

PREVIS: A PERSON-SPECIFIC REALISTIC VIRTUAL SPEAKER

Javier Melenchón Maldonado, Francesc Alías Pujol, Ignasi Iriondo Sanz*

La Salle School of Engineering, Ramon Llull University
Pg. Bonanova 8, 08022 Barcelona, Spain
{jmelen, falias, iriondo}@salleURL.edu

ABSTRACT

This paper describes a 2D realistic talking face. The facial appearance model is constructed with a parameterised 2D sample based model. This representation supports moderated head movements, facial gestures and emotional expressions. Two main contributions for talking heads applications are proposed. First, the image of the lips is synthesized by means of shape and texture information. Secondly, a nearly automated training process makes the talking face personalization easier, due to the use of mouth tracking. Additionally, lips are synchronized in real time with speech that is generated using a SAPI compliant text-to-speech engine.

1. INTRODUCTION

Speech can be considered as a multi-modal signal with a pair of correlated audial and visual modes [1]. The combined use of both modes improves the message perception. The multi-modal synthesis involves an added value in many applications, for instance: a) it increases the understanding rate, especially in noisy environments, b) it enhances accessibility [2] and human computer interaction (HCI) [3][4], and c) it decreases the bit-rate in coding schemes (i.e. MPEG4) [5]. It will also be used in automatic dubbing [6] and realistic avatar animation [7] in the near future.

In the text-to-speech framework, talking heads can be defined as animated faces lip-synchronized with synthetic speech [8]. Their visual mode is usually represented by a 2D or 3D head model. Synthesising a photo-realistic talking head is technically difficult because the human eye is very sensitive to every small deviation from its expected appearance and behaviour.

The 3D modelling of realistic talking heads has been used over the last three decades in attempts to represent efficiently 3D appearances of faces. The 3D approach

began with the pioneering work of Parke [9] and has continued to the present day. Recent work ranges from techniques based on physical structure modelling [7][10] to others based on a texture mapping over a 3D mesh [11][12]. Nowadays, these 3D models are being improved to achieve more photo-realism [13].

The realistic 3D representation of the time-varying facial appearance is computationally expensive because of its high degree of deformability. Moreover, this representation still produces a synthetic look due to the complexity of modelling some elements such as skin and hair [14].

Since the beginning of the last decade, 2D models have been applied to create realistic talking heads. This 2D representation is based on a finite set of images obtained from a human face. One of the first attempts to animate images was the 2D morphing model of Beier and Neely [15], which presented some difficulties related to finding the displacements of the image pixels. Later, the automatic morphing proposed by Ezzat and Poggio [4] solved this problem by means of optical flow methods. However, this process is very sensitive to appearance changes and its image analysis and synthesis require a very high computational cost. Recently, Noh and Neumann [16] have proposed a parameterised morphing model based on radial basis functions (RBF) with manual labelling of feature points on each training image. Other methods are based on a large audiovisual corpus, like the one presented in [6], where each sequence of its corpus is a triphone segment. Triphone-based methods may achieve good video-realism but they require a huge unit database. Another proposal, presented by Cossatto and Graf [3], uses a modular representation of the different facial elements, where mouth shapes are parameterised in a low dimensional space. Nevertheless, manual supervision is required and jerky movements can appear if there are not enough lip images stored in the database.

Nowadays, several advances in computer vision have led to improvements in both 2D and 3D facial models, allowing more accurate talking heads. Following recent work on face tracking and modelling, this paper introduces a new talking face based on a low dimensional parametric 2D facial model (section 2) that can be easily

* This research has been supported by the D.U.R.S.I. of the Generalitat de Catalunya under the grant 2000FI-00679

personalized. Section 3 describes the main features and some results of the implemented HCI application.

2. THE FACIAL MODEL

The proposed 2D facial model is a parameterised sample-based model that uses a set of real human images. The application of this model allows the synthesis of lip appearances, facial gestures and movements related to different emotions.

2.1. The generation of visemes and expressions

To synthesize every frame of our talking head, a background face and a viseme are superimposed. A viseme can be defined as a mouth shape articulating a phoneme. The background face incorporates the non-verbal part of communication (i.e. emotional expressions and facial gestures). We consider six basic human facial expressions of emotion: fear, happiness, sadness, surprise, anger and disgust, following the work of Ekman [17]. Gestures can be seen as the facial movements that emphasize or substitute the verbal part of communication, for instance eye blinking, head nods and shakes.

The overlapping of image background and viseme is achieved by a pixelwise blend (see figure 1) using a weighted window applied to the pixels of the mouth area.

$$w(x, y) = \frac{1}{M} \sqrt[n]{|x|^n + |y|^n} \quad (1)$$

$$I_f(x, y) = I_b(x, y) \cdot w(x, y) + I_v(1 - w(x, y)) \quad (2)$$

A point (x, y) represents the position of a pixel from the lips centroid. The weighted window w (see equation 1) is normalised by a maximum value M and its shape varies according to the n order (typically 1, 2 or 3). In equation 2, I_b is the background image (or a value of the skin colour), I_v represents the image of a viseme and I_f is the full image.

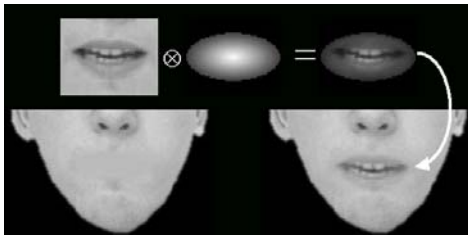


Figure 1. Lips and background image overlapping process.

The model allows the representation of different facial expressions. Each one specifies which sets of background images (gestures) and visemes will be used. For the moment the facial expressions tested with our model are: neutral, happy and sad.

The visemes are stored as a set of bases and coefficients of the principal components and an associative memory is used to translate phonemes to visemes; this conversion depends on the facial expression to be displayed. The image of a viseme is made up of shape and texture, as depicted in figure 2.

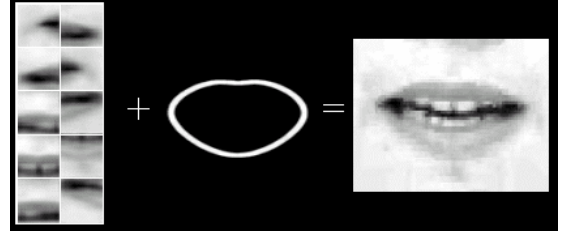


Figure 2. Composition process of a viseme.

We approximate the face by a plane, so the 3D movements of our talking head are recovered with a parametric affine model. Its time-varying parameters describe a cubic spline trajectory that models accurately the acceleration of physical objects [18]. The model only synthesizes moderate head movements like scaling, translation and some kind of rotation. The sequence of movements and gestures are randomly generated at synthesis time.

2.2. Parameter extraction

The synthesis process will use two orthogonal bases to create the appearance (U^s) and shape (U^t). These bases are obtained from performing the Principal Component Analysis (PCA) [19] of the images of the actor's lips [20][21] in a reference sequence; this operation conforms the training process of [21]. The image of a viseme is composed of its shape, modelled by a mean shape (\underline{s}_{mean}) added to a linear combination (\underline{b}_i coefficients) of U^s , and its texture, obtained from a mean texture (\underline{t}_{mean}) plus another linear combination (\underline{c}_i coefficients) of U^t :

$$\begin{aligned} \underline{s}_i &= \underline{s}_{mean} + U^s \underline{b}_i \\ \underline{t}_i &= \underline{t}_{mean} + U^t \underline{c}_i \end{aligned} \quad (3)$$

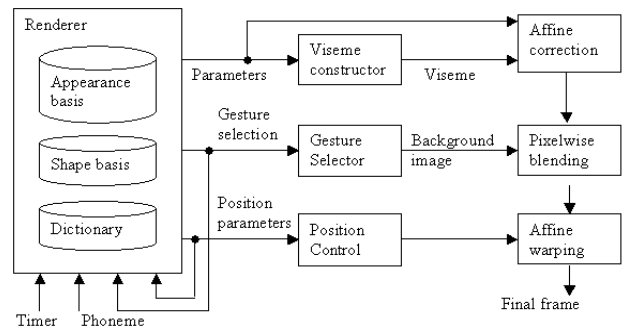


Figure 3. Elements involved in the frame synthesis

These projections (\underline{b}_i , \underline{c}_i) are obtained with the Eigenfiltering for Flexible Eigentracking (EFE) [21], which is similar to the approach presented by Cootes and Taylor [20] called Active Appearance Models (AAM). Unlike AAM, EFE uses a robust function to avoid the influence of outliers in the solution and it is able to achieve sub-pixel precision. EFE outputs the projection of the lips (\underline{b}_i , \underline{c}_i) for every frame of any actor's sequence to be tracked using the previously learned \underline{S}_{mean} , \underline{L}_{mean} , U^s and U^t . The human subject has total freedom of 2D movement during the training process because EFE is able to track all his 2D variations of position. Moreover, it is very robust to illumination changes and outliers, such as shifting a finger in front of the mouth.

In addition, this method takes advantage of the PCA parameterisation and the EFE tracking to simplify the synthesis of non-stored images (i.e. coarticulations or transitions between emotions, see figure 4) and the personalization training process.



Figure 4. Synthetic coarticulation [a]-[δ]

3. THE APPLICATION

We have implemented an audiovisual interface, called PREVIS (Person-specific REalistic Virtual Speaker), which integrates the previously explained facial model with a text-to-speech system. This kind of application, which enhances human-computer communication, will be a useful HCI tool in the future.

In the following paragraphs, the key points of the application are described. It is important to point out that the interface consumes very few system resources, so it can run concurrently with other programs.

3.1. Person-specific training process

One of the most interesting features of PREVIS is the simplicity of its personalizable training procedure, compared with other similar methods [16] allowing person-specific applications. This process is achieved by means of the EFE algorithm [21] providing freedom of 2D movements of the actor.

In our system, a new talking face can be generated by following these four steps: first, the new actor sits down in front of the camera and utters a defined list of words with the desired facial expression; secondly, the EFE tracking utility is run off-line on a PC, which only requires the hand labelling of the first image; thirdly, the correspondence between visemes and phonemes is

specified; and finally, the actor's configuration files (an associative memory and a 20 image sequence per expression) are generated with a tool developed under MATLAB®. Although it is not a fully automatic process, it simplifies the tedious work of creating a new face which is necessary in other approaches [16], so new faces, expressions and languages can be easily added.

3.2. The naturalness of our realistic talking face

The lack of head movements is an important drawback when trying to achieve a natural talking face [18]; therefore we have included some moderate facial displacements coupled with lip animation, whose variability and rhythm can be chosen.

Moreover, human communication is always related to the speakers' emotional state. In this way, the visemes, the background elements (i.e. eyebrows) and the speech prosody are adjusted to the talker's desired state of mind (happiness, sadness or neutrality, manually selected for the moment) at synthesis time.

3.3. How does our talking face speak?

Our application is based on a text-to-speech (TTS) system, instead of being driven by audio input [6]. TTS is able to generate any desired utterance from text. Furthermore, PREVIS has been designed to use SAPI-compliant TTS engines, so it is easy to incorporate new voices and languages. At present, it has been tested with Spanish and Catalan engines developed in La Salle School of Engineering [22][23] and the default English engine distributed with the Microsoft Speech SDK® [24].

PREVIS can easily incorporate a new language: every phoneme notification of the SAPI-compliant engine has to be correctly associated with its related viseme. This process only involves the time cost of changing almost 40 parameters.

The audio-visual synchronization is achieved by means of a client-server scheme (see figure 5), where the TTS engine operates as the server and the application is the client. First, the client sends the text to be synthesized to the server, by calling a SAPI member function. This function allows some embedded *control tags* to modify different speech features [24]. Then, the server synthesises the utterance and it synchronizes the output signal with the phoneme notification by the SAPI interfaces. Finally, from these data, the client generates the associated viseme with the appropriate emotional context, building the talking face to be displayed.

Users are fairly sensitive to audio-visual desynchronization, so visual notifications must be returned promptly and 25 frames per second are necessary to guarantee at least one image per phoneme. Moreover, the engine only notifies the application of the current

phoneme, so the coarticulation between visemes is not implemented.

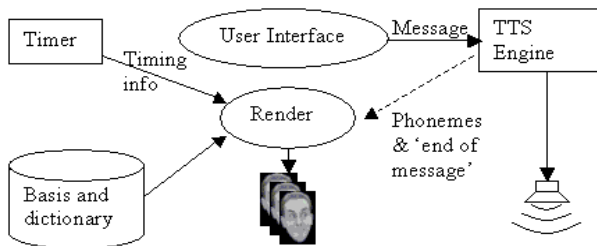


Figure 5. Block diagram of the TTS-application communication

3.4. Results

Next, some screenshots of the application are shown. More audiovisual captures of the application may be viewed by accessing our World Wide Web demonstration page at <http://www.salleurl.edu/~jmelen/demo.html>.



Figure 6. Some screenshots of PREVIS.

4. CONCLUDING REMARKS

The main outcome of this work is the development of a talking face application with low computational cost, which is based on SAPI compliant TTS systems and our new parameterised sample-based facial model. This 2D model improves the realism of the head including moderated movements and expressions and it allows a person-specific training process.

Future plans include the development of a more realistic talking face and the implementation of a fully automatic training process [25]. In order to achieve these goals, we are going to add colour to the images, to include coarticulation between visemes and to implement more expressions and gestures. Furthermore, EFE will be applied to different facial elements in addition to lips to obtain a more versatile talking face. Finally, we are going to evaluate the influence of adding lip synchronization to speech in the comprehension of the messages by means of a perceptual study.

5. REFERENCES

[1] S. Alan, "A Facial Model and Animation Techniques for Animated Speech". *Partial fulfilment of the Requirements for the degree Dr. of Philosophy*. The Ohio State University 2001.

- [2] J. Beskow, M. Dahlquist, B. Granström et al, "The Teleface Project Multi-Modal Speech-Communication For The Hearing Impaired". *Proc Eurospeech '97*.
- [3] E. Cossatto, H. Graf, "Sample-Based Synthesis of Photorealistic Talking Heads", *Proceedings of Computer Animation '98*, pp. 103-110, Philadelphia, Pennsylvania, 1998.
- [4] T. Ezzat, T. Poggio, "Facial Analysis and Synthesis Using Image-Based Models", *Proc 2nd Int. Conf. On Automatic Face and Gesture Recognition*, IEEE CS Press, 1996, pp. 116-121.
- [5] J. Ostermann, A. Puri, "Natural and Synthetic Video in MPEG-4", *Proc. of ICASSP 1998*.
- [6] C. Bregler, M. Covell, M. Slaney, "Video Rewrite: Driving Visual Speech with Audio", *Proc. SIGGRAPH'97*, pp. 353-360.
- [7] C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, P. Cost, "Modelling an Italian Head", *Audio-Visual Speech Processing*, Scheelsminde, Denmark 2001.
- [8] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices". *Nature* 264:746-248, 1976.
- [9] Parke, F.I. "Computer generated animation of faces". University of Salt Lake City. 1972.
- [10] D. Terzopoulos, K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol 15/6, pp. 569-579, 1993.
- [11] J. Ostermann, L.S. Chen, and T.S. Huang, "Animated Talking Head with Personalized 3D Head Model", *Workshop of Multimedia Signal Processing*, pp. 274-279, 1997, Princeton, US.
- [12] T. Kurakate, F. Garcia, H. Yehia, E. Vatikioitis-Bateson, "Facial Animation from 3D Kinematics", *ASJ*, 1997, Sapporo.
- [13] S. Morishima, "Face Analysis and Synthesis", *IEEE Signal Processing Magazine*, vol. 18, No. 3, pp. 26-34, 2001.
- [14] Y. Lee, D. Terzopoulos, and K. Waters. "Realistic modeling for facial animation". In *SIGGRAPH '95 Proceedings*, pages 55-62, Los Angeles, California, 1995.
- [15] T. Beier, S. Neely, "Feature-Based Image Metamorphosis". *Proc. of SIGGRAPH '92*, pp. 35-42, Chicago, IL 1992.
- [16] J. Noh and U. Neumann. "Talking Faces". *IEEE International Conference on Multimedia and Expo (II) 2000*, pp. 627-630, New York, USA.
- [17] P. Ekman, W. Friesen, "Facial Action Cosing System". *Consulting Psychology Press Inc.*, Palo Alto, California 1978.
- [18] G. Maestri, "Digital Character Animation". *New Riders Publishing*, 1996.
- [19] I. T. Jolliffe. *Principal Component Analysis*. New York Springer-Verlag, 1986.
- [20] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *5th ECCV*, 1998.
- [21] F. De la Torre, J. Vitrià, P. Radeva, J. Melenchón, "Eigenfiltering for Flexible Eigentracking" *15th International Conference on Pattern Recognition (ICPR)*. Barcelona, 2000.
- [22] R. Gaus, I. Iriundo, "Diphone based Unit Selection for Catalan TTS Synthesis". *TSD 2000*. Brno, Czech Republic.
- [23] URL: <http://cepheus.salleurl.edu/www/ttsweb.htm>
- [24] URL: <http://microsoft.com/speech/>
- [25] F. De la Torre. "Automatic Learning of Appearance Face Models", *2nd International Workshop on RATFG-RTS*. Vancouver, Canada 2001.