Dynamic Realistic Lip Animation using a Limited number of Control Points

Slim Ouni[1]*, Guillaume Gris[2]
[1] Université de Lorraine / LORIA, [2] Ecole Polytechnique

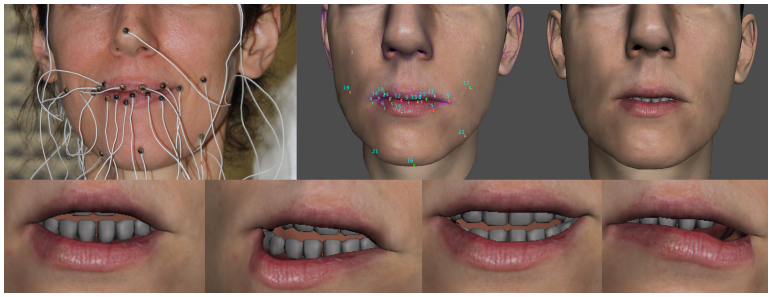

Figure 1: *From Electromagnetic Articulography data acquisition to lip animation*

1 Introduction

One main concern of audiovisual speech research is the intelligibility of audiovisual speech (i.e., talking head). In fact, lip reading is crucial for challenged population as hard of hearing people. For audiovisual synthesis and animation, this suggests that one should pay careful attention to modeling the region of the face that participates actively during speech. Above all, a facial animation system needs extremely good representations of lip motion and deformation in order to achieve realism and effective communication.

The existing methods that model accurately the face with the goal of creating a convincing talking head can produce globally satisfactory results in terms of the static realism of the face rendering, however, some of them fail when dealing with speech-related animation. Very often, the quality of the lip animation during speech is not realistic and, more importantly, not intelligible. Currently, a limiting factor is the complexity of facial movements, where the fine deformation of the lips can be difficult to capture. For instance, only the outer lip contour can be well-captured, however, it is very difficult to capture the inner lips . In this case, markers can often be occluded during protrusion or complete mouth closure. Thus, it is difficult to check whether the lips are completely closed or not, which is crucial for the realization of bilabial sounds, for instance.

For this reason, an alternative approach to improving the accuracy of dynamic tracking of the shape of the lips, would use a tracking system that can provide the appropriate information even when the markers are hidden. Electromagnetic Articulography (EMA) can be a robust technique for providing such information. EMA captures articulatory movements in 3D with a high temporal resolution ($250Hz$) and high spatial resolution ($0.3mm$ RMSE), by tracking tiny sensors attached to speech articulators such as the tongue, teeth, and lips, or any part of the face. The positions of these sensors are calculated by measuring the electrical currents produced within multiple low-intensity electromagnetic fields. This technique is known to present no risk to the health of the speaker, and has been used in the study of speech production for more a decade.

In our work, we have used an articulograph to acquire a set of markers glued on the face (mainly on the lips) and then fitted to a 3D human face model of a human speaker. Finally, we apply an interpolation scheme of the displacement field between the control points. This displacement field describes the deformation of the face surface.

2 Our Approach

We propose to use the EMA technique to dynamically capture the shape of the lips. The EMA sensors are thus to be glued mainly on the lips. The newest articulograph AG501 can track up to 24 sensors simultaneously. This may be enough to control the lips, but it is not sufficient for accurate animation of the entire face. It is more likely that this technique needs to be combined with other motion capture techniques to be able to animate the whole face (e.g., the EMA sensors for the lips, and other motion capture technique for the rest of the face). In our approach, we propose to glue the EMA sensors on the lips and on a very limited region of the face, and then we propose a method for animating the face, but mainly focusing on the lower part of the face related to speech articulation.

The first steps of our approach are: (1) acquiring a 3D model of the face of the speaker (static), using a low-cost 3D scan technique based on a kinect, and (2) acquiring the motion of the face (dynamic) using EMA data. The tracked sensors on the face are the control points that deform the shape of the face. These control points are first fitted to a 3D human face model of a human speaker by minimizing the distance between the control points and the surface of the face model. Finally, we apply an interpolation scheme of the displacement field between the control points. This displacement field describes the deformation of a surface from a limited number of control points. In the case of the face, this method is well adapted to animating realistically and intelligibly the region of the face that is highly correlated with speech, specifically the lips and the lower part of the face, even with a very limited number of control points. In our future work, this technique will be used mainly to improve audiovisual speech synthesis and lip-syncing.

Acknowledgements

This work was supported (in part) by the EQUIPEX Ortolang.

*e-mail:slim.ouni@loria.fr

SIGGRAPH 2015 Posters, August 09 – 13, 2015, Los Angeles, CA.
ACM 978-1-4503-3632-1/15/08.
http://dx.doi.org/10.1145/2787626.2787628