Photo-Realistic Expressive Text to Talking Head Synthesis

Vincent Wan[1], *Robert Anderson*[2], *Art Blokland*[1], *Norbert Braunschweiler*[1], *Langzhou Chen*[1], *BalaKrishna Kolluru*[1], *Javier Latorre*[1], *Ranniery Maia*[1], *Björn Stenger*[1], *Kayoko Yanagisawa*[1], *Yannis Stylianou*[1], *Masami Akamine*[3], *Mark J. F. Gales*[1] *and Roberto Cipolla*[1]

[1]Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK
[2]Department of Engineering, University of Cambridge, Cambridge, UK
[3]Corporate Research and Development Center, Toshiba Corporation, Kawasaki, Japan

vincent.wan@crl.toshiba.co.uk

Abstract

A controllable computer animated avatar that could be used as a natural user interface for computers is demonstrated. Driven by text and emotion input, it generates expressive speech with corresponding facial movements. To create the avatar, HMM-based text-to-speech synthesis is combined with active appearance model (AAM)-based facial animation. The novelty is the degree of control achieved over the expressiveness of both the speech and the face while keeping the controls simple. Controllability is achieved by training both the speech and facial parameters within a cluster adaptive training (CAT) framework. CAT creates a continuous, low dimensional eigenspace of expressions, which allows the creation of expressions of different intensity (including ones more intense than those in the original recordings) and combining different expressions to create new ones. Results on an emotion-recognition task show that recognition rates given the synthetic output are comparable to those given the original videos of the speaker.

Index Terms: visual speech synthesis, expressive and controllable speech synthesis

1. Introduction

For decades, the keyboard and mouse have been the dominant input interfaces to computers and information systems. More recently, touch screen interaction has become ubiquitous, particularly in the domain of mobile devices where screen space is at a premium. Speech interfaces provide an alternative natural interface, however the challenge remains to achieve sufficiently high performance and naturalness in order for most users to view speech as an alternative or complementary input modality. In our quest for the next generation user interfaces, we treat audio-visual communication as a core component. In this work we focus on the synthesis aspect of such a system but we expect future interfaces to be equipped with more cognitive abilities, sensing the user and the environment.

By combining speech and face synthesis, so-called visual speech synthesis, interaction with computers will become more similar to interacting with another person. Expressiveness is an important aspect of any direct interaction between people. The expressions convey additional information to the listener beyond the spoken words. Without expression the interaction may become unnatural and even tedious. It is therefore important that a user interface must be able to handle expressiveness if it is to succeed. While systems exist that produce high quality avatars for neutral speech [7, 10, 12], adding controllable expressions is still challenging [6]. Figure 1 shows our controllable expressive visual text-to-speech synthesiser. The results can be best appreciated in a video, e.g. [13, 14, 15, 16].

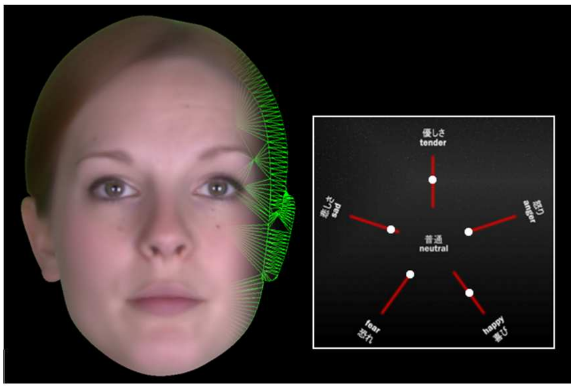


Figure 1: Visual text-to-speech interface. *(left) The face is represented using an extension of a 2D Active Appearance Model, a statistical model of shape and texture. (right) The expression can be controlled with a set of continuous sliders.*

2. System

The key technology behind the talking head is cluster adaptive training (CAT) [2, 4, 5]. It enables flexible control over the expressiveness of both the voice and the face. CAT extends the speech synthesis approach based on hidden Markov models (HMMs). Training of a CAT model requires the collection of a speech and video corpus, where sentences are spoken with a number of different emotions. In addition to a neutral style, our corpus includes angry, happy, sad, tender and fearful expressions. Figure 2 shows an illustration of the facial expressions for each emotion. To model the face, we modify standard active appearance models (AAM) [8], which model shape and texture variation with relatively few parameters. Novel extensions to the AAM incorporate blink and pose states to reduce the artifacts of synthesised sequences [1]. The AAM model parameterises the entire face with 17 dimensional vectors. The speech is parameterised using 45 dimensional Mel-cepstral coefficients, log-F0, 25 Bark-scale band aperiodicities and 19 phase parameters yielding a complex cepstrum representation [3]. First and second time derivatives are added. The face parameters are upsampled using cubic interpolation augmented with first order time derivatives and combined with the speech

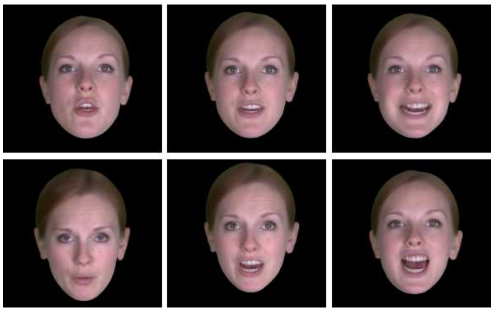


Figure 2: Examples of facial expressions synthesised by the model. *The proposed model can generate different expressions during speech. (top) neutral, tender, happy, (bottom) sad, afraid, angry.*

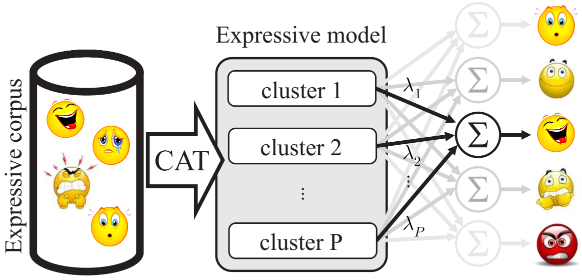


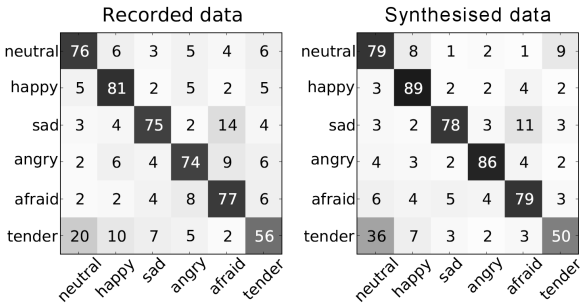
Figure 3: Schematic of Cluster Adaptive Training. *The clusters in CAT form the bases of an emotion-space, and the weights λ_i are the coordinates. Subtle, extreme and new combined expressions are achieved by moving around the space.*

parameters. A CAT model is trained on the combined features. Figure 3 shows the structure of a CAT model consisting of several clusters (six in this case). The basic idea is that the means of the distributions are defined as a linear combination of the means from each cluster. The clusters and the corresponding interpolation weights required for each emotion are all learned as part of the CAT training algorithm. By combining the clusters using the weights learned during training, we can produce the speech and face parameters for each of the expressions in the training set. The interpolation weights can be interpreted as a “space of emotions” with each of the clusters as the bases of the space. By moving around this space we can control the expressiveness in the synthesised speech and face model and so we can produce attenuated or exaggerated versions of the emotions and to “mix” together different emotions to create new emotions not seen in the training. For example, a combination of fear and anger and increased speed produces a panic expression. By training on both speech and facial parameters we guarantee that the face will always be synchronised with the speech, both in terms of lip movement and the chosen emotion (although it is still possible to decouple the two, e.g. sad speech with angry face). The resulting visual speech synthesis is driven entirely from input text and the emotion weight parameters.

3. Performance

The system has been evaluated in two ways [1]:

Emotion recognition study. Results of an emotion recognition experiment are shown in figure 4. Twenty people were asked to identify the emotions depicted in the original audio/video and the synthetic audio/video. The average recognition rates were 73% for the original audio/video, 77% for synthetic audio/video. Recognition rates for synthetic data are comparable, even slightly better than the original footage. This may be due to the stylisation of the emotion in the synthesis.



Recorded data

	neutral	happy	sad	angry	afraid	tender
neutral	76	6	3	5	4	6
happy	5	81	2	5	2	5
sad	3	4	75	2	14	4
angry	2	6	4	74	9	6
afraid	2	2	4	8	77	6
tender	20	10	7	5	2	56

Synthesised data

	neutral	happy	sad	angry	afraid	tender
neutral	79	8	1	2	1	9
happy	3	89	2	2	4	2
sad	3	2	78	3	11	3
angry	4	3	2	86	4	2
afraid	6	4	5	4	79	3
tender	36	7	3	2	3	50

Figure 4: Emotion recognition study. *Recognition rates in percent for real video cropped to the face and for the synthesised avatar. In each case 10 sentences in each emotion were evaluated by 20 people.*

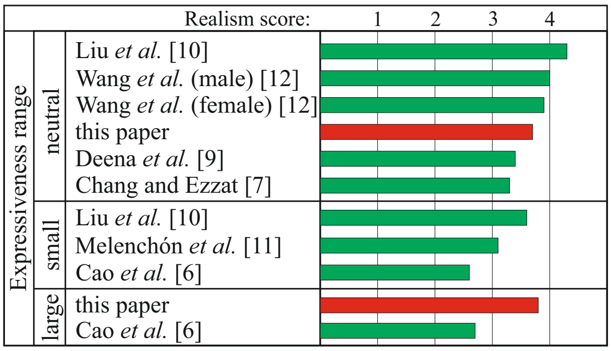


Figure 5: Comparative user study. *Users rated the realism of sample sentences generated using different systems on a scale 1 to 5 where higher values are more realistic. The expressiveness range refers to the level of emotions evaluated.*

Comparison with other talking heads. To compare our talking head with other systems, users were asked to rate the realism of samples of synthesised sentences on a scale of 1 to 5: where 1 is “completely unreal” and 5 is “completely real”. Figure 5 shows the results of the evaluation for different levels of expressiveness, averaged over 100 judgments. For neutral expressions, our avatar performs comparably to others, and in the large expressive range it achieves the higher score.

4. Conclusion

We have successfully developed a highly controllable, expressive avatar which is driven by text and emotion weights. The emotion weights are adjusted using some continuous sliders. The system then synthesises both the speech and face with the desired combination of emotions. Subjective tests show that our system matches human recordings in emotion recognition, and sometimes beats this gold-standard. This system could provide a launch pad for the next generation of user-interfaces which will be endowed with cognitive abilities to convey and perceive emotions in communication.

5. Acknowledgements

We would like to thank Dr. Christopher Newell, School of Arts and New Media, University of Hull, UK and Prof. Paul Barker, Central School of Speech and Drama, University of London, UK for their help and advice on the data collection process.

6. References

[1] Anderson, R., Stenger, B., Wan, V., Cipolla, R.,"Expressive Visual Text-To-Speech Using Active Appearance Models", CVPR 2013, Portland USA, 2013.

[2] Latorre, J., Wan, V., Gales, M. J. F., Chen, L., Chin, K. K., Knill, K. and Akamine, M.,"Speech factorization for HMM-TTS based on cluster adaptive training", in Interspeech 2012, Portland USA, 2012.

[3] Maia, R., Akamine, M., Gales, M. J. F., "Complex cepstrum as phase information in statistical parametric speech synthesis", ICASSP, pages 4581–4584, 2012.

[4] Wan, V., Latorre, J., Chin, K. K., Chen, L., Gales, M. J. F., Zen, H., Knill, K. and Akamine, M. ,"Combining multiple high quality corpora for improving HMM-TTS", in Interspeech 2012, Portland, USA, 2012.

[5] Zen, H., Braunschweiler, N., Buchholz, S., Gales, M. J. F., Knill, K., Krustulović, S. and Latorre, J., "Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization", IEEE Transactions on Audio, Speech & Language Processing, 2011.

[6] Cao, Y., Tien, W., Faloutsos, P. and Pighin, F., "Expressive speech-driven facial animation", ACM TOG, 24(4):1283–1302, 2005

[7] Chang, Y., and Ezzat, T., "Transferable videorealistic speech animation", in SIGGRAPH, pages 143–151, 2005

[8] Cootes, T. F., Edwards, G. J., and Taylor, C. J., "Active appearance models", TPAMI, pages 484–498, 1998

[9] Deena, S., Hou, S., and Galata, A., "Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model", ICMI-MLMI, pages 1-8, 2010.

[10] Liu, K., and Ostermann, J., "Realistic facial expression synthesis for an image-based talking head", in International Conference on Multimedia & Expo, pages 1–6, IEEE 2006.

[11] Melenchón, J., Martínez, E., De la Torre, F., and Montero, J., "Emphatic visual speech synthesis", Trans. Audio, Speech and Lang. Proc., 17(3):459-468, 2009.

[12] Wang, L., Han, W., Soong, F., and Huo, Q., "Text driven 3D photo-realistic talking head", Interspeech, pages 3307-3308, 2011.

[13] BBC News, "Zoe – Cambridge's emotional talking head", 19 March 2013.

[14] Sky News, "Red Dwarf talking head could be a future PA", 19 March 2013.

[15] Reuters, "Meet 'Zoe', the avatar with attitude", 31 March 2013.

[16] Today, "Text better with a face to go with your message", 27 May 2013.