Multimed Tools Appl (2015) 74:9845–9848
DOI 10.1007/s11042-015-2460-5

GUEST EDITORIAL

Expressive talking avatar synthesis and animation

**Lei Xie · Jia Jia · Helen Meng · Zhigang Deng ·
Lijuan Wang**

Published online: 28 January 2015

The talking avatar, an animated speaking virtual character with vivid human-like appearance and real or synthetic speech, has gradually shown its potential in applications involving human-computer intelligent interactions. The talking avatar has rich communication abilities to deliver verbal and nonverbal information by voice, tones, eye-contact, head motion and facial expressions, etc. Avatars are increasingly being used on a variety of electronic devices, such as computers, smart phones, pads, kiosks and game consoles. Avatars also can be found across many domains, such as technical support and customer service, communication aids, speech therapy, virtual reality, film special effects, education and training [6]. Specific applications may include a virtual storyteller for children, a virtual guider or presenter for personal or commercial website, a representative of user in computer games and a funny puppetry for computer-mediated human communications. It is clearly promising that talking avatars will become an expressive multimodal interface in human computer interaction.

L. Xie (✉)
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
e-mail: lxie@nwpu.edu.cn

J. Jia
Department of Computer Science and Technology, Tsinghua University, Beijing, China
e-mail: jjia@mail.tsinghua.edu.cn

H. Meng
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
e-mail: hmmeng@se.cuhk.edu.hk

Z. Deng
Computer Science Department, University of Houston, Houston, USA
e-mail: zdeng@cs.uh.edu

L. Wang
Microsoft Research Aisa, Beijing, China
e-mail: lijuanw@microsoft.com

Despite many years of efforts [1, 4, 8, 14, 15], current research has not yet advanced to the stage where it is possible to synthesize an affective intelligent talking avatar who can express their feelings and emotions through vocal and facial behaviors as natural as our human beings. To this end, recent researches aim to understand complicated human behaviors and generate lifelike speech and facial expressions according to the emotional content to enhance the expressivity of talking avatar.

This special issue aims to bring together researchers engaged in the various research directions of expressive talking avatar synthesis and animation. We received more than 20 high-quality paper submissions and each paper was peer-reviewed by at least three reviewers. After several rounds of review, ten papers were finally selected to be included in this special issue. These papers can be categorized into three topics: avatar animation [3, 9, 13], speech synthesis [11, 12, 16, 18] and human emotion/behavior analysis [5, 10, 17].

Generating lifelike talking avatars remains a challenging task despite decades of research. Photo- or video-realistic appearance is one important goal, which aims to make a talking head looking like a real person [1]. In this special issue, Wang and Soong from Microsoft Research Asia propose a hidden Markov model (HMM) trajectory-guided, real image sample concatenation approach to photo-realistic talking head animation [9]. Firstly, they propose to model and predict the lip movement trajectory with the statistical HMM model and the model is initialized with maximum likelihood training and further refined under minimum generation error criterion. Secondly, they use the trajectory- guided sample selection method, in which the HMM predicted visual trajectory is used as guidance to select the real samples from the image library for a photo-realistic head rendering. Their talking head took part in the LIPS2009 Challenge contest in the AVSP (Audio-Visual Speech Processing) workshop and won the first place in the Audio-Visual match evaluated by many subjects.

Non-verbal cues, e.g., hand gestures, facial expressions and head motions, are used to express feelings, give feedbacks and engage human-human communication. Hence natural head motion is an indispensable factor for a computer-animated talking avatar looking lifelike [1, 7]. To this end, Ding and Xie [3] study the head motion synthesis problem using neural networks. Recently, deep neural networks (DNNs) and deep learning [2] have been successfully used in many tasks. Different from previous approaches that regard the speech-to- head-motion mapping as a classification task, the proposed approach directly learns a speech-to-head-motion regression model using a DNN. Significant performance gain in head motion prediction is reported. Taking the advantage of the rich non-linear learning ability, Wu et al. [13] develop a DNN approach for real-time speech driven talking avatar. Specifically, the input of the system is acoustic speech and the output is articulatory movements on a three-dimensional avatar. Experimental results demonstrate that the proposed acoustic to articulatory mapping approach with DNN can achieve the best performance as compared with general linear model (GLM), Gaussian mixture model (GMM) and conventional artificial neural network (ANN).

Synthesizing natural speech is essential for a lifelike avatar. Wu et al. [12] focus on generating emphatic speech since emphasis plays an important role in highlighting the focus of an utterance to draw the attention of a user. As there are only a few emphasized words in a sentence, the problem of the data limitation is one of the major obstacles in this area. To deal with this data sparseness problem, they propose an HMM based method with limited amount of data. Experiments indicate that the proposed emphatic speech synthesis models improve the emphasis quality of synthesized speech while keeping a high degree of the naturalness. Another paper from Yang et al. [16] aims to give a talking avatar the multilingual speaking ability. Specifically, they consider a real-world low-resource scenario, i.e., using the data from a rich-resource language (Mandarin) to establish a speech synthesis system for a phonetic-related, low-recourse language (Tibetan). With the help of speaker adaptive training strategy, their HMM-based cross-lingual speech synthesis system outperforms the system using only Tibetan speaker dependent models when only a small number of Tibetan training utterances are available.

Voice conversion is quite useful in talking avatar animation, which aims to manipulate the source speaker's voice to let it sound like another target speaker. The primary challenge in voice conversion is how to implement a stable conversion function given the parallel source-target speech utterances. Wu et al. [11] propose an exemplar-based voice conversion approach that assumes a target spectrum can be produced as a weighted linear combination of a set of basis target spectra (exemplars). To do such a regression, coupled source-target dictionaries consisting of acoustically aligned source-target exemplars are assumed to share the same linear combination weights (activations). In Wu's approach, a joint nonnegative matrix factorization (NMF) with sparsity constraint is used to find the activation weights. Objective and subjective experiments confirmed the effectiveness of the proposed NMF approach.

Current emotional speech synthesis technologies are far from mature to achieve human-level expressivity. To address this limitation, Yilmazyildiz et al. [18] provide another way using gibberish speech to enable affective expressivity for robotic agents. Gibberish speech is composed of vocalizations of meaningless strings of speech sounds, which is used by performing artists and cartoon/game makers to express intended emotions. The proposed study has shown that the generated gibberish speech can contribute to a significant extent to studies concerning emotion expression for robotic agents. Besides human robot/avatar interaction, synthetic gibberish speech can be used in segmental evaluation of synthetic speech, testing the effectiveness of affective prosodic strategies and other studies.

It will definitely enhance the immersive experience if a talking avatar is sensitive to user behaviors and emotions. To this end, Wang et al. [10] propose a relevance units machine (RUM) approach for dimensional and continuous speech emotion prediction while Gonzalez et al. [5] aim to recognizing facial actions and their temporal segments based on duration models. User behavior is multimodal in nature, which involves speech, facial expression, head motion and body gestures. In order to improve the avatar's sensitivity, the approach proposed by Yang et al. [17] combines user's multi-modal behaviors with behavior history cues in a dialog management (DM) system. Experiments show that the behavior sensitive DM system makes the avatar be able to sensitive to the users facial expressions, emotional voice and gesture, which enhances user experience in multi-modal human-computer conversation.

We hope that the readers will find these papers informative and interesting. We would like to thank the authors of all submitted papers. We also wish to offer our sincere thanks to the Editor-in-Chief, Professor Borko Furht and to all editorial staffs for their valuable supports throughout the preparation and publication of this special issue. We also thank to the reviewers for their help in reviewing the papers.

References

1. Cosatto E, Ostermann J, Garf HP, Schroeter J (2003) Lifelike talking faces for interactive services. Proc IEEE 91:1406–1429
2. Deng L, Yu D (2014) Deep learning: methods and applications, Now Publishers

3. Ding C, Xie L, Zhu P (2014) Head motion synthesis from speech using deep neural networks. Multimed Tool Appl. doi:10.1007/s11042-014-2156-2
4. Ezzat T, Geiger G, Poggio T (2002) Trainable video realistic speech animation. In: ACM SIGGRAPH, pp. 388–398
5. Gonzalez I, Cartella F, Enescu V, Sahli H (2014) Recognition of facial actions and their temporal segments based on duration models. Multimed Tool Appl. doi:10.1007/s11042-014-2320-8
6. Hura S, Leathem C, Shaked N (2010) Avatars meet the Challenge. Speech Technol. 30–32
7. Le BH, Ma X, Deng Z (2012) Live speech driven head-and-eye motion generators. IEEE Trans Vis Comput Graph 18(11):1902–1914
8. Wang L, Han W, Soong F, Huo Q (2011) Text-driven 3D photo-realistic talking head. In: Interspeech
9. Wang L, Soong FK (2014) HMM trajectory-guided sample selection for photo-realistic talking head. Multimed Tool Appl. doi:10.1007/s11042-014-2118-8
10. Wang F, Sahli H, Gao J, Jiang D, Verhelst W (2014) Relevance units machine based dimensional and continuous speech emotion prediction. Multimed Tool Appl. doi:10.1007/s11042-014-2319-1
11. Wu Z, Chng ES, Li H (2014) Exemplar-based voice conversion using joint nonnegative matrix factorization. Multimed Tool Appl. doi:10.1007/s11042-014-2180-2
12. Wu Z, Ning Y, Zang X, Jia J, Meng F, Meng H, Cai L (2014) Generating emphatic speech with hidden markov model for expressive speech synthesis. Multimed Tool Appl. doi:10.1007/s11042-014-2164-2
13. Wu Z, Zhao K, Wu X, Lan X, Meng H (2014) Acoustic to articulatory mapping with deep neural network. Multimed Tool Appl. doi:10.1007/s11042-014-2183-z
14. Xie L, Liu Z-Q (2007) Realistic mouth-synching for speech-driven talking face using articulatory modelling. IEEE Trans Multimed 9(23):500–510
15. Xie L, Sun N, Fan B (2013) A statistical parametric approach to video-realistic text-driven talking avatar. Multimed Tool Appl 73(1):377–396
16. Yang H, Oura K, Wang H, Gan Z, Tokudai K (2014) Using speaker adaptive training to realize mandarin-tibetan cross-lingual speech synthesis. Multimed Tool Appl. doi:10.1007/s11042-014-2117-9
17. Yang M, Tao J, Chao L, Li H, Zhang D, Che H, Gao T, Liu B (2014) User behavior fusion in dialog management with multi-modal history cues. Multimed Tool Appl. doi:10.1007/s11042-014-2161-5
18. Yilmazyildiz S, Verhelst W, Sahli H (2014) Gibberish speech as a tool for the study of affective expressiveness for robotic agents. Multimed Tool Appl. doi:10.1007/s11042-014-2165-1