A trainable trajectory formation model TD-HMM parameterized for the LIPS 2008 challenge

Gérard Bailly[1], *Oxana Govokhina*[1,2], *Gaspard Breton*[2], *Frédéric Elisei*[1] *& Christophe Savariaux*[1]

[1] Department of Speech and Cognition, GIPSA-Lab, CNRS & Universities of Grenoble, France
[2] Orange R&D, 4 rue du Clos Courtel, BP 59 35512 Cesson-Sévigné - France
(gerard.bailly, frederic.elisei,christophe.savariaux)@gipsa-lab.inpg.fr
ogovokhina@yahoo.fr, gaspard.breton@orange-ftgroup.com

Abstract

We describe here the trainable trajectory formation model that will be used for the LIPS'2008 challenge organized at InterSpeech'2008. It predicts articulatory trajectories of a talking face from phonetic input. It basically uses HMM-based synthesis but asynchrony between acoustic and gestural boundaries – taking for example into account non audible anticipatory gestures – is handled by a phasing model 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. Using different motion capture data, we demonstrate here that the phasing model improves significantly the prediction error and captures subtle context-dependent anticipatory phenomena.

Index Terms: facial animation audiovisual speech synthesis, HMM

1. Introduction

Embodied conversational agents – virtual characters as well as anthropoid robots – should be able to talk with their human interlocutors. They should compute facial movements from symbolic input. Given history of the conversation and thanks to a model of the target language, dialog managers compute a phonetic string with phoneme durations. This minimal information can be enriched with details of the underlying phonological and informational structure of the message, 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 talking head (the shape and appearance models of a talking face or the control model of the robot).

Human interlocutors are very sensitive to discrepancies between the visible and audible consequences of articulation [7, 16] and have strong expectations on articulatory variability [22] 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 and still an unsolved problem.

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 [12]. 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. The TD-HMM system [10] proposes an original model for the re-estimation of phoneme-sized Hidden Markov Models (HMM). The HTS system [20] generate the final articulatory trajectory.

2. State-of the art

Numerous control models have been proposed for audiovisual text-to-speech synthesis [2]. The most popular solution consists in linking an animated head to an existing acoustic text-to-speech system. The trajectory formation model driving the head uses acoustic phoneme boundaries computed by the system to anchor the coarticulation model. Coarticulation is usually predicted using rules [5] or by exploiting an explicit coarticulation model [4, 6] that anchor the positions and spans of the phoneme-specific gestural targets. For predicting tongue movements captured by electromagnetic midsagittal articulography (EMMA), Kaburagi and Honda [14] add dynamic features in the specification of gestural targets of triphones in order to cope with inter-gestural phasing relations.

Data-driven trajectory formation systems automatically capture regularities of the context-dependent gestural realization of phoneme-sized segments [21]. Concatenative audiovisual speech synthesis encapsulates short-term and long-term coarticulation effects by storing multimodal segments depending on the size of the segments. The problem of possible asynchronies is thus pushed in the segmentation and smoothing of boundaries and eventually in the compression/expansion of segments. HMM are now used for speech synthesis and particularly as trajectory formation systems [19, 23]. HMM can in fact capture inter-gestural phasing relations thanks to the state-dependent static and dynamic probability density functions characterizing the HMM states.

A third possibility consists in computing articulation directly from speech signals. Proposals range from frame-based linear [15] or nonlinear models to GMM (Gaussian Mixture Model) -based or HMM-based mapping models that take as input a large speech window surrounding the current analysis frame [19]. 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.

A HMM-based trajectory formation system is here described. It includes a phasing model 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. We show that the modeling of audiovisual asynchrony has an impact on the performance of the whole system.

Figure 1. *125 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.*

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 [9, 17] that consists in collecting precise 3D data on selected visemes. 3D movements of facial fleshpoints (see Figure 1) are acquired using photogrammetry and hand-fitted generic models. Visemes are selected by an analysis-by-synthesis technique [3] 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 [1]. 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. The average modeling error is less than 0.5mm for beads located on the lower part of the face.

4. The trajectory formation system

The principle of speech synthesis by HMM was first introduced by Donovan for acoustic speech synthesis [8] and extended to audiovisual speech by the HTS working group [18]. 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 [20]. Usually, for each phoneme in context, a 3-state left-to-right model is estimated with single Gaussian diagonal output distributions. The state durations of each HMM are usually modeled as single Gaussian distributions. A second training step can also be added to factor out similar output distributions among the entire set of states (state tying). This step is not used here.

The synthesis is then performed as follows. A sequence of HMM states is built by concatenating the context-dependent phone-sized HMM corresponding to the input phonetic string. State durations for the HMM sequence are determined so that the output probabilities of the state durations are maximized (thus usually by z-scoring) Once the state durations have been assigned, a sequence of observation parameters is generated using a specific ML-based parameter generation algorithm [23].

4.2. Comments

States capture parts of the inter-articulatory asynchrony since transient and stable parts of the trajectories of different parameters are not obligatory modeled by the same state (this surely explains why complex HMM structures aiming at explicitly coping with audiovisual asynchronies do not outperform the basic ergodic structure [13]). Within a state articulatory dynamics is captured and is then reflected in the synthesized trajectory. By this way, this algorithm may capture 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.

These coarticulation effects are however 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 but we will show that adding a context-dependent phasing model helps the trajectory formation system to better fit 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 2) that learns 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 by the context-dependent HMM sequence.

We use 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.

This delay is learnt by an iterative process consisting of an analysis-synthesis loop:

1. Standard context-dependent HMM are learnt using acoustic boundaries as delimiters for gestural parameters
2. Once trained, forced alignment of training trajectories is performed (Viterbi alignment in Figure 2).
3. Deviations of the resulting segmentation with acoustic boundaries are collected. The average deviation of the right boundary of each context-dependent HMM is then computed and stored. The set of such mean deviations constitutes the phasing model.
4. New gestural boundaries are computed applying the current phasing model to the initial acoustic boundaries. Additional constraints are added to avoid collapsing: a minimal duration of 30 ms is guaranteed for each phone.

5. Experiments and results

All sentences are used for training. A leave-one-out process for TD-HMM has not been performed since a context-dependent HMM is built only if at least 10 samples are available in the training data; otherwise context-independent phone HMM are trained and used. TD-HMM is compared with concatenative synthesis using multi-represented diphones: synthesis of each utterance is performed simply by using all diphones of other utterances.

Figure 3 compares mean correlations obtained by the concatenative synthesis with the TD-HMM at each iteration. Convergence is obtained after typically 2 or 3 iterations. Figure 4 compares the articulatory trajectories obtained: the most important gain is obtained for silent articulations typically at the beginning (prephonatory gestures) and end of utterances.

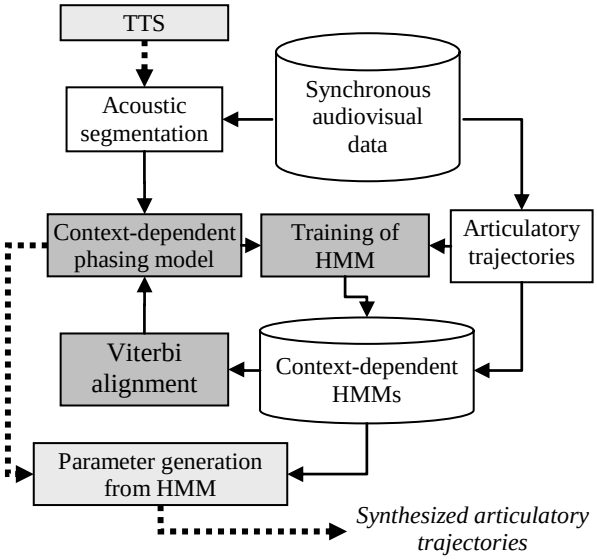

Figure 2. *Training consists in iteratively refining the context-dependent phasing model and HMMs (plain lines and dark blocks). The phasing model computes the average delay between acoustic boundaries and HMM boundaries obtained by aligning current context-dependent HMMs with training utterances. Synthesis simply consists in forced alignment of selected HMMs with boundaries predicted by the phasing model (dotted lines and light blocks).*

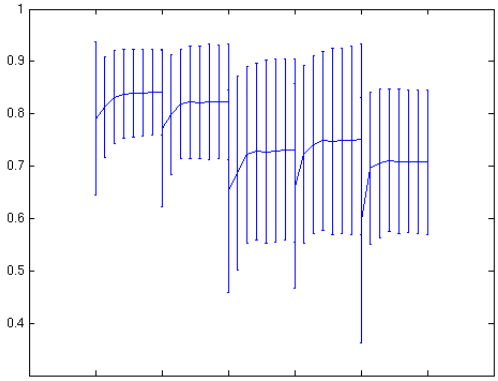

Figure 3: *Mean correlations (together with standard deviations) between original and predicted trajectories for the main five articulatory parameters (jaw rotation, lip rounding, lower and upper lip opening, jaw retraction). First data are predicted by concatenative synthesis using multi-represented diphones [4]. Second data are predicted by HMM using acoustic boundaries. The rest of the data give results obtained after the successive iteration of the estimations of the phasing model.*

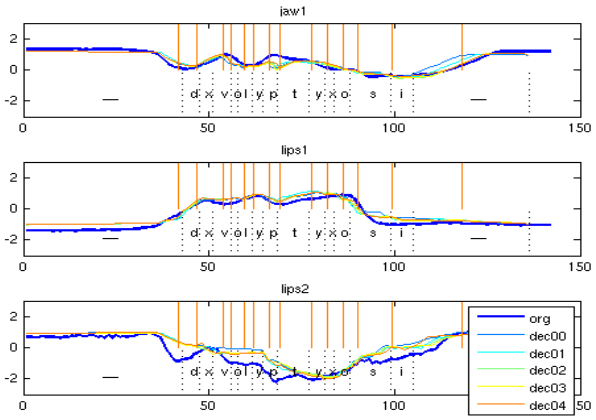

Figure 4: *Comparing natural and synthetic trajectories for the first 3 main articulatory parameters (jaw opening, lip spreading and lower lip opening). Vertical dashed lines at the bottom of each caption are acoustic boundaries while gestural boundaries are given by the top plain lines. Note the large delay of the non audible closure at the end of the utterance.*

6. Conclusions

We have demonstrated here that the prediction accuracy of an HMM-based trajectory formation system is improved by modeling the phasing relations between acoustic and gestural boundaries. The phasing model is learned using an analysis-synthesis loop that iterates HMM estimations and forced alignments with the original data. We have shown that this scheme improves significantly the prediction error and captures both strong (prephonatory gestures) and subtle (rounding) 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 11]; (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.

A more complex phasing model can also be build – using for example CART trees - by identifying phonetic or phonological factors influencing the observed lag between visible and audible traces of articulatory gestures.

We will use this trainable trajectory formation model TD-HMM for the LIPS'08 lipsync challenge and drive the articulated 3D clone (cf. Figure 5) with the articulatory scores computed using 3D data from one English female speaker under analysis at our laboratory.

Acknowledgements

The GIPSA-Lab/MPACIF team thanks Orange R&D for their financial support as well as the Rhône-Alpes region and the PPF “Multimodal Interaction”.

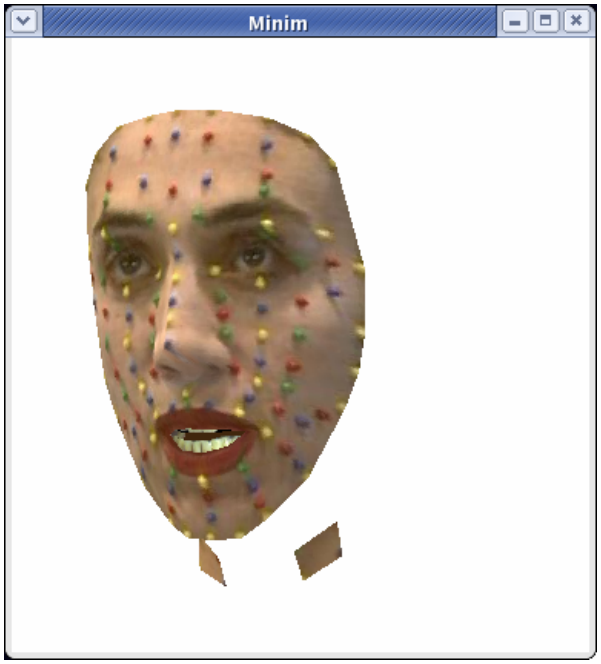

Figure 5: *The articulated 3D clone textured with the front image of Figure 1.*

References

[1] Badin, P., et al., 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] Bailly, G., et al., *Audiovisual speech synthesis.* International Journal of Speech Technology, 2003. **6**: p. 331-346.

[3] Bailly, G., et al. Degrees of freedom of facial movements in face-to-face conversational speech. in International Workshop on Multimodal Corpora. 2006. Genoa - Italy. p. 33-36.

[4] Bailly, G., G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. in IEEE Workshop on Speech Synthesis. 2002. Santa Monica, CA. p. 27-30.

[5] Beskow, J. *Rule-based Visual Speech Synthesis*. in *Eurospeech*. 1995. Madrid, Spain. p. 299-302.

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

[7] Dixon, N.F. and L. Spitz, *The detection of audiovisual desynchrony*. Perception, 1980. **9**: p. 719-721.

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

[9] Elisei, F., et al. Creating and controlling video-realistic talking heads. in Auditory-Visual Speech Processing Workshop. 2001. Scheelsminde, Denmark. p. 90-97.

[10] Govokhina, O., G. Bailly, and G. Breton. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. in ISCA Speech Synthesis Workshop. 2007. Bonn, Germany.

[11] Govokhina, O., et al. TDA: A new trainable trajectory formation system for facial animation. in InterSpeech. 2006. Pittsburgh, PE. p. 2474-2477.

[12] Grant, K.W., V. van Wassenhove, and D. Poeppel. *Discrimination of auditory-visual synchrony*. in *Audio Visual Speech Processing*. 2003. St Jorioz, France. p. 31-35.

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

[14] Kaburagi, T. 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, 1996. **99**(5): p. 3154-3170.

[15] Kuratate, T., et al. Audio-visual synthesis of talking faces from speech production correlates. in EuroSpeech. 1999. p. 1279-1282.

[16] McGurk, H. and J. MacDonald, *Hearing lips and seeing voices.* Nature, 1976. **264**: p. 746-748.

[17] 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.

[18] Tamura, M., et al. Text-to-audio-visual speech synthesis based on parameter generation from HMM. in EUROSPEECH. 1999. Budapest, Hungary. p. 959–962.

[19] Tamura, M., et al. Visual speech synthesis based on parameter generation from HMM: speech-driven and text-and-speech-driven approaches. in Auditory-visual Speech Processing Workshop. 1998. Terrigal, Sydney, Australia. p. 219-224.

[20] Tokuda, K., et al. Speech parameter generation algorithms for HMM-based speech synthesis. in IEEE International Conference on Acoustics, Speech, and Signal Processing. 2000. Istanbul, Turkey. p. 1315–1318.

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

[22] Whalen, D.H., *Coarticulation is largely planned.* Journal of Phonetics, 1990. **18**(1): p. 3-35.

[23] Zen, H., K. Tokuda, and T. Kitamura. An introduction of trajectory model into HMM-based speech synthesis. in ISCA Speech Synthesis Workshop. 2004. Pittsburgh, PE. p. 191-196.