Expressive Talking Head for Interactive Conversational Systems

Paula Dornhofer Paro Costa and José Mario De Martino
School of Electrical and Computer Engineering, Dept. of Computer Engineering and Industrial Automation, University of Campinas, Campinas, São Paulo, Brazil

1 STAGE OF THE RESEARCH

The present doctoral research project is being developed in the following sequential phases:

1. Creation of a motion capture data and audiovisual expressive speech database for Brazilian Portuguese;
2. Study of sample-based synthesis techniques suitable for the reproduction of expressive speech;
3. Development of an expressive talking head for Brazilian Portuguese;
4. Proposal of a strategy to generalize the synthesis methodology to a compact database of samples captured under partially controlled conditions.

The project is entering into its third phase, moving towards the proposal of an expressive speech synthesis methodology with a pilot implementation being developed for Brazilian Portuguese.

2 OUTLINE OF OBJECTIVES

The objective of this project is to develop an expressive speech facial animation synthesis methodology capable of improving the videorealism of the current state-of-the-art expressive talking heads in terms of the range of the expressed emotional states and the ability of reproducing non-verbal communication signaling, with the ultimate goal of creating animated faces capable of inspiring user thrust and empathy .

We aim at applications where intelligent systems are capable of emulating our natural and intuitive face-to-face communication mechanisms in which the talking head act as an embodied conversational agent (ECA), (Figure 1).

Objectives that run in parallel are: to obtain a pilot implementation of a Brazilian Portuguese expressive talking head and to contribute with an comprehensive audiovisual expressive speech corpus for this language.

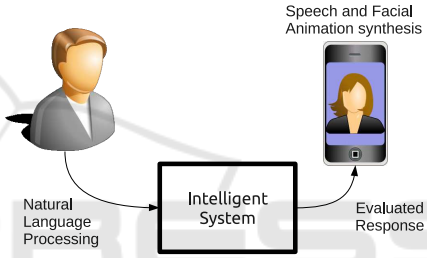

Figure 1: Intelligent system face-to-face communication metaphor.

3 RESEARCH PROBLEM

The automated synthesis of videorealistic talking heads capable of expressing emotions remains a challenging problem in computer graphics. Even the most sophisticated models capable of rendering impressive high quality face images are not capable of inspiring user empathy and thrust without the appropriate modeling of the non-verbal communication signals, a task that still requires human intervention. On the other hand, there is a great demand for expressive talking heads for applications like virtual assistants, tutors, game characters, newscasters, social agents and for controlled experiments in psychology and behavioral sciences.

For centuries, painters, cartoonists, and animators have being able to reproduce realistic images of expressive faces following rules learned from the exhaustive observation of the human behavior. Similarly, the problem of developing computer automated videorealistic facial animation can be stated as the problem of making the machine learn how the humans behave while expressing different emotional and cognitive states and deriving models to reproduce them – a typical supervised machine learning task.

This doctoral project concentrates on two main aspects of this problem: the computational model for emotions and the visual modeling of the face.

Categorical emotion models, like the “big six” (Ekman, 1972) universally recognizable facial expressions (anger, disgust, fear, happiness, sadness,

Dornhofer Paro Costa P. and De Martino J..
Expressive Talking Head for Interactive Conversational Systems.

surprise) have been widely applied to the development of expressive avatars. However, we argue that this modeling lacks completeness when the objective is the implementation of interactive embodied conversational agents. In order to capture user empathy, for example, a virtual airport assistant informing a flight cancellation, should not express any of the "big six" plain emotions but a set of more complex facial expressions capable of dealing with the user frustration, indicating that the system recognizes the flight cancellation as an undesirable event. For this purpose, "appraisal" models, that take into consideration the evaluation process that leads to an emotional response, seem to provide a more embracing characterization of emotions. In particular, we adopt the model proposed by Ortony, Clore and Collins (OCC model) since it presents a concise but comprehensive vocabulary of 22 emotions that arises as reactions to events, agents or objects (Ortony et al., 1988).

Another important question is how to synthesize photorealistic appearances and to reproduce the dynamics of the speech combined to the expression of emotions. Put in other words, how to delude human observers: specialists trained since the born to detect the smallest variations on the signals conveyed by the voice, the face and the body. In this work we explore the image-based, or 2D, synthesis technique as a mean to obtain inherently photorealistic expressive faces, avoiding the typical synthetic looking of model-based (3D) facial animation systems.

4 STATE OF THE ART

Since the pioneering work of Parke (Parke, 1972), many others have contributed with different approaches to improve the level of videorealism of synthetic talking faces: (Bregler et al., 1997), (Ezzat and Poggio, 1998), (Brand, 1999), (Cosatto and Graf, 2000), (Pasquariello and Pelachaud, 2002), (Ezzat et al., 2002).

In the last decade, it can be observed an emerging interest in adding to the synthetic talking heads the capability of expressing emotions.

In (Chuang and Bregler, 2005), for example, the authors focus in the difficulties to edit motion capture data. In their proposal, they take expressionless speech performance as input, analyze the content and modify the facial expression according to a statistical model. The expressive face is then retargeted onto a 3D character using blendshape animation. The paper presents the results of the methodology using as training data three short video sequences including three basic expressions: neutral, angry, and happy.

In (Beskow and Nordenberg, 2005), an expressive 3D talking head is implemented using an MPEG-4 compatible model. An amateur actor was recorded portraying five different emotions (happy, sad, angry, surprised, and neutral) and a Cohen-Massaro coarticulation model was trained for each emotion.

A 3D expressive speech-driven facial animation system is also presented in (Cao et al., 2005). In this system the inputs are the speech to be animated and a set of emotional tags. A high-fidelity motion capture database was built with a professional actor representing five emotions: frustrated, happy, neutral, sad and angry. The motion capture data, together with the timed phonetic transcript of the recorded utterances, were used to construct what the authors call an "anime" graph, where an "anime" corresponds to a dynamic definition of a viseme. The synthesis consists in searching the best path in this graph through the minimization of a cost function that penalizes discontinuity in unnatural visual transitions.

Four emotions (neutral, happy, angry, sad) were captured with a motion capture system in (Deng et al., 2006). The resulting material was used to build a coarticulation model and a visual appearance model that are combined to generate a 3D facial animation at synthesis time.

As novel approach to model emotions in a facial animation system, in (Jia et al., 2011) the authors parameterize eleven emotions (neutral, relax, submissiveness, surprise, happiness, disgust, contempt, fear, sorrow, anxiety and anger) according to the PAD (Pleasure, Arousal, Dominance) dimensional emotion model. In this system, the acoustic features of the speech are used to drive an MPEG-4 model.

More recently, (Anderson et al., 2013) present a 2D VTTS (visual text-to-speech) which is capable of synthesizing a talking head given an input text and a set of continuous expression weights. The face is modeled using an active appearance model (AAM), from a corpus containing six emotions: neutral, tender, angry, afraid, happy and sad.

These works illustrate, through a diverse range of approaches, the challenges imposed by this research problem.

5 METHODOLOGY

5.1 Corpus

In order to study different aspects of expressive speech, ten professional actors, Brazilian Portuguese native speakers, were divided in two types of experiments. The first experiment consisted in asking the

Figure 2: Motion capture.

actors to represent the emotions of the OCC model. For that purpose, 22 recording scripts were designed coherently with the cognitive state description given by the model ensuring that each speech has occurrences of all Brazilian Portuguese context-dependent visemes (visual phonemes), as proposed in (De Martino et al., 2006). Each speech segment was recorded with a neutral expression, followed by the actor representation of the corresponding emotional state. In a second experiment, the objective was to obtain data to investigate how the expressive signals observed in the face are modulated by different personality traits. For that purpose, we adopted the simpler modeling of the "big six" emotions and the actors were asked to represent them with three different extroversion characteristics: shy, neutral/equilibrated and extroverted. In this experiment, the speech segments were the same for all emotions and personalities.

Each experiment consisted of a motion capture session and a video session. During the mocap session, 63 markers distributed over the actor's face/head were tracked by 8 infrared Vicon cameras (Figure 2). In the video session, the actor repeated the same performance in front of a chroma-key background, without markers, makeup, or facial and hair adornments. Both experiments resulted in a set of motion capture and video samples of expressive faces of 4 female actresses and 6 male actors with ages ranging from 20 to 60 years old.

5.2 Visual Face Model

Taking samples from the audiovisual corpus as training data, our methodology involved the exploratory study of the Active Appearance Model (AAM) as a facial synthesis model.

Active appearance modeling is capable of parameterizing an image database of facial expressions in terms of their shapes and texture parameters, making possible, given some constraints, the generation of facial expressions and head poses that were not present in the original database. Besides, applying an analysis by synthesis approach, the derived AAM model can be used as a machine learning tool to analyze the expressive speech dynamics of the original corpus and to derive rules to reproduce it.

A well known problem of AAM, however, is its smoothing effect on final images and the blurred aspect typically observed inside the mouth, a region that experiences a great appearance variation but that has no fiducial anchor points. In this study we explored different setups for deriving AAM model in order to determine the limits of this model given the characteristics of our corpus.

In the typical implementation of AAM, the coordinates of the feature points $(x_1, y_1), ..., (x_k, y_k)$ are organized as a shape vector $s = [x_1, y_1, ..., x_k, y_k]$. Principal component analysis (PCA) is applied to the training base of shape vectors and the shape model is represented by the linear combination of the mean facial shape s_0 and n facial shape eigenvectors:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i \quad (1)$$

After the definition of a shape model, each training facial image is warped to obtain the mean facial shape $\mathbf{s}_0$. The shape-free facial images are transformed in texture vectors $\mathbf{A}$. A second PCA is performed and the texture is modeled as a linear combination of the mean facial appearance $\mathbf{A}_0$ and m eigenfaces:

$$\mathbf{A} = \mathbf{A}_0 + \sum_{i=1}^{m} \lambda_i \mathbf{A}_i \quad (2)$$

When the parametric shape and appearance models are used independently, the model is called an independent-AAM. To remove correlation between shape and texture model parameters and to make the model representation more compact a third PCA can be performed on the concatenated shape and texture parameters. In this case a combined-AAM is obtained.

The study was conducted using as training database of 873 images with semi-automatically 56 labeled feature points (Figure 3). The images in the database cover all the Brazilian Portuguese visemes for each OCC emotion, together with neutral expression samples.

We explored multiple modeling building scenarios in order to evaluate the videorealism attained by each approach:

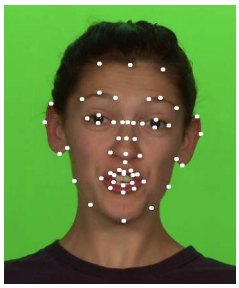

Figure 3: Feature points used to define the facial shape.

Figure 4: Comparing corresponding synthesized frames with two different AAM building scenarios. Repair that in both images the region inside the lips is blurred but the image at the left has greater definition of the teeth and the tongue than the image at the right. The left image was synthesized using a mixed approach: an independent model was used to synthesize the shape, while a combined model was used to synthesize the appearance.

- full face independent-AAM;
- parts based independent-AAM;
- full face combined-AAM;
- parts based independent-AAM.

Multiple masks configurations and a mixed approach (an independent model was used to synthesize the head trajectory and a combined model to generate the facial texture) were also tested (Figure 4).

5.3 Expressive Talking Head Synthesis

Once defined the facial model, further studies are needed to define the methodology to drive the facial animation synthesis. While some approaches derive the parameters to guide the facial animation from the speech, we assume that the speech to be animated could be provided by a TTS, a typical subsystem of an interactive and intelligent computational system.

From this assumption, we aim to develop a synthesis system that receives as inputs the speech to be animated timed phonetic transcription, a tag about the emotion to be animated and a tag about the personality or simply, an intensity modulator of the emotion to be expressed.

A challenging problem regarding this approach is: while the timed phonetic transcription provides information that enables the determination of the dynamics of speech articulatory movements and, consequently, the shape of the face, it does not provide information about the appearance to be assumed. We plan to solve this problem through regression models that are able to correlate the animation articulatory targets with the appearance of samples from the image database.

5.4 Generalization and Evaluation

As a part of our research, we aim to derive a synthesis strategy that could be generalized for smaller image database with fewer samples of facial expressions and possibly captured under uncontrolled conditions. This poses the problem of learning the trajectories of facial expressions transitions in the original database and trying to reproduce such trajectories in an image space with missing samples, or “holes” in the path. Once again, we plan to use regression models to fill in the blanks.

Finally, an essential and challenging step of our methodology is the subjective evaluation process, since there are no widely accepted evaluation protocols of expressive talking heads. In this case, we plan to to assess two main aspects of the synthesized animations: their level of videorealism and their capability of expressing a range of predefined emotions with fidelity.

6 EXPECTED OUTCOME

As described in the preceding sections this project expects to propose an innovative synthesis methodology for expressive talking heads. We also aim to develop a Brazilian Portuguese pilot implementation and use it to drive audiovisual expressive speech evaluations.

The implementation to be pursued is illustrated in Figure 5, where the facial animation system receives as input: parameters driven from the speech (such as the articulatory target or acoustic features that can drive the animation), a tag of emotion to be synthesized (that can be informed, for instance, by an intelligent system) and a personality trait parameter (or alternatively, a mood parameter) that modulates the visual synthesis.

As a secondary outcome, we expect to make the Brazilian Portuguese expressive corpus publicly

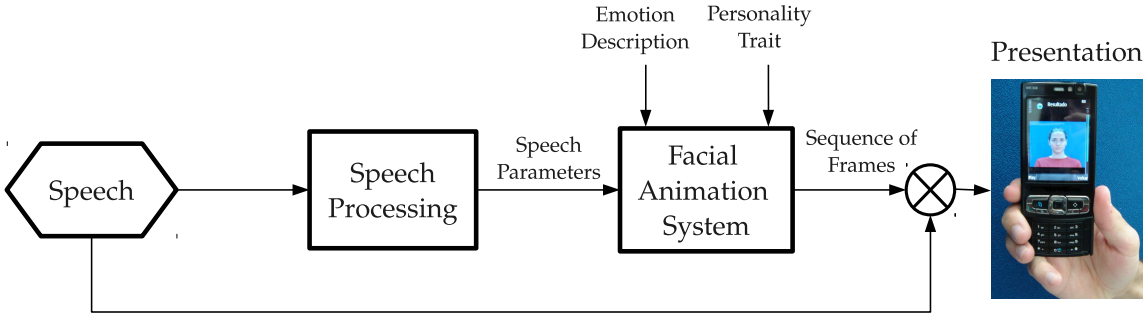

Figure 5: Expressive Talking Head System.

available as an organized database of emotions, enabling the development of other prospective works. Finally, we also expect to propose a generalization synthesis methodology to make possible the creation of simple expressive talking heads from a few images of a real face.

ACKNOWLEDGEMENTS

This work is supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 141366/2010-9.

REFERENCES

Anderson, R., Stenger, B., Wan, V., and Cipolla, R. (2013). Expressive Visual Text-to-Speech Using Active Appearance Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3382–3389.

Beskow, J. and Nordenberg, M. (2005). Data-driven synthesis of expressive visual speech using an MPEG-4 talking head. In *INTERSPEECH*, pages 793–796. ISCA.

Brand, M. (1999). Voice puppetry. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Bregler, C., Covell, M., and Slaney, M. (1997). Video Rewrite: driving visual speech with audio. In *SIGGRAPH*, pages 353–360.

Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. H. (2005). Expressive speech-driven facial animation. *ACM Transactions on Graphics*, 24(4):1283–1302.

Chuang, E. and Bregler, C. (2005). Mood swings: expressive speech animation. *ACM Transactions on Graphics*, 24(2):331–347.

Cosatto, E. and Graf, H. P. (2000). Photo-Realistic Talking-Heads from Image Samples. *IEEE Transactions on Multimedia*, 2(3):152–163.

De Martino, J. M., Magalhães, L. P., and Violaro, F. (2006). Facial animation based on context-dependent visemes. *Computer & Graphics*, 30:971–980.

Deng, Z., Neumann, U., Lewis, J., Kim, T.-Y., Bulut, M., and Narayanan, S. (2006). Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1523–1534.

Ekman, P. (1972). Universals and Cultural Differences in Facial Expressions of Emotion. In *Proceedings of the Nebraska Symposium on Motivation*, number 19, pages 207–282. Lincoln University of Nebraska Press.

Ezzat, T., Geiger, G., and Poggio, T. (2002). Trainable videorealistic speech animation. In *SIGGRAPH*, pages 388–398.

Ezzat, T. and Poggio, T. (1998). MikeTalk: A Talking Facial Display Based on Morphing Visemes. In *CA*, pages 96–102.

Jia, J., Zhang, S., Meng, F., Wang, Y., and Cai, L. (2011). Emotional Audio-Visual Speech Synthesis Based on PAD. *IEEE Transactions on Audio, Speech & Language Processing*, 19(3):570–582.

Ortony, A., Clore, G., and Collins, A. (1988). *Cognitive Structure of Emotions*. Cambridge University Press.

Parke, F. I. (1972). Computer generated animation of faces. In *ACM'72: Proceedings of the ACM annual conference*, pages 451–457, New York, NY, USA. ACM Press.

Pasquariello, S. and Pelachaud, C. (2002). Greta: A simple facial animation engine. *Soft Computing and Industry*, pages 511–525.