Learning optimal audiovisual phasing for a HMM-based control model for facial animation

Oxana Govokhina, Gérard Bailly, Gaspard Breton

Learning Optimal Audiovisual Phasing for an HMM-based Control Model for Facial Animation

Oxana Govokhina[1,2], Gérard Bailly[1] and Gaspard Breton[2]

[1] GIPSA-Lab Dpt. Speech & Cognition CNRS/INPG/UJF/Stendhal 38041 Grenoble - France
{Oxana.Govokhina,Gerard.Bailly}@gipsa-lab.inpg.fr

[2] France Telecom R&D, 4 rue du Clos Courtel, BP 59 35512 Cesson-Sévigné - France
Gaspard.Breton@orange-ftgroup.com

Abstract

We propose here an HMM-based trajectory formation system that predicts articulatory trajectories of a talking face from phonetic input. In order to add flexibility to the acoustic/gestural alignment and take into account anticipatory gestures, a phasing model has been developed that predicts the delays between the acoustic boundaries of allophones to be synthesized and the gestural boundaries of HMM triphones. The HMM triphones and the phasing model are trained simultaneously using an iterative analysis-synthesis loop. Convergence is obtained within a few iterations. We demonstrate here that the phasing model improves significantly the prediction error and captures subtle context-dependent anticipatory phenomena.

1. Introduction

Embodied conversational agents – virtual characters as well as anthropoid robots – should be able to compute facial movements from symbolic input in order to engage in conversation with human partners. This symbolic input minimally consists in the phonetic string with phoneme durations. It can be enriched with more phonological information, facial expressions, or paralinguistic information that has an impact on speech articulation (mental or emotional state). A trajectory formation model has thus to be built that computes articulatory parameters from such a symbolic specification of the speech task. These articulatory parameters will then drive the plant (the shape and appearance models of a talking face or the control model of the robot).

Human interlocutors are sensitive to discrepancies between the visible and audible consequences of articulation [1, 2] and have strong expectations on articulatory variability [3] resulting from the under-specification of articulatory targets and planning. The effective modeling of coarticulation in speech is therefore a challenging issue for trajectory formation systems.

Audiovisual speech synthesizers should therefore cope not only with the modeling of adequate inter-articulatory coordination but also with the correct synchronization of audible and visible articulation [4]. Central to all speech synthesizers using rules, stored segments or trajectory formation models to generate speech from phonological input is the choice of speech landmarks. In most systems acoustic boundaries between phones are used as such landmarks for prosody characterization or generation. We question here the relevance of these landmarks for the generation of gestural scores.

2. State-of the art

Several strategies can be proposed to build audiovisual text-to-speech synthesis [5]. The most straightforward solution simply consists in driving a trajectory formation model from the phoneme string and phoneme durations computed by an existing text-to-speech system. The trajectory formation model then uses acoustic phoneme boundaries to anchor the gestural score and the coarticulation model if necessary. Coarticulation is usually predicted using rules [6] or by exploiting an explicit coarticulation model [7, 8] that anchor the positions and spans of the phoneme-specific gestural targets. Interestingly, Kaburagi and Honda [9] have proposed to add dynamic features in the specification of gestural targets in order to cope with inter-gestural phasing relations.

Data-driven trajectory formation systems have also been proposed to automatically capture regularities of the context-dependent gestural realization of phoneme-sized segments [10]. Concatenative audiovisual speech synthesis encapsulates coarticulation effects by storing multimodal segments. The problem of possible asynchronies is thus pushed in the segmentation and smoothing of boundaries and eventually in the compression/expansion of segments if required. Although HMMs are intrinsically generation engines that are tuned to emit a set of training observations, they have been used only recently for speech synthesis and particularly as trajectory formation systems [11, 12]. HMMs can in fact capture inter-gestural phasing relations thanks to the state-dependent static and dynamic probability density functions characterizing the sub-phonemic observations. Although HMM structures have been proposed [13] to take into account larger audiovisual asynchronies, the benefit for audiovisual recognition scores is highly discussed [14]. We should also mention a third possibility that consists in computing articulation directly from speech signals. Proposals range from frame-based linear [15] or nonlinear models to GMM-based or HMM-based mapping models that take as input a large speech window surrounding the current analysis frame [11]. The key problem is here to determine the span of coarticulation and hope that the mapping model will learn context-dependent phasing patterns from training data.

We study here an HMM-based trajectory formation system and claim that audiovisual asynchrony has an impact on its performance. A phasing model has thus been developed that predicts the delays between the acoustic boundaries of allophones to be synthesized and the gestural boundaries of HMM triphones that are proposed by unconstrained HMM alignment.

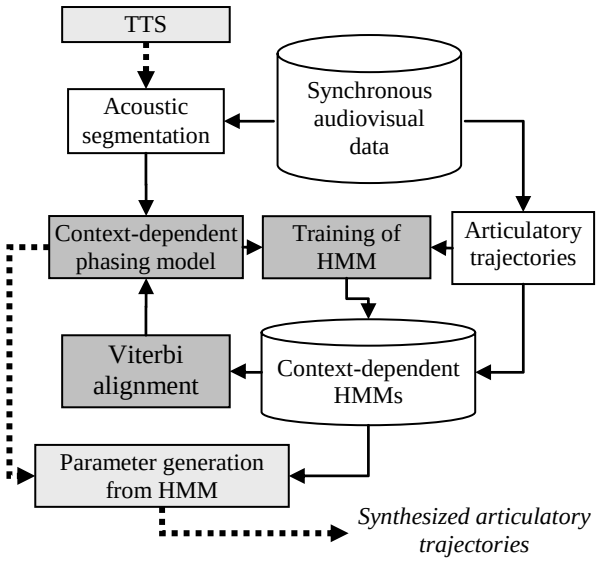

Figure 1. Training consists in iteratively refining the context-dependent phasing model and HMMs (plain lines and dark blocks). The phasing model computes the average delay between acoustic boundaries and HMM boundaries obtained by aligning current context-dependent HMMs with training utterances. Synthesis simply consists in forced alignment of selected HMMs with boundaries predicted by the phasing model (dotted lines and light blocks).

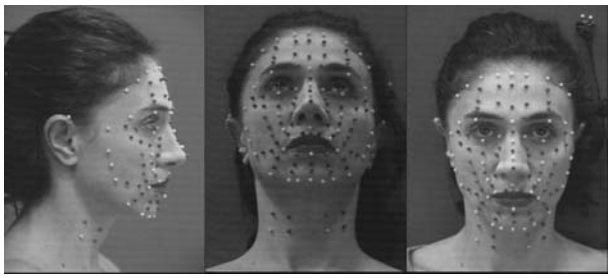

Figure 2. 125 colored beads have been glued on the subject's face along Langer's lines so that to cue geometric deformations caused by main articulatory movements when speaking.

3. Data and articulatory model

In order to be able to compare up-to-date data-driven methods for audiovisual synthesis, a main corpus of 697 sentences pronounced by a female speaker was recorded. Using a greedy algorithm, the phonetic content of these sentences was designed in order to maximize statistical coverage of triphones (differentiated also with respect to syllabic and word boundaries).

We used the motion capture technique developed at ICP [16, 17] that consists in collecting precise 3D data on selected visemes. 3D movements of facial fleshpoints (see Figure 2) are acquired using photogrammetry and hand-fitted generic models. Visemes are selected by an analysis-by-synthesis technique [18] that combines robust automatic tracking with semi-automatic correction.

Our shape models are built using a so-called guided Principal Component Analysis (PCA) where a priori knowledge is introduced during the linear decomposition. We in fact compute and iteratively subtract predictors using carefully chosen data subsets [19]. For speech 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 and basic facial expressions but only components related to speech articulation are considered here.

We use here only the first 230 sentences for training and 10 sentences for testing. The average modeling error for training frames is less than half a millimeter for beads located on the lower face.

4. The trajectory formation system

The principle of speech synthesis by HMM was first introduced by Donovan for acoustic speech synthesis [20]. This was extended to audiovisual speech by the HTS working group [21]. The HMM-trajectory synthesis technique comprises training and synthesis parts.

4.1. Basic principles

An HMM and a duration model for each state are first learned for each segment of the training set. The input data for the HMM training is a set of observation vectors. The observation vectors consist of static and dynamic parameters, i.e. the values of articulatory parameters and their temporal derivatives. The HMM parameter estimation is based on ML (Maximum-Likelihood) criterion [22]. The ML estimation is achieved using a particular EM (Expectation Maximization) algorithm known as the Baum-Welch recursion algorithm. Usually, for each phoneme in context, a 3-state left-to-right model with single Gaussian diagonal output distributions. The state durations of each HMM are usually modeled as single Gaussian distributions. A second training step may also be added to factor out similar output distributions among the entire set of states (state tying).

The synthesis is performed as follows. The phonetic string to be synthesized is first chunked into segments and a sequence of HMM states is built by concatenating the corresponding segmental HMMs. State durations for the HMM sequence are determined so that the output probabilities of the state durations are maximized (thus usually by z-scoring) From the HMM sequence with the proper state durations assigned, a sequence of observation parameters is generated using a specific ML-based parameter generation algorithm [12].

4.2. Comments

This trajectory formation system exploits the dynamic parameters both in training and synthesis: the generated trajectory reflects both the means and covariances of the output distributions of a number of frames before and after each of the frames. By this way, this algorithm may incorporate implicitly part of short-term coarticulation patterns and inter-articulatory asynchrony. Larger coarticulation effects can also be captured since triphones intrinsically depend on adjacent phonetic context.

Note however that these coarticulation effects are anchored to acoustic boundaries that are imposed as synchronization events between the duration model and the HMM sequence. Intuitively we can suppose that context-dependent HMM can easily cope with this constraint. We show here that adding a context-dependent phasing model helps the trajectory formation system to better adjust to observed trajectories.

4.3. Adding and learning a phasing model

We propose to add a phasing model to the standard HMM-based trajectory formation system (see Figure 1) that consists in learning the time lag between acoustic and gestural units i.e. between acoustic boundaries delimiting allophones and gestural boundaries delimiting pieces of the articulatory score observed/generated by the context-dependent HMM sequence.
We test here a very simple phasing model: a unique time lag is associated with each context-dependent HMM. This lag is computed as the mean delay between acoustic boundaries and unconstrained alignment of triphones with articulatory trajectories of training utterances.

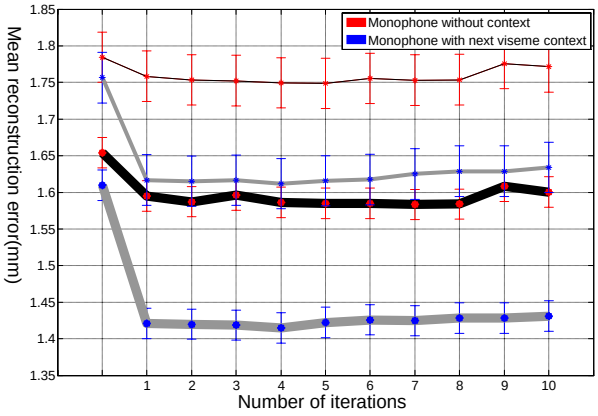

Figure 3: Mean reconstruction error as a function of number of iterations for context independent (black) and context-dependent phone HMMs (light gray). Results for training vs. test utterances are displayed respectively with thick vs. thin lines. Convergence is very fast and the phasing model benefits even more from contextual information.

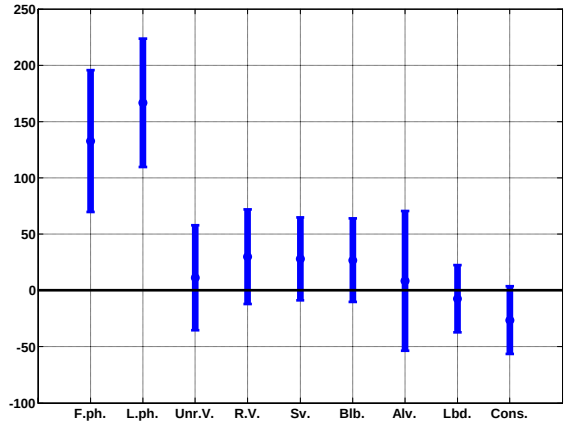

Figure 4: Average duration (ms) increase/decrease of the gestural segment with reference to its acoustic duration according to position and phoneme category. From left to right: first and final segment of the utterance, unrounded, rounded vowels, semivowels, bilabials, alveolars, labiodentals and remaining consonants.

5. Results

Figure 3 shows the significant decrease of prediction error when the phasing model is introduced in the HMM-based trajectory formation model. The convergence is obtained within 2 iterations: regularization constraints guarantying minimum durations of segments should be applied at least one time to avoid degeneration of the model.
Figure 4 shows that most gestural expansions occur at initial and final positions in the utterance (capturing prephonatory gestures and termination of phonation). Slow vocalic gestures generally expand whereas rapid consonantal gestures shrink: this is completely in accordance to the well-known numerical model of coarticulation proposed by Öhman [23] that superposes and blends vocalic and consonantal tongue gestures. The trajectory formation model places boundaries between segments so that dynamic information contained by observation probabilities of flanking HMM states best capture the variations of gestural speeds at the boundaries. Figure 5 gives an example of the necessary compromise between speech and duration: the large rounding gesture due to the semi-vowel [ɥ] is adequately predicted by the proposed system because the phasing model expands the duration of the gesture compared to the observed acoustic duration of the sound.

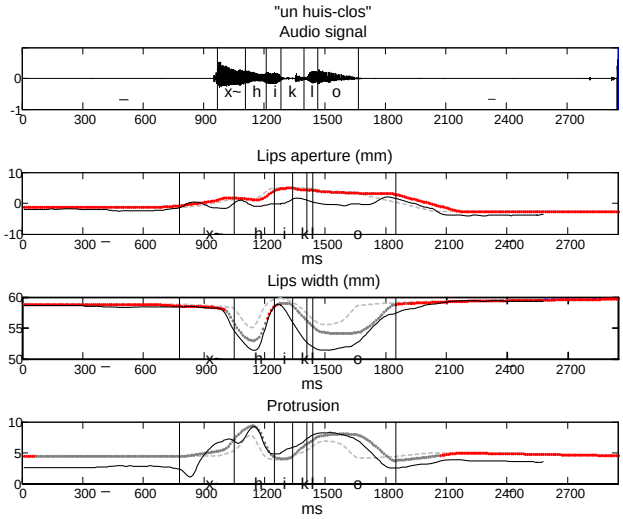

Figure 5. Comparing prediction of lip geometry by context-dependent HMMs trained either using acoustic (light gray) or gestural boundaries (dark gray) with original test data (black). The utterance is: ”un huis clos” [œ̃ɥiklo]. Note the expansion of initial and final movements (enabling the large final rounding movement) as well as the expansion of the semivowel [ɥ] with the following [i] shifted forward in time.

6. Conclusions

We have demonstrated here that the prediction accuracy of an HMM-based trajectory formation system can be greatly improved by modeling the phasing relations between acoustic and gestural boundaries. The phasing model is learned using an analysis-synthesis loop that uses constrained and unconstrained HMM alignments with the original data. We have shown that this scheme improves significantly the prediction error and captures subtle context-dependent anticipatory phenomena.

The interest of such an HMM-based trajectory formation system is double: (a) it provides accurate and smooth articulatory trajectories that can be used straightforwardly to control the articulation of a talking face or used as a skeleton to anchor multimodal concatenative synthesis [see notably the TDA proposal in 24]; (b) it also provides gestural segmentation as a by-product of the phasing model. These gestural boundaries can be used to segment original data for multimodal concatenative synthesis. This segmentation can also be used for asynchronous audiovisual speech recognition.

References

[1] N. F. Dixon and L. Spitz, "The detection of audiovisual desynchrony," *Perception*, vol. 9, pp. 719-721, 1980.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[3] D. H. Whalen, "Coarticulation is largely planned," *Journal of Phonetics*, vol. 18, pp. 3-35, 1990.

[4] K. W. Grant, V. van Wassenhove, and D. Poeppel, "Discrimination of auditory-visual synchrony," presented at Audio Visual Speech Processing, St Jorioz, France, 2003.

[5] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, pp. 331-346, 2003.

[6] J. Beskow, "Rule-based Visual Speech Synthesis," presented at Eurospeech, Madrid, Spain, 1995.

[7] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, D. Thalmann and N. Magnenat-Thalmann, Eds. Tokyo: Springer-Verlag, 1993, pp. 141-155.

[8] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," presented at IEEE Workshop on Speech Synthesis, Santa Monica, CA, 2002.

[9] T. Kaburagi and M. Honda, "A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes," *Journal of the Acoustical Society of America*, vol. 99, pp. 3154-3170, 1996.

[10] C. Weiss, "Framework for data-driven video-realistic audio-visual speech synthesis," presented at Int. Conf. on Language Resources and Evaluation, Lisbon, 2004.

[11] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda, "Visual speech synthesis based on parameter generation from HMM: speech-driven and text-and-speech-driven approaches," presented at Auditory-visual Speech Processing Workshop, Terrigal, Sydney, Australia, 1998.

[12] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," presented at ISCA Speech Synthesis Workshop, Pittsburgh, PE, 2004.

[13] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," presented at Human Language Technology Conference, San Diego, CA, 2002.

[14] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans. on Speech and Audio Processing*, 2005.

[15] T. Kuratate, K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia, "Audio-visual synthesis of talking faces from speech production correlates," presented at EuroSpeech, 1999.

[16] L. Revéret, G. Bailly, and P. Badin, "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," presented at International Conference on Speech and Language Processing, Beijing, China, 2000.

[17] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," presented at Auditory-Visual Speech Processing Workshop, Scheelsminde, Denmark, 2001.

[18] G. Bailly, F. Elisei, P. Badin, and C. Savariaux, "Degrees of freedom of facial movements in face-to-face conversational speech," presented at International Workshop on Multimodal Corpora, Genoa - Italy, 2006.

[19] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images," *Journal of Phonetics*, vol. 30, pp. 533-553, 2002.

[20] R. Donovan, "Trainable speech synthesis," in *Univ. Eng. Dept.* Cambridge, UK: University of Cambridge, 1996, pp. 164.

[21] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," presented at EUROSPEECH, Budapest, Hungary, 1999.

[22] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[23] S. E. G. Öhman, "Numerical model of coarticulation," *Journal of the Acoustical Society of America*, vol. 41, pp. 310-320, 1967.

[24] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw, "TDA: A new trainable trajectory formation system for facial animation," presented at InterSpeech, Pittsburgh, PE, 2006.