The UWB 3D Talking Head Text-Driven System Controlled by the SAT Method Used for the LIPS 2009 Challenge

Zdeněk Kroul, Miloš Železný

Department of Cybernetics, Faculty of Applied Sciences,
University of West Bohemia, Plzeň, Czech Republic
{zdkrnoul,zelezny}@kky.zcu.cz

Abstract

This paper describes the 3D talking head text-driven system controlled by the SAT (Selection of Articulatory Targets) method developed at the University of West Bohemia (UWB) that will be used for participation in the LIPS 2009 challenge. It gives an overview of methods used for visual speech animation, parameterization of a human face and a tongue, and a synthesis method. A 3D animation model is used for a pseudo-muscular animation schema to create visual speech animation usable for lipreading.
Index Terms: facial animation, audio-visual speech synthesis, audio-to-visual mapping

1 Introduction

The research on audio-visual speech synthesis has been done at UWB for several last years [1, 2, 3]. At present, our talking head system automatically converts input text into audiovisual speech in the form of an animation of a 3D human head model. An input sequence of phonemes is transformed into a continuous stream of animated lip and tongue movements. We can change an appearance or parameterization of the 3D model as well as the type of the synthesis method. In previous work, we concentrated on a study of labial co-articulation and a development of such speech output that can be used as a lipreading support.

2 Talking head system overview

The presented system performs mapping from text to visual speech. It uses an acoustic speech synthesizer, which has been developed independently. In this paper we concentrate on the visual speech synthesis system. The visual speech synthesizer is based on a parametrically controllable 3D model of a human head. It uses a generic 3D model of a head that is adapted to have a look of a specific person using special 3D scanning method. Movable parts can be animated by a set of control points. Surrounding of these points is interpolated to achieve smooth face movements. Synthesis itself is concatenative using the method of selection of articulatory targets. Co-articulation is modeled by visual unit selection method. The 3D model is rendered using OpenGL.

3 Animation schema

The animation schema is based on geometric representation of a human face and other necessary parts of a head in 3D space by polygonal meshes. We consider computation of deformations of outermost layer of skin only. We model deformations of face skin (epidermis) and a tongue without subsurface layers. The main aim of the animation schema is to provide synthesized visual speech as a lipreading support.

The principle of the animation schema is based on feature points and deformation zones. A deformation zone is considered as an area of the 3D face surface given by a set of vertices and the feature point as one 3D point in this area. A shift of its position causes changes in whole deformation zone. It uses one polygonal mesh only that is often formed in neutral face pose in comparison with approaches using interpolation of multiple face poses. Low demands on the shape of the polygonal mesh make it very flexible as well. However, for better approximation of a face shape around lips, deformations commutated around a curve appear to be more suitable. Therefore, we have designed an animation schema which counts with both mentioned cases. The animation schema is based on 3D cubic spline curves, each defined by several control points. We can find similarity with the control of the tongue model defined in 2D space. On the other hand, the advantage of a 3D cubic spline curve is that it uses feature points directly matched with animation parameters of a 3D head model.

3.1 Parameterization of visual speech

The parameterization of lips or a tongue is given by the layout of the feature points on the polygonal meshes. We use a low level parameterization defined in MPEG4 standard [4]. We describe a relation of the feature points to several spline curves to cover the deformations caused by muscles under the epidermis layer. The lips are parameterized by 16 feature points and two closed curves. One curve connects eight points from the FAP group 8 except the point 8.1 to approximate the outer lip contour. The inner lip contour is created using the FAP group 2. The deformation of dermis caused by jaw rotation is approximated by one open curve constructed from FAPs 2.1, 2.13 and 2.14. The centers of cheeks are controlled by two isolated feature points FAPs 5.1 and 5.2. The tongue model is controlled by two curves defined in *sagital* and *transversal* planes. The first one is defined by five feature points. The tongue tip and tongue dorsum correspond to FAPs 6.1 and 6.2. The second one controls the width of tongue body and is constructed from five feature points. The sides of tongue body are controlled by FAPs 6.3 and 6.4 and the feature point on the tongue tip is shared with the first curve.

Data analysis of measured FAPs can produce high level parameterization of visual speech. We have used the principal component analysis (PCA) to reduce feature point space to four dimensions. Currently we have these parameters: *lip opening*, *lip protrusion*, *upper lip raising* and *jaw rotation*.

3.2 Selection of articulatory targets

The synthesis method is based on the regression tree technique. The method predicts the articulatory targets as one continuous value given for each phoneme and articulatory parameter separately. The articulatory targets are placed at the centers of the speech segments and interpolation is used to create the continuous trajectory in given frame rate (usually 25 frames per seconds).

4 Conclusion

In this paper, we described the 3D talking head text-driven system controlled by the SAT method developed at UWB. The summarization of new features is presented. The presented animation schema allows more precise approximation of the inner and outer lip contours as well as deformations of the tongue mesh. Parameterization of visual speech is based on the MPEG4 standard. The articulatory trajectories are suitable for experiments with data-driven synthesis methods.

5 Acknowledgements

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/P609, by the Ministry of Education of the Czech Republic, project No. ME08106, and by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416.

References

[1] M. Železný, Z. Krňoul, P. Císař, and J. Matoušek, "Design, implementation and evaluation of the czech realistic audio-visual speech synthesis," *Signal Procesing, Special section: Multimodal human-computer interfaces*, vol. 86, pp. 3657–3673, 2006.

[2] Z. Krňoul, M. Železný, L. Müller, and J. Kanis, "Training of coarticulation models using dominance functions and visual unit selection methods for audio-visual speech synthesis," in *Proceedings of INTERSPEECH 2006 - ICSLP*, Bonn, 2006.

[3] Z. Krňoul and M. Železný, "Innovations in czech audio-visual speech synthesis for precise articulation," in *Proceedings of AVSP 2007*, Hilvarenbeek, Netherlands, 2007.

[4] M. Escher, I. Pandzic, and N. M. Thalmann, "Facial deformations for MPEG-4," in *Proceedings of the Computer Animation*. IEEE Computer Society, 1998, p. 56.