

# On Combining the Facial Movements of a Talking Head

T.D. Bui<sup>1</sup>, D. Heylen<sup>2</sup>, A. Nijholt<sup>2</sup>, M. Poel<sup>2</sup>

<sup>1</sup> College of Technology, Vietnam National University, Hanoi, Vietnam

<sup>2</sup> Human Media Interaction, University of Twente, Enschede, The Netherlands

## Abstract

We present work on Obie, an embodied conversational agent framework. An embodied conversational agent, or talking head, consists of three main components. The graphical part consists of a face model and a facial muscle model. Besides the graphical part, we have implemented an *emotion model* and a mapping from emotions to facial expressions. The *animation* part of the framework focuses on the combination of different facial movements temporally. In this paper we propose a scheme of combining facial movements on a 3D talking head.

## Keywords

Emotions, Emotional Facial Expressions, Facial Movement Combination, Embodied Conversational Agents.

## 1 Introduction

The combination of research in the field of computer graphics, autonomous agents, and speech and language technology has led to the development of embodied agents [2,3,20,21]. The emerging technology of embodied agents can realize different promising applications including human-like interfaces to improve the interaction between human and computer; simulated virtual characters for different applications such as entertainment, education, and the like; and believable animated characters to increase the interestingness of computer games.

In this paper, we present work on Obie, an embodied conversational agent framework (see *Figure 4*). An embodied conversational agent, or talking head, consists of various components. The graphical part of the framework consists of a face model and a facial muscle model which allow realistic facial expressions to be produced in real-time on a standard pc. Besides the graphical part, we need a system that accounts for the actions (dialogue) and emotions of the agent. We have implemented an emotion model and a mapping from emotions to facial expressions. The animation part of the framework focuses on the combination of different facial movements temporally. We concentrate on the dynamic aspects of facial movements and the combination of facial expressions in different channels that are responsible for different tasks.

Section 2 discusses the emotion model and the mapping from emotions to emotional facial expressions in the framework. Section 3 focuses on the animation part of the framework. Some illustrations are presented in Section 4.

## 2 Emotions and emotional facial expressions

Emotions play an essential role in making embodied conversational more believable [19]. For our agents to experience emotions, we have implemented an emotion model in our framework [5]. Emotions can be expressed in

different ways: facial expressions, gestures, speech, etc. Among these, facial movements play a very important role in interpreting emotions. We have therefore proposed a fuzzy rule based system to map an emotion state to emotional facial expressions [4].

Probably, the most popular descriptive work on how emotions are expressed on faces is done by Ekman and Friesen [10]. It discusses how several emotions as well as their blends are displayed on the face. We have based ourselves on this work to map emotion representations onto the contraction level of facial muscles. We focus on two aspects of generating emotional facial expressions. First, we take into account the continuous changes in expressions of an emotion depending on the intensity by which it is felt. Secondly, we propose a way to specify combinations of expressions due to more than one emotion, i.e., blends, in accordance with the literature mentioned. We have used a fuzzy rule-based system, which allows us to incorporate qualitative as well as quantitative information. Fuzzy rules can capture descriptions which are described in natural language as well as vague concepts like “slight sadness”, “more intense sadness”, etc. Moreover, the fuzzy rule-based approach can assure the smooth mapping between emotions and facial expressions.

Following Ekman and Friesen [10], we consider the following six so-called “basic “ emotions: **Sadness, Happiness, Anger, Fear, Disgust** and **Surprise**. These are claimed to be associated with prototypical facial expressions that are said to be universal in that research suggests that they are associated consistently with these across different cultures. Ekman and Friesen also describe in detail what the expressions for these emotions and certain blends look like. Emotion feelings may differ in intensity. In [10], it is pointed out how for each of the basic emotions the expression can differ depending on the intensity of the emotion. It is therefore important for us to build our system on a representation that takes intensities into account. We have used their descriptions as the basis for our fuzzy rules.

## 3 The facial animation

The Obie framework is mainly concerned with the facial expression behavior of an embodied agent. The expressions are displayed on a face model, which we developed to allow high quality and realistic facial expressions [6]. The face model is implemented together with a muscle model that produces realistic deformation of the facial surface, handles multiple muscle interactions correctly and produces bulges and wrinkles in real-time.

It is particularly important for the framework to deal with the problem of combining different facial movements temporally. Existing facial animation systems usually

combine these movements by adding them up [1,8,15,17]. Unfortunately, when and how a movement appears and disappears, and how co-occurring movements are integrated (co-articulation effects, for instance) are difficult to quantify [12]. In addition, the problems of overlaying and blending facial movements in time, and the way felt emotions are expressed in facial activity during speech, have not been studied thoroughly [16]. Thus, the activity of human facial muscles is far from simply additive. A typical example would be smiling while speaking. The Zygomatic Major and Minor muscles contract to pull the corner of the lip outward, resulting in a smile. The viseme corresponding to the phoneme “@U” in the word “Hello” requires the contraction of the lip funneler Orbicularis Oris, which drives the lips into a tight, pursed shape. However, the activation of the Zygomatic Major and Minor muscles together with the lip funneler Orbicularis Oris would create an unnatural movement. We call these “conflicting” muscles. The activation of a muscle may require the deactivation of other muscles [10]. Depending on the priority of the tasks to be performed on the face, appropriate muscles are selected for activation. In most of the cases, the visual speech has higher priority than the smile. In some situations, the smile may have higher priority than the visual speech, for instance, when the subject is too happy to utter the speech naturally.

Based on psychological literature [10,9] and research in facial movement measurement and synthesis [13,18], we have proposed a scheme of combining facial movements on a 3D talking head. There are several types of movements, such as conversational signals, emotion display, etc. We call these channels of facial movement. We concentrate on the dynamic aspects of facial movements and the combination of facial expressions in different channels that are responsible for different tasks.

Although facial movements happen continuously, most of them are known from electromyography (EMG) studies to occur in distinct phases [13]. The flow of movements can then be broken up into so-called “atomic” movements. We define an “atomic” movement as a group of muscle contractions that share the same function (e.g., conversation signal, emotion display), start time, end time, and onset and offset duration. Each “atomic” facial movement belongs to a specific channel, which contains only non-conflicting movements. Atomic movements within a channel occur sequentially, although they may overlap each other at their beginning and ending. This classification is also based on the function of the movements [9].

In our system, we distinguish six channels:

- Channel 1 contains **manipulators**, which are movements to satisfy biological requirements of the face. In our system, we consider eye blinking to wet the eyes as manipulators. These movements are random rather than repeated with fixed rate as in [18]. The random eye blinking is generated based on the algorithm proposed in [14].
- Channel 2 contains **lip movements** when talking (represented as viseme segments). Lip movements are generated from the text that is going to be spoken by the talking head. The text is converted to phoneme segments (phoneme with temporal information --

starting and ending time). The phonemes are converted to corresponding visemes. Each viseme is equipped with a set of dominance functions of parameters participating in the articulation of the speech segment. We use dominance functions from [8] for each viseme segment.

- Channel 3 contains **conversational signals**. These are movements to accentuate or emphasize speech, or to provide feedback from a listener. They can occur on pauses due to hesitation or to signal punctuation marks (such as a comma or an exclamation mark). They are used to improve the interaction between the speaker and the listener. In some systems the generation of conversational signals has been done by analyzing the text [18] or speech [1].
- Channel 4 contains **emotion displays**, which are **emotional expressions** or **emotion emblems**. **Emotional expressions** are movements to express felt emotions of the speaker. On the other hands, **emotion emblems** express emotions that are being mentioned, or instance, a disgust expression when talking about something disgusting.
- Channel 5 contains **gaze movements** and Channel 6 contains **head movements**. Gaze and head movements are generated to support eye contact or to point to something during conversation. Head movements are also used to replace verbal content (e.g., nodding the head for saying yes). As the eyes and the head do not stay in the same place all the time, we use a noise generating function to create random subtle movements to make the talking head livelier.

We follow Pelechoud et al. [18] to synthesize facial movements in three phases: onset, apex, and offset. We used Essa's work [13] on analysis, identification and synthesis of facial expressions to design the temporal pattern of facial movements. Essa used exponential curves to fit the onset and offset portions of each parameter. Based on the suggested functions by Essa, we derived two functions for the onset and offset portion of a parameter activity. An example of the temporal pattern of a facial movement is shown in Figure 1.

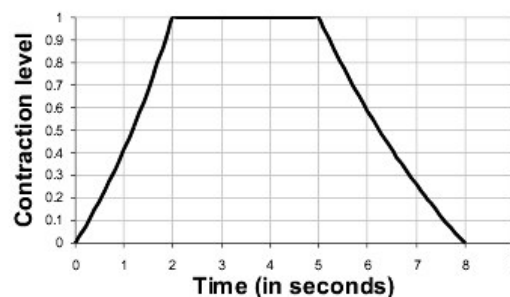


Figure 1. The activity function of a facial movement.

In order to combine facial movements, first, we concatenate the movements in the same channel. This is done by modulating the activity of each muscle involved in the movements in that channel, to create transition effects between movements. This combination only applies to individual muscles.

We use the dominance model [7] to create the co-articulation effect of lip movements when talking. Co-

articulation is the blending effect that surrounding phonemes have on the current phonemes. For the combination of movements from other channels, we have proposed an algorithm to produce smooth transitions between movements. An example of combining the Jaw Rotation of two movements in the same channel is shown in Figure 2.

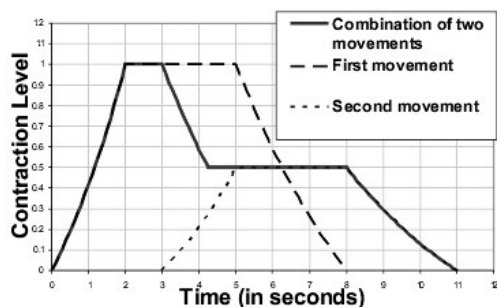


Figure 2. Combination of the Jaw Rotation of two movements in the same channel.

The movements from all channels are then combined taking into account the resolution of possible conflicting muscles. At a certain moment in time, when there is a conflict between parameters in different animation channels, the parameters involved in the movement with higher priority will dominate the ones with lower priority. The activity of those parameters around that time is also affected so that they cannot activate or release too fast. An example of combining the activity of Zygomatic major and Orbicularis Oris is shown in Figure 3.

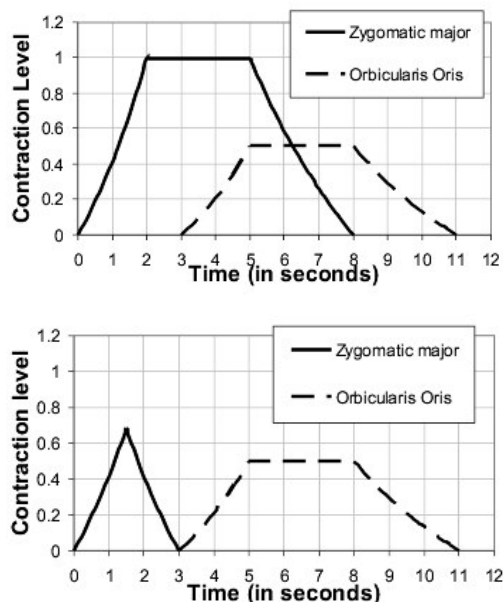


Figure 3. The activity of Zygomatic major and Orbicularis Oris before (top) and after (bottom) applying the combination.

## 4 Illustration

Figure 4 shows (frame by frame) our face model uttering the sentence “Oh, really?”. An emotion display happens during the utterance, which is a full surprise display,

starting at time 0 and lasting 2 seconds. The surprise display is mapped to the contraction of the parameter Jaw Rotation with the fuzzy rule based system. The lip movements during speech are calculated based on empirical data from Rutgers university [8]. The figure shows the smooth and natural combination of visual speech and the emotion display.

## 5 Conclusion

In this paper, we have discussed a framework for an embodied conversational agent. The framework is represented as a talking head. We focus on the emotion facial expressions of an agent as well as the dynamic in the combination of facial movements. Our framework has been built based on psychological literature as well as work on facial behavior measurement. With this framework, we can create smooth and natural animation for our embodied conversational agent.

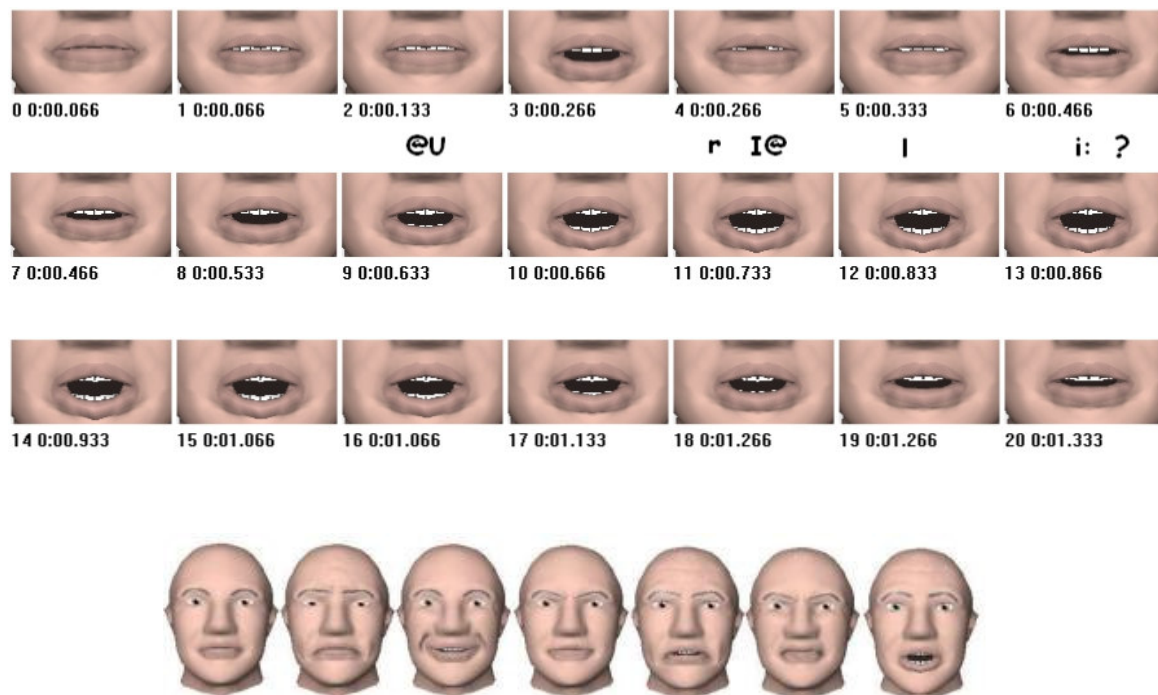
In the future, we want to verify our framework with more empirical data from research on facial behavior measurement. We also want to do human judgment studies to evaluate our framework.

Dirk Heylen, Anton Nijholt, and Mannes Poel continue this research within the Humaine Network of Excellence.

## REFERENCE

1. Albrecht, J. Haber, M. Schröder, and H. P. Seidel (2002). Automatic generation of non-verbal facial expressions from speech. In *Proceedings Computer Graphics International (CGI) 2002*, pp. 283–293.
2. E. André, G. Herzog, and T. Rist (1997). Generating multimedia presentations for robocup soccer games. In H. Kitano, ed., *RoboCup '97: Robot Soccer World Cup I*, pp. 200–215. Springer-Verlag, New York.
3. B. Blumberg, M. Downie, Y. A. Ivanov, M. Berlin, M. P. Johnson, and B. Tomlinson (2002). Integrated learning for interactive synthetic characters. In *SIGGRAPH 2002*, pp. 417–426.
4. T. D. Bui, D. Heylen, M. Poel, and A. Nijholt (2001). Generation of facial expressions from emotion using a fuzzy rule based system. In *Proc. AI 2001*, pp. 83–95.
5. T. D. Bui, D. Heylen, M. Poel, and A. Nijholt (2002). Parlee: An adaptive plan-based event appraisal model of emotions. In *Proc. KI 2002*, pp. 129–143.
6. T. D. Bui, D. Heylen, and A. Nijholt (2003). Improvements on a simple muscle-based 3d face for realistic facial expressions. In *Proc. CASA-2003*, pp. 33–40.
7. Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In Magnenat Thalmann, N. and Thalmann, D., ed., *Models and Techniques in Computer Animation*, pp. 139-156. Springer, Tokyo.
8. D. C. DeCarlo, M. S. Revilla, and J. Venditti (2002). Making discourse visible: Coding and animating conversational facial displays. In *Computer Animation 2002*.
9. P. Ekman (1989). The argument and evidence about universals in facial expressions of emotion. In H. Wagner and A. Manstead, ed., *Handbook of Social Psychophysiology*. Wiley, Chichester, New York.

10. P. Ekman and Friesen, W. V. (1975). *Unmasking the Face: A Guide To Recognizing Emotions From Facial Clues*. Prentice-Hall, Englewood Cliffs, New Jersey.
11. P. Ekman and Friesen, W. V. (1978). *Facial Action Coding System*. Consulting Psychologists Press, Palo Alto, CA.
12. P. Ekman, T.S. Huang, T.J. Sejnowski, and J.C. Hager, ed. (1993). *Final Report to NSF of the Planning Workshop on Facial Expression Understanding*. Human Interaction Laboratory, University of California, San Francisco, 1993.
13. Essa, I. A. (1994). *Analysis, Interpretation, and Synthesis of Facial Expressions*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.
14. L. Itti, N. Dhavale, and F. Pighin (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*.
15. S. A. King, R. E. Parent, and B. Olsafsky (2000). An anatomically-based 3d parametric lip model to support facial animation and synchronized speech. In *Proceedings of Deform 2000*, pp. 7–19.
16. C. Latta, N. Alvarado, S. S. Adams, and S. Burbeck (2002). An expressive system for animating characters or endowing robots with affective displays. In *AISB 2002 Annual Conference, Symposium on Animating Expressive Characters for Social Interactions*.
17. C. Pelachaud, M. L. Viaud, and H. Yahia (1993). Rule structured facial animation system. *IJCAI*, 2:1610–1615.
18. C. Pelechaud, N. I. Badler, and M. Steedman (1996). Generating facial expressions for speech. *Cognitive Science*, 20:1–46.
19. Picard, R. (1997). *Affective Computing*. MIT Press, Cambridge, MA.
20. J. Rickel and W. L. Johnson (1998). Steve: A pedagogical agent for virtual reality. In *Proceedings of the Second International Conference on Autonomous Agents*.
21. M. Theune, S. Faas, D. Heylen, and A. Nijholt (2003). The virtual storyteller: Story creation by intelligent agents. In *Proceedings TIDSE 03*, pp. 204–215. Fraunhofer IRB Verlag.



**Figure 4.** Our talking head utters the sentence “Oh! Really?” while displaying surprise and a sample of expressions.