of FAPs for animating any MPEG-4 compliant face model. The Facial Definition Parameters (FDPs) are defined

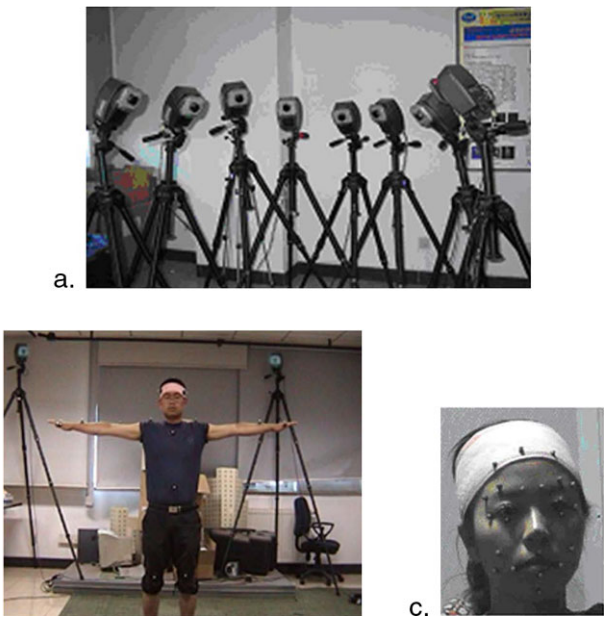

Fig. 2 (**a**) Motion Analysis devices (**b**) Placement of markers on body. (**c**) Placement of 25 markers on face

by the locations of the feature points and are used to customize a given face model to a particular face. Each FAP value is simply the displacement of a particular feature point from its neutral position expressed in terms of the Facial Animation Parameter Units (FAPUs). The FAPUs correspond to fractions of distances between key facial features. Thus, once we have the displacements of the feature points from the neutral position, it is easy to extract the FAPs corresponding to the given facial animation. Detailed description about MPEG-4 standard and FAPs can be found in [31]. The distances between some key facial features are calculated as FAPUs in the reference frame according to MPEG-4 standard. Then each FAP value is measured by the displacement of a particular feature point in each frame. In this work, 25 FAPs in MPEG-4 are extracted to represent the lip movements.

4.2 Acoustic features encoding and collaborative filtering

The HTK tools [32] are utilized to compute 12 cepstral coefficients of MFCCs [33] and the energy term. In this work, MFCC's delta and acceleration components are considered to model temporal dynamic changes in the speech signal. The acoustic signals are processed using a 20 ms Hamming window with overlapping frames of 10 ms. The frame rate of FAPs is upsampled from 75 to 100 frames per second. Then, the synchronized FAPs and MFCCs with their delta and acceleration components are combined to form MFCC-FAP corpus.

Collaborative filtering algorithm works by search a large group of samples to finding a smaller set similar to the input sample. It looks at features corresponding to each sample in this smaller set and combines them to create a ranked list of features. These ranked features are weighted to form new features. During this period, MFCCs with their derivatives are used as samples and the corresponding FAPs are utilized as the features.

The algorithm for lip-sync Synchronization is procedure as following:

(1) Similarity metric (formula (1)) are used to find the similar samples for a new input frame from the samples and the corresponding features database. In (1), x_i and y_i are the items in two vectors that both have the same dimension. r is the similarity score between these two vectors. Higher value of r indicates being more similar.

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

(2) Sort the similar sample set according to the similarity values to form the first k items of the sorted results.

(3) The features of each sample in the ranked sample set are scored by multiplying the similarity value for each ranked sample.

(4) These features are divided by the sum of all the similarities as formula (2) shows. In (2), f_{ij} and w_{ij} are the feature vector and the similarity value of the ith sample in the jth sample set.

$$f_i = \left(\sum_{j=1}^{k} w_{ij} f_{ij} \right) \Big/ \left(\sum_{j=1}^{k} w_{ij} \right) \tag{2}$$

(5) Finally, f_i in formula (2) is the generated feature vector of the ith new input sample in our system.

5 Visual prosody for head motion synthesis

We synthesize head motion by textual analysis and interpolation technique on time sequence. Firstly, like what we have introduced in Sect. 4, we cluster head movement patterns that show up in each of seven different emotional states performed by one actress, and investigate how they relate with text's features. A data mining approach is proposed to establish the relationship between input text and constructed database. Our workflow for head motion is divided into following two steps:

5.1 Features collection

The head motion database is recorded by an actress. Like what we collection features for lip synchronization, 489 sentences are selected and each sentence is spoken in different

emotional states, which are selected according to the meaning of the sentences and spoken mimicking that situation. Having only one performer, it is easy to maintain the style and personality in each emotional state. It is also convenient to cluster the head motion patterns in each emotional state.

All the features are extracted by the textual analysis module of a Chinese text-to-speech (TTS) system [34]. Grammatical and prosodic features are utilized to modeling the relationship with head motion. In this work, the prosodic features contain the stress point in a sentence, the boundary type, the position of a syllable in a sentence, tone type in Mandarin Chinese, part-of-speech of word and the length of the prosodic word. All these features combine to form visual prosody of head motion [35].

All of the textual features are organized in the syllable level. Note that we assume that the head motion is consistency in one syllable, which may smooth some sharp head motion. However it does not influence the total trend of head motion.

5.2 Clustering of head motion features

The trajectories of the 25 markers contains all of the head motion information. The rigid head motion parameters, including Euler angles and translations, are extracted from those trajectories. Head translations are measured by integrating among all the markers' distance to the maker which is placed at nose. The rotation matrix is computed using the method in [36] and the three Euler angles could be obtained from the rotation matrix. The boundaries of each syllable in a sentence are aligned with HTK [32] tools based on the acoustic speech recorded by the audio recording device. These boundaries are utilized to help find the available head motion frames in the synchronized motion head data.

We use a two-layer clustering method to find the elementary head motion patterns from the 18 dimensional rigid head motion features in each emotional state. K-Means clustering is firstly be utilized to extract the head motion patterns in each emotional state. The Pearson correlation score is applied to indicate the similarity between two frames of head motion features. A hierarchical clustering is used to model how these groups are related after using K-means clustering to create a set of groups. Then, some adjacent groups are merged depending on a threshold. The similar groups always influence the accuracy in modeling mapping issue as shown later. Merging some of them reduce the situation that some samples are likely to be classified into similar groups.

After the head motion patterns are determined, the available frames of head motion features in each sentence are assigned to the nearest group, which are used as the training parameters as the synchronized textual features for Classification and Regression Trees (CART) modeling.

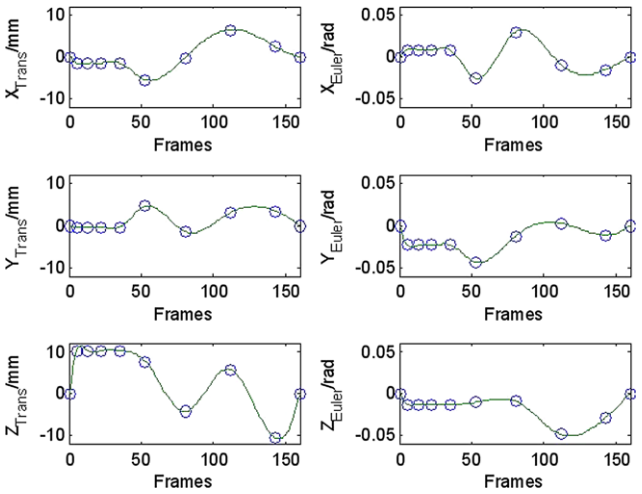

Fig. 3 Synthesized curves smoothed by spline interpolation in the sentence *zhe li de jing se hao mei a* (The scenery is very beautiful)

Figure 3 shows the synthesized curves of translations and Euler angles using the model in surprise state on a sample sentence. 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 where the angle around the x-axis dominates to show a nod in head motion.

6 Speech-independent animation and multimodal marker language

6.1 Body, pose and gesture animation

In our system, body and face animation, including pose, gesture movements are non-closely related to speech. We call them Speech-Independent Animation and they are driven by the selection actions from the database recorded from motion capture devices. Here, the motion capture devices (Fig. 2(a)) track 28 markers on the body of an actor (Fig. 2(b)) with 8 cameras. We obtain the 3D trajectories for each marker points as the output of the tracking system. For simplification, we divide human's actions into seven mood states: neutral, happiness, sadness, anger, disgust, surprise, fear. In each emotional state, the actor plays strong and weak actions respectively. As s result, there are two copies for any action unit on two different states. Figure 4 lists capture frames for two kind of happiness (a) (b) and anger (c) (d) actions on strong and weak degrees. The actions in same column belong to same emotional state, and are different on action speed and magnitude.

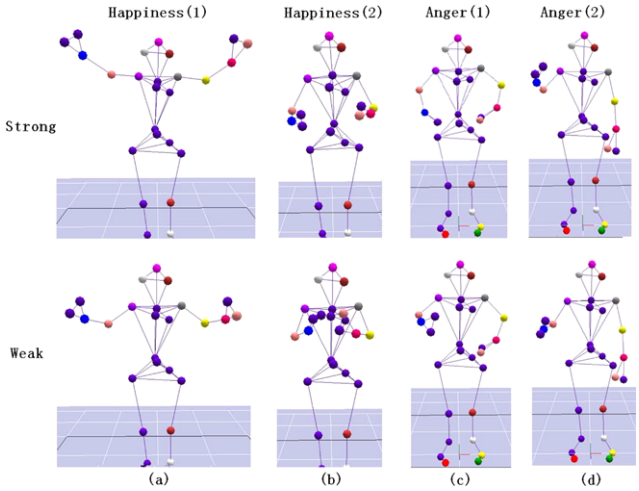

Fig. 4 Several selected frames of tracked markers on actor for happiness, anger actions with strong and weak emotional state

6.2 Multimodal Marker Language

We propose a Multimodal Marker Language (MML) to drive agent based on the techniques introduced in previous sections. Being different from the traditional Behavior Marked Language(BML), MML is a high level language. It's format is: < EmtnId = ∗ DgrId = ∗ GazeId = ∗ Text = "..">.

The value of field *EmtnId* is between 1 and 7, which presents one of the seven basic emotions and the value of *DgrId* is 1 and 0, presenting emotional degree on strongness or weakness respectively. *GazeId* is the virtual agent's eyes movement directions, which have five properties: *LeftUp*, *RightUp*, *Straight*, *LeftBottom*, *RightBottom*. The field *Text* is followed by a sentence. This sentence will be spoke by agent, and lip-synchronization and head motion are automatically generated from visual prosody of *Text*. As a result, there are no fields defined for lip and head in MML. This makes MML simple in form.

7 Experiences

7.1 Comparison synthesized lip movement with original recorded data

We evaluate lip movement and head motion by comparison them with actor's raw actions. Figure 5 lists the original and final created trajectories of FAP 3 open_jaw for a sentence. The generated FAP trajectory (red curve) reconstructs the original FAP curve (blue curve) well in the overall temporal and spatial trends of the whole curve. Additionally, the generated FAP trajectory has more details and slightly smaller amplitudes than the original one. The big differences between these two curves in the beginning and end stages are made because the silence is removed from the MFCC-FAP

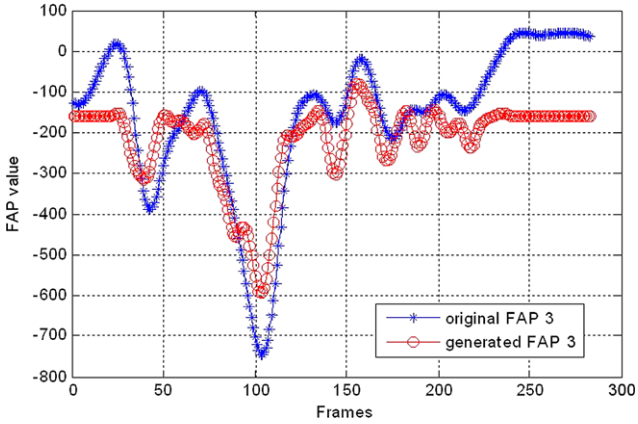

Fig. 5 Trajectories of the original and generated FAP 3 for a sentence on lip-movement

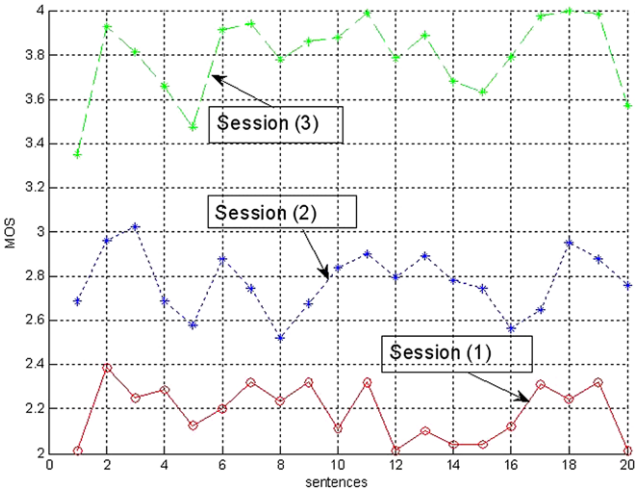

Fig. 6 MOS scores of 20 sentences on body, face and gaze animation

corpus and replaced by the same FAP value in the synthesis stage. We can see from Fig. 5 that our system generate nature lip-movement.

7.2 Subjective evaluation on body, face, gaze and head animation

At the same time, subjective evaluation on body, face, gaze animation and head motion of 3D taking agent are given by Mean Opinion Score (MOS) scores, which is adopted for different sentences on: (1) pure acoustic output without any animation on body, face, gaze animation and head motion; (2) 3D talking with the random animation of above channels; (3) the agent generated by our multimodal approach.

For each sentence, the subjects were asked to score the expressivity of 3D agent on a five level mean opinion score (MOS) scale. The results show in Fig. 6. The average MOS scores of three sessions are 2.2 (session (1)), 2.8 (session (2)), 3.8 (session (3)). The agent without any animation

Fig. 7 Several selected frames for agent speaking *It's sunny tomorrow* with neural emotion state

is always considered unnatural and displays an apparent synthetic appearance. The one with random animation seems to float over the background, but gets a higher score than that without animation. When the features synthesized with the multimodal approach proposed in this paper is added, the 3D talking agent shows an appearance close to the realistic face animation and improves the MOS by 1.0 points than that with random animation on body and face.

7.3 Visual agent

Depending on MML controller techniques that we introduced in previous sections, we build a 3D talking agent with free Cal3d platform [37] on PC with 2.6 G CPU and 2 G RAM. As Cal3D support skeleton animation for body and face, the recorded action units or tracked marker points on body and face could be directly transferred to virtual agent on different emotional demanding. For head motion and lip movement, these two channels' animation are driven by the input *text* in MML. Figure 7 present several frames for the 3D talking agent for the sentence *It's sunny tomorrow*.

8 Conclusions

This paper analysis the channels from the emotional state and the personalized performance of one actor and validates the feasibility of creating body, face, gaze animation from capture data and head motion, lip and head movement from visual prosodic features. Multimodal approaches are effective to enhance the expressivity of agent. And MML is proved to be a effective way to coordinate all channels for talking agent. The synthesized 3D talking agent with speech and lip, head, body, face, gaze animation give satisfied results.

Though natural 3D talking agent is achieved in our work, there are some improvements needed to be done. The first one is the emotional state of each sentence in MML needs to be assigned by user, a work of predicting the emotional state from the textual context information should be added. Secondly, all the body and face action units in database are recorded in advance, which could not modified by users in run-time. We aim to create more action units from the fixed action in database by motion morph or planning techniques in future.

References

1. http://en.wikipedia.org/wiki/Ananova (2011) Accessed 16 August
2. Wik P, Hjalmarsson A (2009) Embodied conversational agents in computer assisted language learning. Speech Commun 51(10):1024–1037
3. http://www.mmdagent.jp/ (2011) Accessed 16 August
4. http://www.speech.kth.se/multimodal/ (2011) Accessed 16 August
5. 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. In: Proceedings of SIGGRAPH, pp 73–80
6. Van Welbergen H, Reidsma D, Ruttkay ZM, Zwiers Elckerlyc J (2010) A BML realizer for continuous, multimodal interaction with a virtual human. J Multimodal User Interfaces 3(4):271–284 ISSN 1783-7677
7. Cerekovic A, Pejsa T, Pandzic IS (2009) RealActor: character animation and multimodal behavior realization system. In: IVA, pp 486–487
8. 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, Berlin
9. Courgeon M, Rebillat M, Katz B, Clavel C, Martin J-C (2010) Life-sized audiovisual spatial social scenes with multiple characters: MARC SMART-I2. In: Proceedings of the 5th meeting of the French association for virtual reality
10. Park SI, Shin HJ, Shin SY (2002) On-line locomotion generation based on motion blending. In: Proc of the ACM SIGGRAPH/eurographics symposium on computer animation, New York, NY, USA. ACM Press, New York, pp 105–111
11. Baerlocher P (2001) Inverse kinematics techniques for the interactive posture control of articulated figures. PhD thesis, Swiss Federal Institute of Technology, EPFL
12. Cassell J, Vilhjalmsson HH, Bickmore TW (2001) Beat: the behavior expression animation toolkit. In: Proceedings of SIGGRAPH, pp 477–486
13. Gu E, Badler N (2006) Visual attention and eye gaze during multipartite conversations with distractions. In: Proc of intelligent virtual agents (IVA'06), Marina del Rey, CA
14. 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, New York, NY, USA. ACM Press, New York, pp 251–260
15. Kuffner JJ, Latombe JC (2000) Interactive manipulation planning for animated characters. In: Proc of pacic graphics'00, Hong Kong
16. Kallmann M (2005) Scalable solutions for interactive virtual humans that can manipulate objects. In: Artificial intelligence and interactive digital entertainment (AIIDE), Marina del Rey, CA
17. Carolis BD, Pelachaud C, Poggi I, de Rosis F (2001) Behavior planning for a reflexive agent. In: Proceedings of the international joint conference on artificial intelligence (IJCAI'01), Seattle
18. Graf HP, Cosatto E, Ostermann J, Schroeter J (2003) Lifelike talking faces for interactive services. Proc IEEE 91:1406–1429
19. Graf HP, Cosatto S., Huang F (2002) Visual prosody: facial movements accompanying speech. In: Fifth IEEE international conference on automatic face and gesture recognition
20. Chuang E, Bregler C (2005) Mood swings: expressive speech animation. ACM Trans Graph 24:331–347
21. Bodenheimer B, Rose C, Rosenthal S, Pella J (1997) The process of motion capture: Dealing with the data. In: Thalmann, computer animation and simulation. Eurographics Animation Workshop. Springer, New York, p 318

22. 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, pp 111–118
23. 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 23(3):506–513
24. Thiebaux M, Marsella S, Marshall AN, Kallmann M (2008) SmartBody behavior realization for embodied conversational agents AAMAS. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, international foundation for autonomous agents and multiagent systems, 2008, pp 151–158
25. Mancini Greta M, Pelachaud C (2007) Dynamic behavior qualifiers for conversational agents. In: Intelligent virtual agents, IVA'07, Paris
26. Lewis J (1991) Automated lip-sync: Background and techniques. J Vis Comput Animat 2:118–122
27. Wen Z, Hong P, Huang TS (2002) Real-time speech-driven face animation with expressions using neural networks. IEEE Trans Neural Netw 13:916–927
28. Xin L, Tao J, Yin P (2009) Realistic visual speech synthesis based on hybrid concatenation method. IEEE Trans Audio Speech Lang Process 17:469–477
29. Che J, Yang M, Mu K, Tao J (2010) Real-time speech-driven lip synchronization. In: 4th International universal communication symposium, pp 377–381
30. Oki BM, Goldberg D, Nichols D, Terry D (1992) Using collaborative filtering to weave an information tapestry. Commun ACM 35:61–70
31. Tekalp AM, Ostermann J (2000) Face and 2-d mesh animation in mpeg-4. Signal Process Image Commun 15:387–421
32. Young S et al (2000) The HTK book (v3.0). Cambridge University Engineering Department, Cambridge
33. Xu M, Duan L-Y, Cai J et al (2004) HMM-based audio keyword generation. In: Advances in multimedia information processing, 5th Pacific rim conference on multimedia
34. Jia H, Liu F, Tao J (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
35. Mu K, Tao J, Che J, Yang M (2010) Mood Avatar: Automatic Text-Driven Head Motion Synthesis. In: 12th international conference on multimodal interfaces and 7th workshop on machine learning for multimodal interaction, Beijing, China
36. Deng Z, Neumann U, Busso C, Narayanan S (2005) Natural head motion synthesis driven by acoustic prosodic features. Comput Animat Virtual Worlds 16:283–290
37. http://home.gna.org/cal3d/ (2011) Accessed 16 August