Movement Realism in Computer Facial Animation

Steve Maddock, James Edge and Manuel Sanchez
Dept. Computer Science, University of Sheffield, UK
{s.maddock,j.edge,m.sanchez}@dcs.shef.ac.uk

Abstract

Static and dynamic realism are both important in computer facial animation. However, for embodied conversational agents in interfaces, and for agents or avatars in virtual environments, movement and behaviour are more important than photorealism. In this paper we summarise our work on two systems that feature aspects of this dynamic realism: coarticulation for visual speech and facial tissue deformation producing expressions and wrinkles.

1 Introduction

Cartoonists have always recognised that the behaviour of a character is more important than a photorealistic look. Roboticists also attend to this, and have even coined the term 'uncanny valley' [Mori70], as illustrated in Figure 1. Assuming we can extend this to synthetic characters, we seem to be able to increase their human-like appearance up to a threshold beyond which the empathy decreases sharply and the character becomes 'uncanny'. As the character becomes more human-like, the empathy returns.

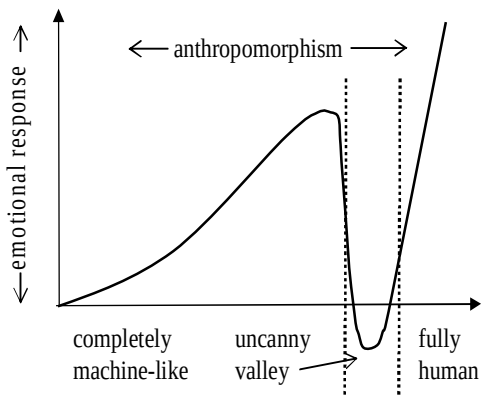

Figure 1: Emotional response against robot realism.

Research in virtual environments also supports this differentiation between behaviour and photorealism. Correct avatar gaze, over photorealism, improves immersive experience [Bail01, Bail04], as long as the character has a sufficient level of realism [Gara03]. In addition, Blascovitch *et al* [Blas02] argue that photographic realism is only important if it relates specifically to behavioural realism, e.g. recognisable eyebrows are needed in order to convey a frown. Also, Casaneuva and Blake [Casa01] found that co-presence was improved by using avatars with gestures and facial expressions. In contrast to others though, they also found that "more realistic avatars generated higher levels of co-presence".

Considering these ideas in relation to facial animation, we suggest that the word 'behaviour' needs to be split into two parts: (i) the physical movement of the face; (ii) the character which the face is a part of. Thus lip shape, wrinkles and 'wobbling fat' are considered rather than whether or not the character behaves in a human-like way, showing some form of intelligence. Thus we are not focussing on such things as conversational mechanics (e.g. see [Sand00] or [Cass01]) or the correlation between the raising of the eyebrows and voice pitch [Cave96]. We know that speech perception is improved if natural visual movement accompanies it [Mass98, Mass00], that "static expressive faces are rated less attractive than moving expressive faces" [Knap02] and "during the computation of identity the human face recognition system integrates both types of information: individual non-rigid facial motion and individual facial form" [Knap03]. However, it could be argued that these three examples encapsulate aspects of AI in that natural facial gestures such as gaze and eyebrow movement are incorporated.

We will describe two systems that we have developed that attend to aspects of movement realism. The first system is a pose-target approach to facial animation (e.g. see [Park72]) and the second system uses motion capture data. For the first system, described in section 2, we have developed a constrained optimisation approach to facial animation [Edge04a; Edge04c] (based on Witkin and Kass's work on articulated body animation [Witk88]) that attends to the particular idiosyncracies of speech movement. Thus, the mouth moves in a realistic way.

For the second system, we consider the way that the skin moves when making an expression. We use motion capture data to give realistic skin movement [Sanc03, Edge04b, Sanc04]. Since we use motion capture data, we get realistic pace of movement and correlation of movement in different areas of the face. On top of the motion capture data, we layer wrinkles in real-time [Sanc04] to enhance the observed effect of such movement. Section 3 will describe this. Finally, Section 4 presents some conclusions.

2 Visual Speech

Perhaps the most important aspect of facial animation is visual speech, which has lagged behind progress in audio speech synthesis. Speech audio is often considered as a sequence of atomic units called phonemes. For each audio phoneme, a corresponding visual equivalent called a viseme can be used that capture an extreme variation in speech articulation – we only observe some of this articulatory system, e.g. lips. A typical approach to speech animation is then to interpolate between a sequence of visemes. However, a simple interpolation-based approach is inadequate as it will not capture the natural variation in speech movements. The disparity between the atomic viseme-based representation of speech movements and natural articulation is referred to as coarticulation.

In natural speech, the extent to which viseme 'targets' are met is directly affected by context. Thus the shape of the

mouth for a particular phoneme is affected by preceding (e.g. boot vs. beet) and following (e.g. stew vs. stick – the lips are rounded when saying the /t/ of stew) phonemes. Essentially a viseme has an effect over a duration of the animation – its static appearance belies its dynamic influence. The standard approach to coarticulation for the pose-target approach to facial animation is to use dominance functions [Cohe93]. We have developed a new approach that uses constraint-based animation [Edge04a; Edge04c].

In our work a viseme (or target or collection of parameters), is regarded as a distribution of vocal tract shapes. The 'ideal' viseme acts as a centre-of-mass within the distribution pulling the speech trajectory towards it – as we have already mentioned, coarticulation effects and speed of talking will usually mean that we do not make the ideal viseme shape with our mouth. Each target parameter is weighted in accordance with its relative dominance for a particular viseme. An ideal closed mouth shape is more likely to be met than an open mouth shape, and is something that we would notice visually.

In order to generate viseme transitions for a given speech utterance we apply a technique similar to the spacetime constraints method used for articulated body animation [Witk88]. For a sequence of visemes, we optimise the distance between the speech trajectory and the ideal viseme centres subject to a set of constraints. The

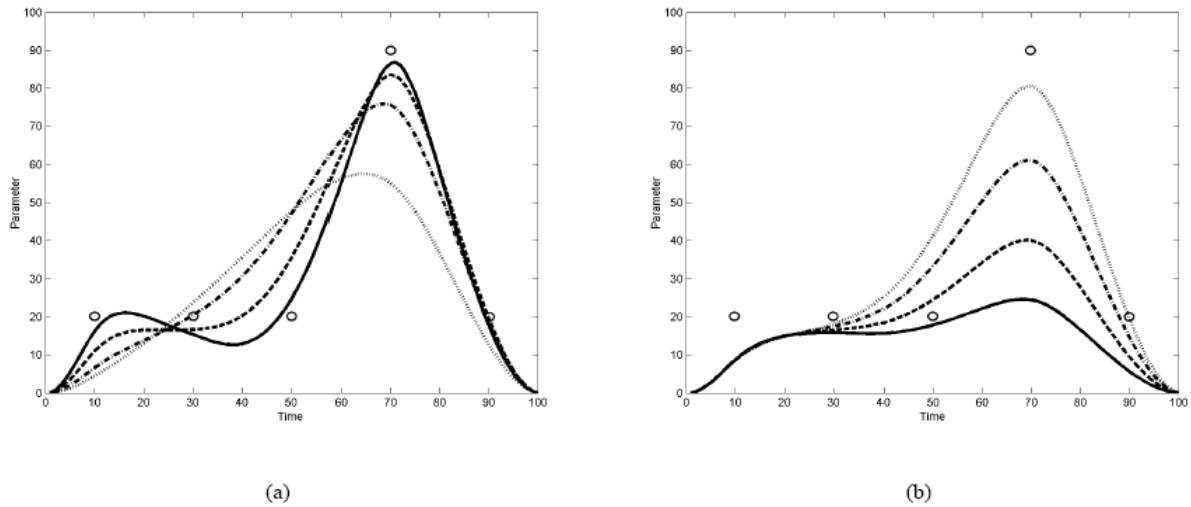

Figure 2: Varying parameters for an example trajectory through five visemes, represented as small open circles. (a) This demonstrates the effect of increasingly constraining parametric acceleration on the final speech trajectory (the dotted trajectory is most constrained, solid is least); this effectively dampens the motion preventing all targets (small open circles) from being met. (b) Decreasing the dominance of the fourth target reduces its effect over the trajectory (notice that this effect is exerted over several surrounding visemes). A combination of these two types of variable can be used to model basic speech coarticulation.

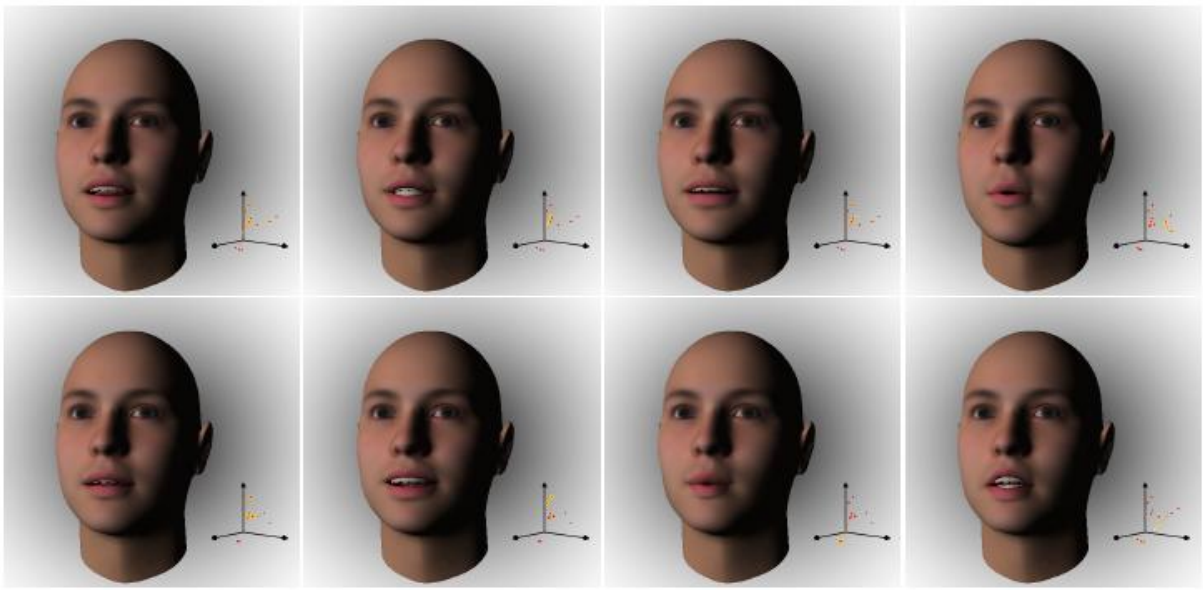

Figure 3: Frames from an animation incorporating coarticulation.

parametric acceleration of the trajectory is limited to prevent visemes from being interpolated. As the trajectory is increasingly constrained visemes will be less well met, with highly dominant visemes exerting greatest influence over the final movement. Figure 2 illustrates this process, and Figure 3 shows some frames from an example animation[1].

3 Wrinkles

For static synthetic faces, realistic skin rendering is usually achieved by using carefully created texture maps or photo textures. This works well, although significant improvements can be gained by using more realistic reflection models that attend to sub-surface scattering (see, for example, [Hanr93] or the imitative real-time approach of [Gree04]). Such texture maps can include wrinkles, giving a face a more weathered or experienced look. However, these do not capture the formation of wrinkles due to skin movement.

We have developed a complete facial animation system that uses motion capture data (MoCap) to animate any synthesized face. The first stage is to capture the movement of a discrete set of markers on a real face. The second stage is to retarget this motion to deform a detailed target mesh representing a face This is achieved using radial basis functions and a surface-to-surface control technique called BIDS [Sanc04], which uses a Bezier-triangle surface spanning the set of motion capture points (as illustrated in Figure 4).

Using MoCap to drive facial animation suffers from the sparseness of data captured about the face. Since only the positions of a discrete set of markers are captured, the fine tissue detail, such as wrinkling, is not included. A layering approach can be used to include such detail. In Sanchez *et al* [Sanc04], we construct a model that relates the wrinkling effects observed on a specific performer to the differential strain sustained by the facial tissue. Then, in a real-time process, as BIDS is used to deform a face mesh, we can evaluate the strain of the specific movement, e.g. an expression, and use the model to produce a normal map representing the wrinkles to layer on. Figure 5 illustrates deformation using BIDS without the wrinkle model. Figure 6 demonstrates how BIDS plus the wrinkle model improves the realism of the results[2].

4 Conclusions

We have described two separate systems that attend to two different aspects of movement realism in computer facial animation. The first system models the realistic movement of the mouth in visual speech, capturing the coarticulation effects that lead to more natural-looking visual speech. The second system uses motion capture data and layers on wrinkling effects to increase the realism of observed facial expression movement. We have also used the motion capture system for visual speech [Sanc03, Edge04b], capturing phrases and words,

Figure 4: The BIDs control surface formed by triangulating the motion capture markers.

Figure 5:The motion on the face on the left is transferred to the synthetic face on the right using BIDS. The skin on the synthetic face does not wrinkle.

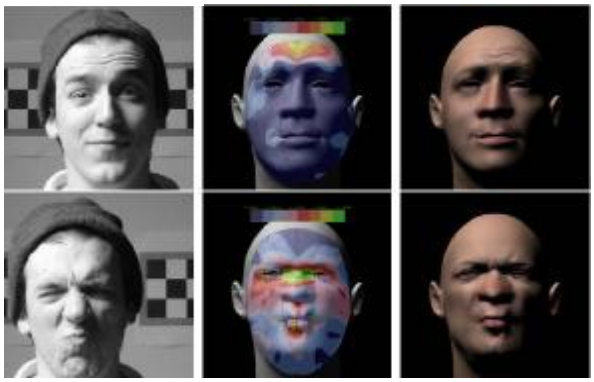

Figure 6: The motion on the faces on the left is transferred to the synthetic faces on the right using BIDS. A real-time strain analysis of the movement of the BIDS control surface is shown in the middle. This is used to add wrinkles to the synthetic face.

where we could argue that coarticulation is handled within individual dynamic units, although not between units.

We have also commented on the need to attend to realism of facial movement. However, there seems to be no evidence as to how realistic facial movement such as correct mouth shape and wrinkles has to be. It may be that it is the overall character 'behaviour' that is more important than such facial movement. The character must be seen to react in an intelligent and human-like manner, e.g. gaze, eye blinks, lips 'synchronised-enough' with speech. Perhaps even a more-than-real conversational agent may be worth exploring, with exaggerated movements (as in cartoons), with overlaid graphics, e.g. speed lines, or some form of psychorealism as in Landreth's Ryan (see [Robe04]) where the face is "augmented with growths that amplify his emotions" [Robe04].

[1] http://www.dcs.shef.ac.uk/~steve/HCI05/quotes.mpg

[2] http://www.dcs.shef.ac.uk/~steve/HCI05/v2mocap.mpg

References

Bail01 Bailenson, J.N., J. Blascovich, A.C. Beall, J.M. Loomis (2001). Equilibrium revisited: Mutual gaze and personal space in virtual environments. PRESENCE: Teleoperators and Virtual Environments, 10. 583-598.

Bail04 Bailenson, J.N., & J. Blascovich (2004). Avatars. Encyclopedia of Human-Computer Interaction, Berkshire Publishing Group, 64-68.

Blas02 Blascovich, J., J. Loomis, A. Beall, K. Swinth, C. Hoyt, J.N. Bailenson (2002). Immersive virtual environment technology as a methodological tool for social psychology. Psychological Inquiry, 13, 103-124.

Casa01 Casanueva, J. S. and E.H. Blake (2001). The Effects of Avatars on Co -presence in a Collaborative Virtual Environment. Technical Report CS01-02-00, Department of Computer Science, University of Cape Town, South Africa.

Cass01 Cassell, J. (2001). Representation and Intelligence in Embodied Conversational Agents. AI Magazine 22 (3), 67-83.

Cave96 Cave, C., I. Guaitella, R. Bertrand, S. Santi, F. Harlay, R. Espesser (1996). About the relationship between eyebrow movements and F0 variations. Proceedings of ICSLP 96. Wilmington, DE: Univ. of Delaware. 2175-2178.

Cohe93 Cohen, M. and D. Massaro (1993). Modeling coarticulation in synthetic visual speech. In Proceedings Computer Animation '93, 139–156.

Edge04a Edge, J. and S. Maddock (2004). Constraint-based Synthesis of Visual Speech, ACM SIGGRAPH'04 Technical Sketch.

Edge04b Edge, J., M. Sanchez, S. Maddock (2004). Using motion capture data to animate visual speech, Symposium on Language, Speech and Gesture for Expressive Characters, March 30-31, 2004, part of the AISB 2004 Convention: Motion, Emotion and Cognition, University of Leeds, UK, March 29 - April 1, 2004, pp.66-74.

Edge04c Edge, J. (2004). Techniques for the Synthesis of Visual Speech, PhD thesis, Department of Computer Science, University of Sheffield, UK, September 2004.

Gara03 Garau, M., M. Slater, V. Vinayagamoorhty, A. Brogni, A. Steed, M.A. Sasse (2003). The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. Proceedings of the SIG-CHI conference on Human factors in computing systems, April 5-10, 2003, Fort Lauderdale, FL, USA, 529 - 536.

Gree04 Green, S. (2004). Real-Time Approximations to Subsurface Scattering, GPU Gems, Addison-Wesley 2004.

Hanr93 Hanrahan, P. and W. Kreuger (1993). Reflections from Layered Surfaces due to Subsurface Scattering, Proc. SIGGRAPH 1993, 165-174.

Knap02 Knappmeyer, B., I.M. Thornton, N. Etcoff, H.H. Bülthoff (2002). The influence of facial motion on the perception of facial attractiveness. Second Annual Meeting of the Vision Science SocietySarasota, Florida (May 2002).

Knap03 Knappmeyer, B., I.M. Thornton and H.H. Bülthoff. (2003). The use of facial motion and facial form during the processing of identity. Vision Research 43(18), 1921-1936.

MacD05 MacDorman, K.F. (2005). Androids as an Experimental Apparatus: Why is there an uncanny valley and can we exploit it? Proc CogSci-2005 Workshop: Toward Social Mechanisms of Android Science, 25-26 July 2005 in Stresa, Italy, pp. 106-118

Mass98 Massaro, D.W., (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, Massachusetts: MIT Press

Mass00 Massaro, D.W., M.M. Cohen, J. Beskow, R.A. Cole (2000). Developing and evaluating conversational agents. Chapter 10 of Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., eds., 2000. Embodied Conversational Agents. MIT Press, 2000, ISBN 0-262-03278-3.

Mori70 Mori M. (1970). Bukimi no tani [The uncanny valley]. Energy, 7(4), pp 33-35. (see [MacD05] for a translation by Karl F MacDorman and Takashi Minato)

Park72 Parke. F. (1972). Computer generated animation of faces. Proc. ACM National Conference, No. 1, pp.451–457. ACM, 1972.

Robe04 Robertson, B. (2004). Psycholrealism: Animator Chris Landreth creates a new form of documentary filmmaking. Computer Graphics World July, 2004.

Sanc03 Sanchez Lorenzo, M., J.D. Edge, S. King, and S. Maddock (2003). Use and Re-use of Facial Motion Capture Data, Proc. Vision, Video and Graphics 2003, University of Bath, July 10-11, 2003, pp. 135-142.

Sanc04 Sanchez, M., J.D.Edge, S.C.Maddock (2004). Realistic Performance-driven Facial Animation using Hardware Acceleration, Department of Computer Science Research Memorandum CS-04-10, University of Sheffield.

Sand00 Sanders, G. A., and J. Scholtz (2000). Measurement and Evaluation of Embodied Conversational Agents. Chapter 12 of Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., eds., 2000. Embodied Conversational Agents. MIT Press, 2000, ISBN 0-262-03278-3.

Witk88 Witkin, A., and M. Kass (1988). Spacetime constraints. In Proceedings of the 15th annual conference on Computer graphics and interactive techniques (Siggraph'88), 159–168.