A new trainable trajectory formation system for facial animation

Oxana Govokhina [1,2], Gérard Bailly [2], Gaspard Breton [2] and Paul Bagshaw [2]
[1] Institut de la Communication Parlée, 46 av. Félix Viallet, F38031 Grenoble
[2] France Telecom R&D, 4 rue du Clos Courtel, F35512 Cesson-Sévigné

Abstract

A new trainable trajectory formation system for facial animation is here proposed that dissociates parametric spaces and methods for movement planning and execution. Movement planning is achieved by HMM-based trajectory formation. Movement execution is performed by concatenation of multi-represented diphones. Movement planning ensures that the essential visual characteristics of visemes are reached (lip closing for bilabials, rounding and opening for palatal fricatives, etc) and that appropriate coarticulation is planned. Movement execution grafts phonetic details and idiosyncratic articulatory strategies (dissymetries, importance of jaw movements, etc) to the planned gestural score.

Introduction

The modelling of coarticulation is in fact a difficult and largely unsolved problem (Hardcastle and Hewlett 1999). The variability of observed articulatory patterns is largely planned (Whalen 1990) and exploited by the interlocutor (Munhall and Tohkura 1998). Since the early work of Öhman on tongue movements (1967), several coarticulation models have been proposed and applied to facial animation. Bailly et al (Bailly, Gibert et al. 2002) implemented some key proposals and confronted them to ground-truth data: the concatenation-based technique was shown to provide audiovisual integration close to natural movements. The HMM-based trajectory formation technique was further included (Govokhina, Bailly et al. 2006). It outperforms both objectively and subjectively the other proposals. In this paper we further tune the various free parameters of the HMM-based trajectory formation technique using a large motion capture database (Gibert, Bailly et al. 2005) and compare its performance with the winning system of Bailly et al study. We finally propose a system that aims at combining the most interesting features of both proposals.

Audiovisual data and articulatory modelling

The models are benchmarked using motion capture data. Our audiovisual database consists of 238 (228 for training and 10 for test) French utterances

Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece.

spoken by a female speaker. Acoustic and motion capture data are recorded synchronously using a Vicon© system with 12 cameras (Gibert, Bailly et al. 2005). The system delivers the 3D positions of 63 infra-red reflexive markers glued on the speaker's face at 120 Hz (see Figure 1). The acoustic data is segmented semi-automatically into phonemes. An articulatory model is built using a statistical analysis of the 3D positions of 63 feature points. The *cloning* methodology developed at ICP (Badin, Bailly et al. 2002; Revéret, Bailly et al. 2000) consists of an iterative Principal Component Analysis (PCA) performed on pertinent subsets of feature points. First, jaw rotation and protrusion (*Jaw*1 and *Jaw*2) are estimated from the points on jaw line and their effects subtracted from the data. Then the lip rounding/spreading gesture (*Lips*1), the proper vertical movements of upper and lower lips (*Lips*2 and *Lips*3), of the lip corners (*Lips*4) and of the throat (*Lar*1) are further subtracted from the residual data. These parameters explain 46.2, 4.6, 18.7, 3.8, 3.2, 1.6 and 1.3% of the movement variance.

The analysis of geometric targets of the 5690 allophones produced by the speaker (see Figure 2) reveals confusion trees similar to previous findings (Odisio and Bailly 2004). Consequently 3 visemes are considered for vowels (grouping respectively rounded [ʊψɔ̃], mid-open [ieøoã] and open vowels [aɛɔœɶɛ̃]) and 4 visemes for consonants (distinguishing respectively bilabials [pbm], labiodentals [fv], rounded fricatives [ʒʃ] from the others).

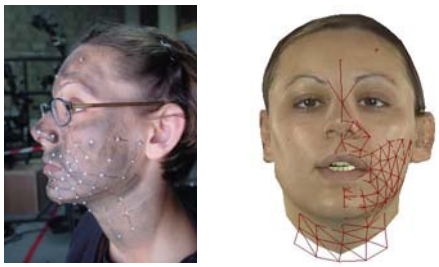


Figure 1: Motion capture data and videorealistic clone mimicking recorded articulation.

HMM-based synthesis

The principle of HMM-based synthesis was first introduced by Donovan for acoustic speech synthesis (Donovan 1996). This was extended to audiovisual speech by the HTS working group (Tamura, Kondo et al. 1999).

Training. 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 derivatives. The HMM parameter estimation is based on ML (Maximum-Likelihood) criterion (Tokuda, Yoshimura et al. 2000). Here, for each pho-

neme in context, a 3-state left-to-right model with single Gaussian diagonal output distributions and no skips is learned.

Synthesis. 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 probability of the state durations is maximized. From the HMM sequence with the proper state durations assigned, a sequence of observation parameters is generated using a specific ML-based generation algorithm (Zen, Tokuda et al. 2004).

Note that HHM synthesis imposes some constraints on the distribution of observations for each state. The ML-based parameter generation algorithm requires Gaussian diagonal output distributions. It thus best operates on an observation space that has compact targets and characterizes targets with maximally independent parameters. We compared the dispersion of visemes obtained using different observation spaces: articulatory vs. geometric. Only lip geometry (aperture, width and protrusion) is considered. Despite its lower dimension, the geometric space provides less confusable visemes.

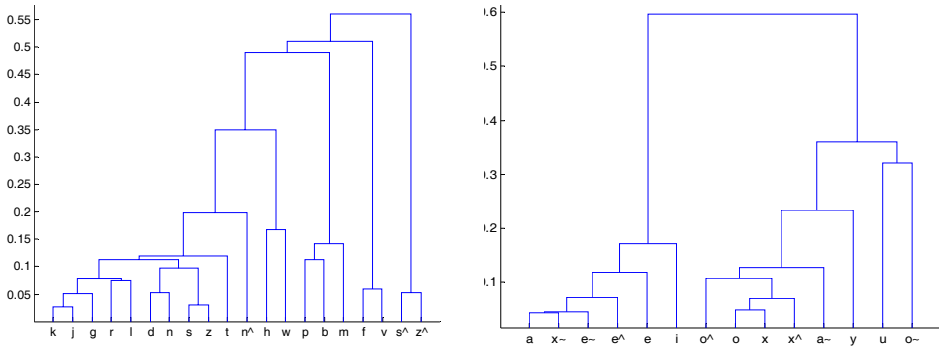


Figure 2. Grouping phonemes into viseme classes according to geometric confusability. Left: consonantal targets. Right: vocalic targets.

Detailed analysis

We compared phoneme-HHM with and without contextual information for selection. Table 1 summarizes our findings: anticipatory coarticulation is predominant, grouping context into visemes does not degrade performance. This contextual information enables the HMM system to progressively capture variability of allophonic realizations (see Figure 3). Syllable boundaries are known to influence coarticulation patterns. For this data, however, adding presence/absence of syllabic boundary does not improve the results (see bottom of Table 1). Sentence-internal (syntactic) pauses behave quite differ-

ently from initial and final pauses: initial pauses are characterized visually by prephonatory lips opening that reveals presence of initial bilabial occlusives if any; final pauses are characterized by a complete closure whereas the mouth often remains open during syntactic pauses especially when occurring between open sounds. We show that the viseme class immediately following the syntactic pause provides efficient contextual information for predicting lip geometry (see Table 2).

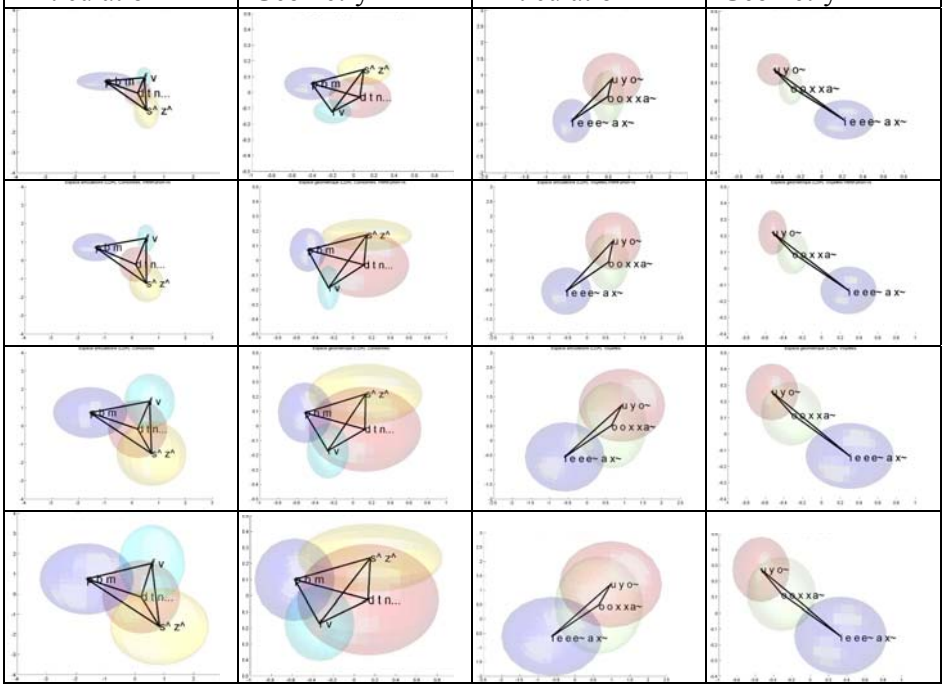
Consonantal visemes		Vocalic visemes	
Articulation	Geometry	Articulation	Geometry
			

Figure 3. Projecting the consonantal and vocalic visemes on the first discriminant plane (set using natural reference) for various systems and two different parametric spaces: articulatory versus geometric. From top to bottom: phoneme HMM, phoneme HMM with next segment information, TDA and natural reference.

Table 1: Adding contextual information to an initial context-independent phoneme HMM. Mean correlation (±standard deviation) between observed and predicted trajectories using different phoneme HMM systems for geometric space; coverage (nb. of segments which number of samples is superior to ten divided by total nb. of segments) and mean nb. of samples (±standard deviations) are computed.

Phoneme HMM	Correlation	Coverage	Mean nb. of samples
Without context	0.77±0.07	1.00	164±112
Prev. phoneme	0.78±0.09	0.13	20±11
Next phoneme	0.83±0.06	0.13	20±11
Next viseme	0.83±0.07	0.23	31±35
adding syllable	0.84±0.06	0.12	28±26

Table 2: Mean correlations (±standard deviations) between the targets of the sentence-internal pauses and the targets of next (or previous) segment.

Target	Articulation	Geometry
Next	0.76±0.04	0.80±0.07
Previous	0.43±0.10	0.40±0.13

Table 3: Mean correlations (±standard deviations) between observed and predicted trajectories using different systems and representations.

System	Articulation	Geometry
Phoneme-HMM	0.61±0.11	0.77±0.07
Contextual phoneme-HMM	0.69±0.10	0.83±0.07
Concatenation of diphones	0.61±0.15	0.78±0.07
Concatenation with HMM selection	0.63±0.15	0.81±0.06
TDA	0.59±0.16	0.81±0.06

The proposed trajectory formation system

TDA (Task Dynamics for Animation), the proposed trajectory formation system, combines the advantages of both HMM- and concatenation-based techniques. The proposed system (see Figure 4) is motivated by articulatory phonology and its first implementation by the task dynamics model (Saltzman and Munhall 1989). Articulatory phonology put forward underspecified gestures as primary objects of both speech production and perception. In the task dynamics model, context-independent underspecified gestures first give spatio-temporal gauges of vocal tract constrictions for each phoneme. Then a trajectory formation model executes this gestural score by moving articulatory parameters shaping the vocal tract. In this proposal, the gestural score specifying the lip geometry (lip opening, width and protrusion) is first computed by HMM models. Then execution of this score is performed by a concatenation model where the selection score penalizes

segments according to their deviation from this planned geometry. The stored segments are thus characterized both by lip geometry for selection and by detailed articulation (jaw, separate control of upper and lower lips as well as rounding, etc) for the final generation.

Planning gestures by HMM synthesis. HMM-based synthesis outperforms both in objective and subjective terms concatenative synthesis and phoneme or diphone HMMs, when all these systems are trained to generate directly articulatory parameters. When trained on geometric parameters, these systems generate targets that are more discriminated and the correlation between original trajectories and those generated by all systems is substantially higher when considering geometry (see Table 3). This confirms previous studies that promote constrictions as the best characteristics for speech planning (Bailly 1998).

Executing gestures by concatenative synthesis. While diphone HMMs generate smooth trajectories while preserving visually relevant phonetic contrasts, concatenative synthesis has the intrinsic properties of capturing inter-articulatory phasing and idiosyncratic articulation. Concatenative synthesis also intrinsically preserves the variability of natural speech.

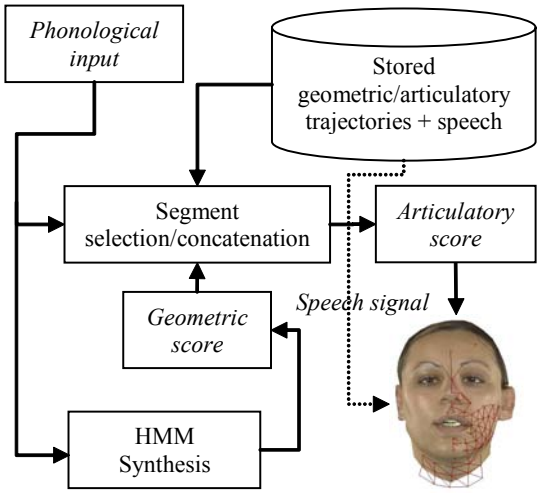


Figure 4: The proposed trajectory formation system TDA. A geometric score is thus computed by HMM-based synthesis. Segments are then retrieved that best match this planned articulation. Articulatory trajectories also stored in the segment dictionary are then warped, concatenated and smoothed and drive the talking head. Since the speech signal is generated using the same warping functions, audiovisual coherence of synthetic animation is preserved.

Performance analysis

Table 3 summarizes the comparative performance of the different systems implemented so far. Performance of the concatenation system is substantially increased when considering a selection cost using target parameters computed HMM trajectory planner. This is true whenever considering geometry or articulatory planning space. The performance of the current implementation of the TDA is however deceptive: the articulatory generation

often degrades the quality of the planned geometric characteristics. If the TDA compensates well for the bad planning of movement during syntactic pauses, it often degrades the timing (see Figure 5). We are currently reconsidering the procedure that warps stored articulatory segments to planned gestures.

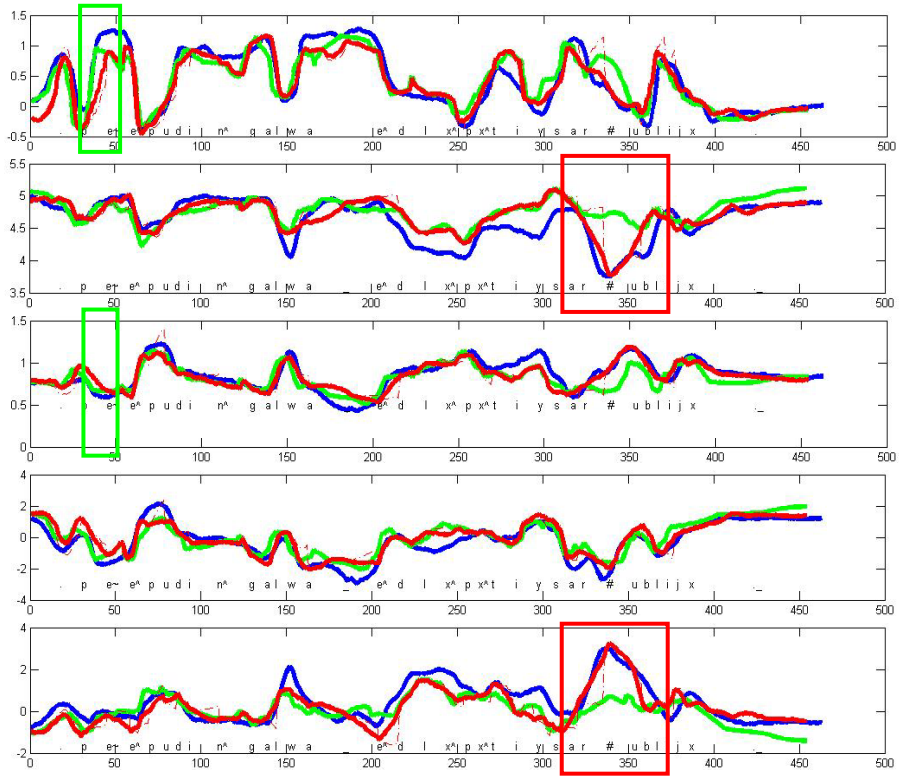


Figure 5. Comparing trajectory formation systems (blue: natural reference; red: concatenation/selection TDA; green: contextual phoneme-HMM) with a natural test stimulus (blue). From top to bottom: geometric parameters: lip aperture, width and protrusion; articulatory parameters: jaw aperture, lips rounding/spreading. Major discrepancies between TDA and contextual phoneme-HMM are enlighten.

Conclusions and perspectives

The TDA system is a trajectory formation system for generating speech-related facial movement. It combines a HMM-based trajectory formation system responsible for planning long-term coarticulation in a geometric space with a trajectory formation system that selects and concatenates segments that are best capable of realizing this gestural score. Contrary to most pro-

posals, this system builds on motor control theory – that identifies distinct modules for planning and execution of movements – and implements a theory of control of speech movements that considers characteristics of vocal tract geometry as primary cues of speech planning.

In the near future we will exploit in a more efficient way the information delivered by the HMM-based synthesis e.g. adding timing and spatial gauges to the gestural score in order to guide more precisely the segment selection.

References

Badin, P., G. Bailly, L. Revéret, M. Baciu, C. Segebarth and C. Savariaux (2002). "Three-dimensional linear articulatory modelling of tongue, lips and face based on MRI and video images." Journal of Phonetics **30** (3): 533-553.

Bailly, G. (1998). "Learning to speak. Sensori-motor control of speech movements." Speech Communication **22** (2-3): 251-267.

Bailly, G., G. Gibert and M. Odisio (2002). Evaluation of movement generation systems using the point-light technique. IEEE Workshop on Speech Synthesis, Santa Monica, CA: 27-30.

Donovan, R. (1996). Trainable speech synthesis. PhD thesis. Univ. Eng. Dept. Cambridge, UK, University of Cambridge**:** 164 p.

Gibert, G., G. Bailly, D. Beautemps, F. Elisei and R. Brun (2005). "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 **118** (2): 1144-1153.

Govokhina, O., G. Bailly, G. Breton and P. Bagshaw (2006). Evaluation de systèmes de génération de mouvements faciaux. Journées d'Etudes sur la Parole, Rennes - France: accepted.

Hardcastle, W. J. and N. Hewlett (1999). Coarticulation: Theory, Data, and Techniques. Cambridge, UK, Press Syndicate of the University of Cambridge.

Munhall, K. G. and Y. Tohkura (1998). "Audiovisual gating and the time course of speech perception." Journal of the Acoustical Society of America **104**: 530-539.

Odisio, M. and G. Bailly (2004). "Tracking talking faces with shape and appearance models." Speech Communication **44** (1-4): 63-82.

Öhman, S. E. G. (1967). "Numerical model of coarticulation." Journal of the Acoustical Society of America **41**: 310-320.

Revéret, L., G. Bailly and P. Badin (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. International Conference on Speech and Language Processing, Beijing - China: 755-758.

Saltzman, E. L. and K. G. Munhall (1989). "A dynamical approach to gestural patterning in speech production." Ecological Psychology **1** (4): 1615-1623.

Tamura, M., S. Kondo, T. Masuko and T. Kobayashi (1999). Text-to-audio-visual speech synthesis based on parameter generation from HMM. EUROSPEECH, Budapest, Hungary: 959–962.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura (2000). Speech parameter generation algorithms for HMM-based speech synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey: 1315–1318.

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

Zen, H., K. Tokuda and T. Kitamura (2004). An introduction of trajectory model into HMM-based speech synthesis. ISCA Speech Synthesis Workshop, Pittsburgh, PE: 191-196.