TDA: A new trainable trajectory formation system for facial animation

Oxana Govokhina[(1,2)] *Gérard Bailly*[(1)] *Gaspard Breton*[(2)] *Paul Bagshaw*[(2)]

[(1)] Institut de la Communication Parlée, 46, av. Félix Viallet, 38031 Grenoble Cedex, France
[(2)] France Telecom R&D, 4 rue du Clos Courtel, BP 59, F35512 Cesson-Sévigné Cedex
{oxana.govokhina,gerard.bailly}@icp.inpg.fr, {gaspard.breton,paul.bagshaw}@francetelecom.fr

Abstract

A new trainable trajectory formation system - named TDA - 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. This module essentially plans configurations of lip geometry (aperture, spreading and protrusion). Movement execution is performed by concatenation of multi-represented diphones. This module is responsible for selecting and concatenating detailed facial movements that best obey to the target kinematics of the geometry previously planned. 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. This planning scheme is compared to alternative planning strategies using articulatory modeling and motion capture data.

Index Terms: visual speech synthesis, facial animation.

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 speak with human partners. This symbolic input minimally consists in the phonetic string with phonemic 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 build 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 very sensitive to discrepancies between the visible and audible consequences of articulation [5, 12] and have strong expectations on articulatory variability [18, 23] resulting from the underspecification of articulatory targets and planning. The proper modeling of coarticulation in speech benefits to the intelligibility of the agent and is in fact a challenging issue for trajectory formation systems.

We propose here a trajectory formation system that builds on the task dynamics model [17] but combines two up-to-date trajectory formation systems operating on two different representation spaces of the movement for handling separately movement planning and execution.

2. State-of-the-art

The most popular trajectory formation system used in facial animation computes a sequence of contextual articulatory targets that are connected or weighted by temporal functions. The most simple model consists in blending prototypical phonemic targets according to phoneme-specific coarticulation functions [4]. Similarly Reveret et al [16] have adapted the more speech-specific Ohman's model [14] that distinguishes between vowel- and consonant-specific coarticulation functions. Statistical N-phones models only considering one target per phoneme have also been proposed. The trajectory is then built by adding general constraints on movements such as minimal jerk [7]. For tongue control, Okadome et al [15] added dynamic parameters sampled at targets in order to cope with inter-gestural phasing. This could be considered as the first step towards more general statistical modeling of phoneme-specific gestures in context using the generation abilities of HMM [19]

The availability of large motion capture data and the possibility of storing and retrieving video segments has also popularized concatenation-based techniques [10, 13].

Please note finally the use of inverse acoustics for lip-sync [24] that should possibly exploit the acoustic trace of coarticulation to recover the visible one.

Using point-light displays, Bailly et al [2] compared key proposals using common training and test material. More recently Govokhina et al [8] added an HMM-based trajectory formation to the set of key proposals driving a full videorealistic talking head. In both studies HMM- and concatenation-based techniques reach a level of adequacy close to original motion. The long-term coherence of the trajectory offered by HMM-based technique seems however to outperform a pure concatenation.

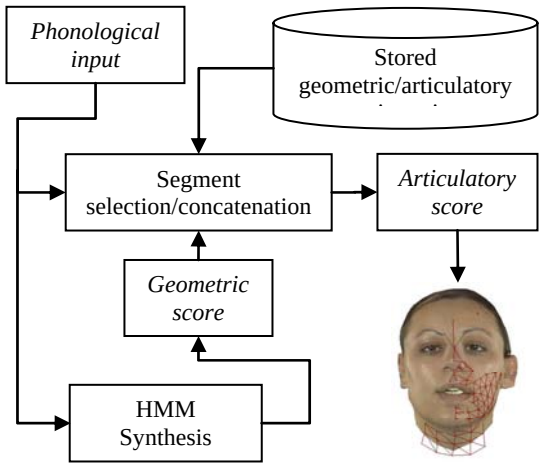

Figure 1. The proposed trajectory formation system TDA.

3. Trainable trajectory formation systems

We sketch below the basic features of these two trajectory formation systems.

3.1. Concatenation of multi-represented segments

Synthesis by concatenation consists in selecting and concatenating pre-recorded segments. Phonological features are first used to select candidate segments: besides a simple phonemic match, phonotactic constraints (phonemic context, position in the syllable or word) and/or compliance with higher-level phonological structure [20] are often added. Then a DTW algorithm find an optimal path through this lattice of

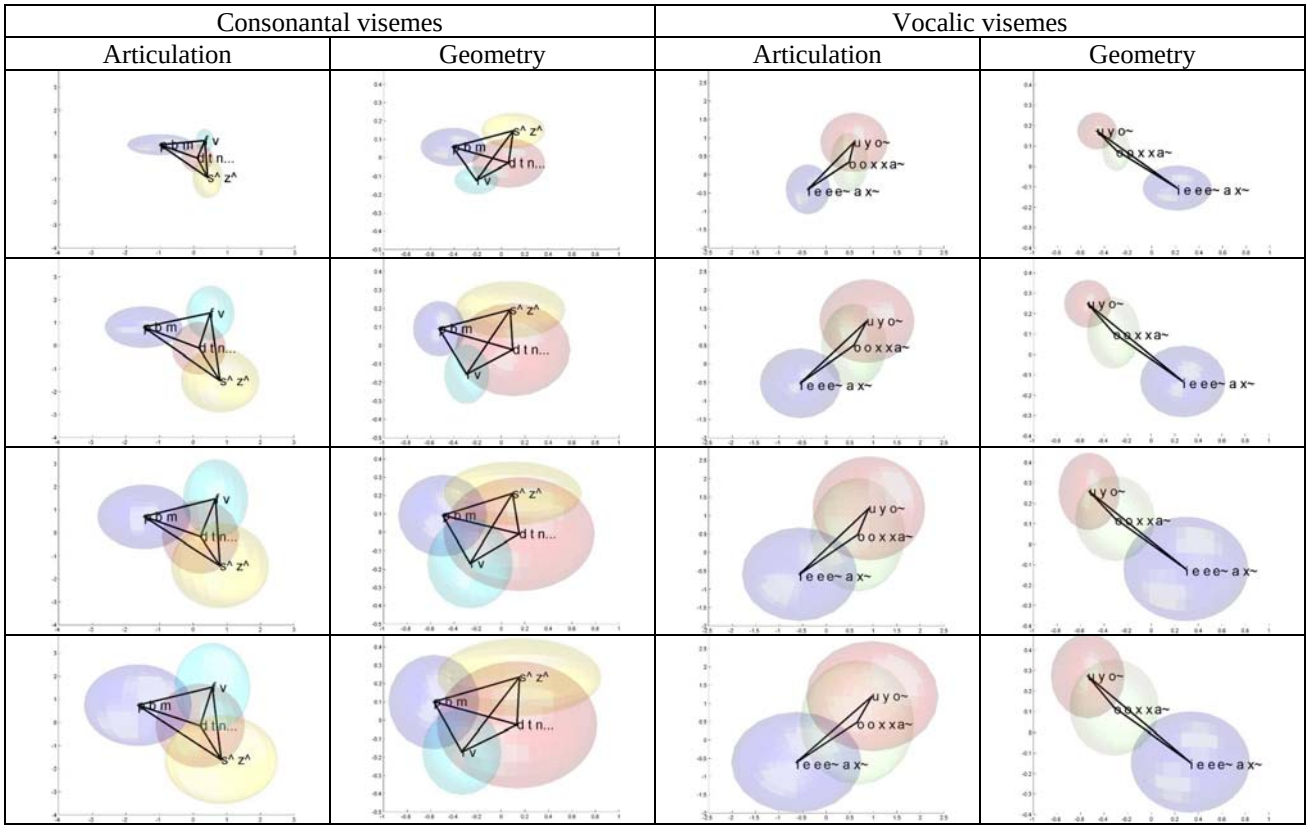

Figure 2. Projecting the target consonantal and vocalic visemes on the first discriminant plane (set using natural reference) for various systems and two different parametric representations: articulatory versus geometric. From top to bottom: phoneme HHM, diphone HMM, concatenative synthesis and natural reference. Targets are more discriminated using the geometric representation using whatever system.

candidate units that best gathers units with minimal selection and concatenation costs.

Two penalties are often considered to build the selection cost: a penalty that depends on the degree of adequacy of the considered unit with the phonological constraints and a penalty that considers the distance between parameters characterizing the selected unit with parameters computed with an external prediction model. So parameters computed by a prosodic model are often considered [11].

The concatenation cost depends on the match between parameters of adjacent units across the boundary. Both static and dynamic cues are often considered.

3.2. HMM-based synthesis

The principle of speech synthesis by HMM was first introduced by Donovan for acoustic speech synthesis [6]. This was extended to audiovisual speech by the HTS working group [19]. The HMM-trajectory synthesis technique comprises training and synthesis parts.

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 [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 and no skips is learned. The state durations of each HMM are modeled by single Gaussian distributions. A second training step may also be added to factor out similar output distributions among the entire set of states.

Synthesis. 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 probability of the state durations are maximized [25]. 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 [26]. This algorithm exploits the dynamic parameters included in the observations in training as well as in 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 long-term coarticulation patterns.

3.3. Comments

Note here that HHM synthesis imposes some constraints on the distribution of observations for each state. The ML-based parameter generation algorithm requires single Gaussian diagonal output distributions [note however the use of a Gaussian mixtures in 21]. It will thus best operate on an observation space that has compact targets and characterize targets with maximally independent parameters.

4. The proposed trajectory formation system

TDA (Task Dynamics for Animation), the trajectory formation system we propose, combines the advantages of both HMM- and concatenation-based techniques. The proposed system is motivated by articulatory phonology [3] and its first implementation by the task dynamics model [17]. Articulatory phonology put forward underspecified gestures as primary

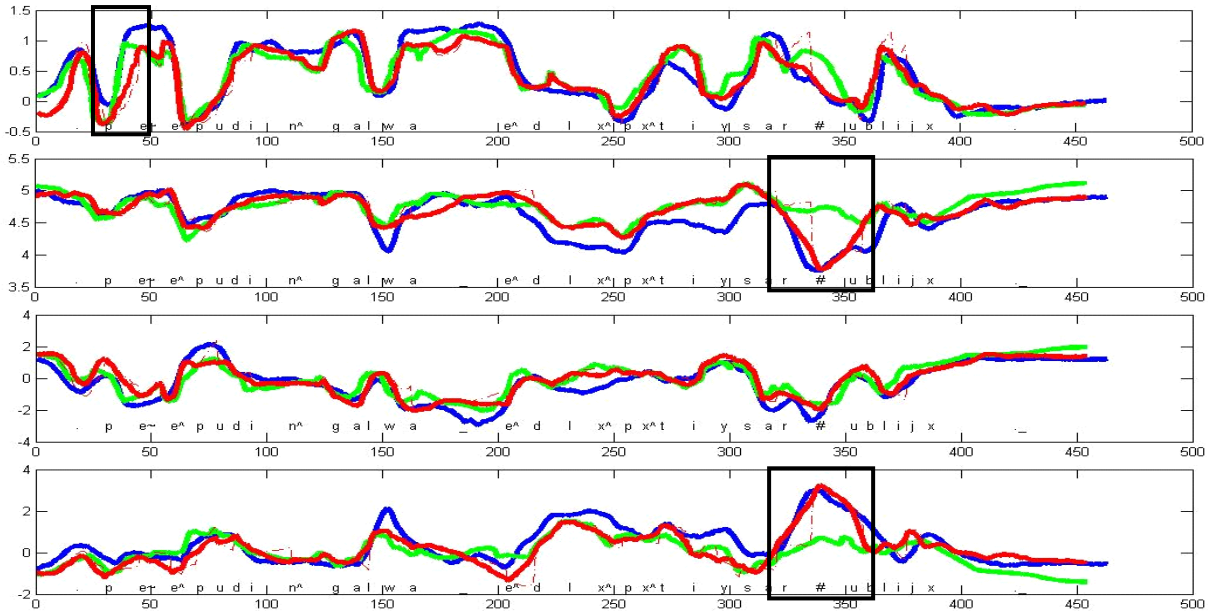

Figure 3. Comparing trajectory formation systems (blue: natural reference; red: concatenation/selection TDA; green: contextual phoneme-HMM) with a natural test stimulus (blue).From top to bottom: geometric parameters: lip aperture, width and protrusion; articulatory parameters: jaw aperture, lips rounding/spreading. Major discrepancies between TDA and contextual phoneme-HMM are enlighten

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 our 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. The synopsis of the system is depicted Figure 1.

Table 1: 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 guide	0.63±0.15	0.81±0.06
TDA	0.59±0.16	0.81±0.06

Planning gestures by HMM synthesis. We have shown elsewhere [9] that trajectory formation based on context-dependent phone HMMs (context is here limited to the following viseme – 3 visemes for vowels and 4 for consonants, see Figure 2) 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 also targets that are more discriminated (see Figure 2). The correlation between original trajectories and those generated by all systems is substantially higher when considering geometry (see Table 1 and Figure 3). This confirms previous studies that promote constrictions as the best characteristics for speech planning [1].

Executing gestures by concatenative synthesis. If 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 (see Figure 2 and Figure 3).

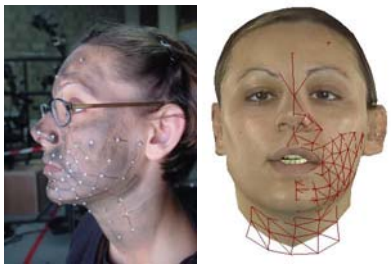

Figure 4: Motion capture data and videorealistic clone mimicking recorded articulation.

5. Evaluation

We compare the adequacy of different trajectory formation systems in generating articulatory trajectories that best integrate with a natural audio sound given its phonemic transcription and phoneme durations (see Table 1). The TDA system is compared to the original articulation and four other systems that directly generate articulatory trajectories: phoneme-HMM, contextual phoneme-HMM and concatenation of diphones with or without weighting candidates by their similarity with trajectories first computed by the contextual phoneme-HMM. The 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 3). We are

currently reconsidering the procedure that warps stored articulatory segments to planned gestures

6. 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 proposals, 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.

This clear dichotomy between planning and execution provides a possible route towards sets of MPEG4 compatible talking faces where the encoder specifies geometric features and the decoder is responsible for computing the speaker-specific facial deformations given these constraints.

The TDA system parallels proposals made for acoustic synthesis where a prosodic model helps a concatenative speech synthesis system for selecting appropriate acoustic segments and maintains a global structural coherence of the synthetic stimuli. In the future we will examine the possibility to extend our approach to joint audiovisual synthesis and enhance the ability of our planning module to cope with long-term coarticulation patterns.

References

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

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

[3] Browman, C.P. and Goldstein, L.M. (1989) *Articulatory gestures as phonological units.* Phonology, **6**: p.201-251.

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

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

[6] Donovan, R. (1996) *Trainable speech synthesis*. PhD thesis, in *Univ. Eng. Dept.* University of Cambridge: Cambridge, UK. 164 pages.

[7] Ezzat, T., Geiger, G., and Poggio, T. (2002) *Trainable videorealistic speech animation.* ACM Transactions on Graphics, **21**(3): p.388-398.

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

[9] Govokhina, O., Bailly, G., Breton, G., and Bagshaw, P. (2006) *A new trainable trajectory formation system for facial animation*. in *ISCA Workshop on Experimental Linguistics*. Athens, Greece

[10] Huang, F.J., Graf, H.P., and Cosatto, E. (2002) *Triphone-based unit selection for concatenative visual speech synthesis*. in *International Conference on Acoustics, Speech and Signal Processing.* Orlando, FL. p.1315–1318.

[11] Hunt, A.J. and Black, A.W. (1996) *Unit selection in a concatenative speech synthesis system using a large speech database*. in *International Conference on Acoustics, Speech and Signal Processing.* Atlanta, GA. p.373-376.

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

[13] Minnis, S. and Breen, A.P. (1998) *Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis.* in *International Conference on Speech and Language Processing*. Beijing, China. p.759-762.

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

[15] Okadome, T., Kaburagi, T., and Honda, M. (1999) *Articulatory movement formation by kinematic triphone model*. in *IEEE International Conference on Systems Man and Cybernetics,*. Tokyo, Japan. p.469-474.

[16] Revéret, L., Bailly, G., and Badin, P. (2000) *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*. Beijing - China. p.755-758.

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

[18] Sereno, J.A., Baum, A.R., Cameron Marean, G., and Lieberman, P. (1987) *Acoustic analysis and perceptual data on anticipatory labial coarticulation in adults and children.* Journal of Acoustical Society of America, **81**: p.512-519.

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

[20] Taylor, P. and Black, A.W. (1999) *Speech synthesis by phonological structure matching*. in *EuroSpeech*. Budapest, Hungary. p.1531-1534.

[21] Toda, T., Black, A.W., and Tokuda, K. (2004) *Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis*. in *International Speech Synthesis Workshop*. Pittsburgh, PA. p.26-31.

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

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

[24] Yehia, H., Kuratate, T., and Vatikiotis-Bateson, E. (2000) *Facial animation and head motion driven by speech acoustics.* in *5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*. Kloster Seeon, Germany. p.265-268.

[25] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1998) *Duration modeling for HMM-based speech synthesis.* in *International Conference on Spoken Language Processing*. Sydney. p.29–32.

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