

Cloning speakers' articulation, shape and appearance

Gérard Bailly, Pierre Badin, Frédéric Elisei, Oxana Govokhina,

Christophe Savariaux & Yuliya Tarabalka

Department of Speech and Cognition, GIPSA-Lab, CNRS & Universities of Grenoble, France

Contact author : gerard.bailly@gipsa-lab.grenoble-inp.fr

Abstract

We briefly describe here the methodology we have developed at GIPSA for cloning speakers, i.e. capturing and modeling speaker-specific control, shape and appearance models. We will mention here evaluation experiments that have been conducted to assess the perceptual benefit of visual synthesis used in three different applications: lipsync/text-to-audiovisual synthesis, audiovisual conversion and tongue reading.

Index Terms: facial animation, audiovisual speech synthesis, HMM

1. Introduction

We all share a common anatomy that enables us to produce sounds by shaping sound sources and the vocal tract shape. We all have a jaw, a tongue, lips: this partially explain regularities of languages. We almost develop speech skills the same way: starting vocalizing, babbling, pointing... ending up mastering the subtle placing and shaping of our articulators of our ambient language.

We are also all different: no other person as you articulate [u] with lips in a round or a tucked [f] like you. The challenge of facial animation is to cope with invariance of function and variability of realizations keeping coherence of behavior and respecting personality.

Since the pioneer work by Benoit et al [8], we have maintained a strong activity on articulatory modeling and control. Shape models for the lips, the face [15] as well as for the tongue [1, 2], the velum [16] and the eyelids [11] have been developed using always basic principles: massive speaker-specific data collected on a few subjects thanks various experimental settings and recording devices (photogrammetry, EMA, MRI, etc.), knowledge driven analysis and modeling and evaluation. Note also our work on cued speech [12, 13] where gestures of the head and facial articulation should be combined with hand movements.

2. Data and articulatory models

The shape model that produces the facial geometry from the articulatory score is an crucial component. We essentially combine three techniques for building shape models of speaker-specific organs:

- Fleshpoints marking and tracking: we basically glue colored beads on the face (see *Figure 1*) or ElectroMagnetic Articulograph (EMA) coils on the tongue and velum
- Semi-automatic positioning/parameterization of generic models of organs: generic lips, eyes, teeth, back side of the head are fitted to silhouettes or surfaces. 3D-to-3D matching procedures are also used [9]
- Linear/nonlinear models linking position of fleshpoints and surface geometry are then built (see *Figure 3*)

Our shape models are then built from the collection of positions of fleshpoints using a so-called guided Principal Component Analysis (PCA) where a priori knowledge is introduced during the linear decomposition. We actually compute and iteratively subtract predictors using carefully chosen data subsets [1]. For facial movements, this methodology enables us to extract six components directly related to jaw, proper lip movements and clear movements of the throat linked with underlying movements of the larynx and hyoid bone. The resulting articulatory model also includes components for head movements (see *Figure 2*) and basic facial expressions [4, 5]. The average RMS modeling error is less than 0.5mm for all speakers cloned so far.



Figure 1. Colored beads have been glued on the subject's face along Langer's lines so as to cue geometric deformations caused by main articulatory movements when speaking.

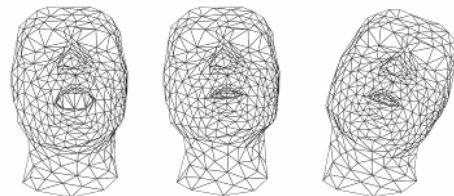


Figure 2: Some elementary articulations for the face and the head that statistically emerge from the motion capture data of speaker CD using guided PCA.

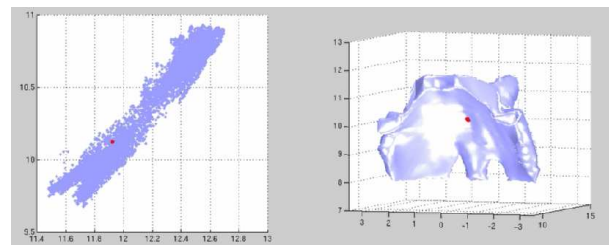


Figure 3: linking movement of a single EMA coil to 3D deformation of the velar region

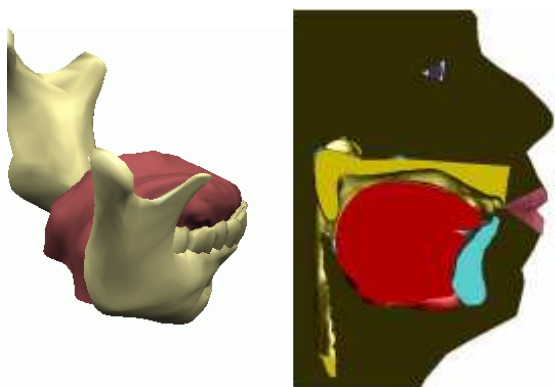


Figure 4: Left: the 3D tongue model developed from MRI data. Right: controlling the tongue movements from EMA kinematic data.

3. Controlling articulation

We have developed several control models able to compute articulatory scores either from a phonetic input or from a speech signal. One of the major challenges of multimodal synthesis is the intra-coordination between constituents of each modality (formants, excitations for acoustics, articulators for facial animation) and the inter-coordination between modalities.

From phonetic input. The trainable PHMM trajectory formation system [6] complements a standard trajectory HMM with a delay model. Both are trained with an iterative analysis-synthesis loop combining HMM training and forced alignment. PHMM improves both the modeling accuracy and subjective scores [7].

From motion capture. Since our virtual talking heads are driven by parameters directly computed from movements of fleshpoints, such models can be driven from motion capture. We have tested ability of naïve viewers to "read" movements of virtual tongue/face driven by motion capture data (cf. Figure 4). Results show that tongue reading needs training and benefits at very low SNR [3]

From speech. We have also developed and tested GMM-based techniques for audiovisual speech conversion [14, 17]. Performances show that our modeling approach captures essential visual cues that can both be used for audiovisual speech synthesis and recognition.

4. Models of appearance

Active appearance models [10] have popularized data-driven approaches for texturing shape models. Thanks to sub-millimeter modeling of the facial shape, numerous samples of shape-free images can be gathered that capture texture variation of fleshpoints: the typical blurring of AAM due to semi-automatic positioning of feature points on a few dozen visemes does not occur. We have demonstrated [4] that audiovisual recognition of reconstructed videos does not statistically differ from the original.

5. References

- [1] Badin, P., G. Bailly, L. Revéret, M. Baciú, C. Segebarth, and C. Savariaux, *Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images*. Journal of Phonetics, 2002. **30**(3): p. 533-553.
- [2] Badin, P. and A. Serrurier. *Three-dimensional linear modeling of tongue: Articulatory data and models*. in *International Seminar on Speech Production (ISSP)*. 2006. Ubatuba, SP, Brazil. p. 395-402.
- [3] Badin, P., Y. Tarabalka, F. Elisei, and G. Bailly. *Can you "read tongue movements"?* in *Interspeech*. 2008. Brisbane, Australia. p. 2635-2637.
- [4] Bailly, G., A. Bégault, F. Elisei, and P. Badin. *Speaking with smile or disgust: data and models*. in *Auditory-Visual Speech Processing Workshop (AVSP)*. 2008. Tangalooma - Australia. p. 111-116.
- [5] Bailly, G., F. Elisei, P. Badin, and C. Savariaux. *Degrees of freedom of facial movements in face-to-face conversational speech*. in *International Workshop on Multimodal Corpora*. 2006. Genoa - Italy. p. 33-36.
- [6] Bailly, G., O. Govokhina, G. Breton, F. Elisei, and C. Savariaux. *The trainable trajectory formation model TD-HMM parameterized for the LIPS 2008 challenge*. in *Interspeech*. 2008. Brisbane, Australia. p. 2318-2321.
- [7] Bailly, G., O. Govokhina, F. Elisei, and G. Breton, *Lip-synching using speaker-specific articulation, shape and appearance models*. Journal of Acoustics, Speech and Music Processing, submitted.
- [8] Benoît, C. and B. Le Goff, *Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP*. Speech Communication, 1998. **26**: p. 117-129.
- [9] Béram, M., G. Bailly, M. Chabanas, M. Desvignes, F. Elisei, M. Odisio, and Y. Pahan, *Towards a generic talking head*, in *Towards a better understanding of speech production processes*, J. Harrington and M. Tabain, Editors. 2006, Psychology Press: New York. p. 341-362.
- [10] Cootes, T.F., G.J. Edwards, and C.J. Taylor, *Active Appearance Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. **23**(6): p. 681-685.
- [11] Elisei, F., G. Bailly, and A. Casari. *Towards eyegaze-aware analysis and synthesis of audiovisual speech*. in *Auditory-visual Speech Processing*. 2007. Hilvarenbeek, The Netherlands. p. 120-125.
- [12] Gibert, G., G. Bailly, D. Beautemps, F. Elisei, and R. Brun, *Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech*. Journal of Acoustical Society of America, 2005. **118**(2): p. 1144-1153.
- [13] Gibert, G., G. Bailly, and F. Elisei. *Evaluating a virtual speech cuer*. in *InterSpeech*. 2006. Pittsburgh, PE. p. 2430-2433.
- [14] Heracleous, P., D. Beautemps, V.-A. Tran, H. Loevenbruck, and G. Bailly, *Exploiting visual information for NAM recognition*. IEICE Electronics Express, 2009. **6**(2): p. 77-82.
- [15] Revéret, L., G. Bailly, and P. Badin. *MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation*. in *International Conference on Speech and Language Processing*. 2000. Beijing, China. p. 755-758.
- [16] Serrurier, A. and P. Badin, *A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data*. Journal of the Acoustical Society of America, 2008. **123**(4): p. 2335-2355.
- [17] Tran, V.-A., G. Bailly, H. Loevenbruck, and C. Jutten. *Improvement to a NAM captured whisper-to-speech system*. in *Interspeech*. 2008. Brisbane, Australia. p. 1465-1468.