Virtual Talking Heads and audiovisual articulatory synthesis

Pierre Badin, Gérard Bailly, Frédéric Elisei, and Matthias Odisio
Institut de la Communication Parlée, CNRS – INPG – Université Stendhal, Grenoble, France
E-mail: (badin, bailly, elisei, odisio)@icp.inpg.fr – Web: http://www.icp.inpg.fr

ABSTRACT

Our approach to audiovisual articulatory synthesis involves the development of *Virtual Talking Heads* that integrate the articulatory, aerodynamic and acoustic phenomena underlying speech production. Specifically, these Talking Heads are faithful *clones* of the speakers whose data the various models are based on. Our contribution presents some of the results achieved at ICP in this domain: 3D orofacial linear articulatory models, made possible by recent progresses in medical imaging and video processing techniques; aerodynamic and acoustic models, and basic glottis / oral constriction coordination principles; development of various strategies to determine the articulatory control parameters (coarticulation models vs. simple concatenative strategies); text-to-audiovisual speech synthesis. We finally make some suggestions for future developments.

1. INTRODUCTION

Articulatory synthesis of speech goes back to the sixties, when Coker et al. [1] developed at the Bell Laboratories what is likely to be the first complete articulatory synthesizer, and used it in a text-to-speech synthesis system for American English. Different versions of this synthesiser have been for many years the core of studies in speech production and perception at Haskins Laboratories [2]. More recently, Kröger et al. [3] designed an articulatory text-to-speech synthesis system for German that seems the most elaborate at present. Kaburagi & Honda [4] have also built a system that synthesises articulatory trajectories from phoneme sequences and produces speech based on an articulatory-acoustic dictionary that associates articulators shapes and corresponding spectra for a Japanese subject.

The novelty of recent years in articulatory modelling is the advent of 3D articulators and face models made possible by progresses in speech production imaging. In addition, a better understanding of aerodynamic speech production phenomena has allowed us to develop better speech synthesisers and to yield more realistic synthetic speech. The present paper describes recent advances in our "cloning" approach to speech production modelling, i.e. orofacial articulatory and aeroacoustic models based on real subjects data that are integrated into a *Virtual Talking Heads*, and various attempts to control them to produce high quality audiovisual speech.

2. OROFACIAL MODELLING

The speech production apparatus is made of a large number of neuromuscular components, which possess a large dimensionality but are functionally coupled in order to produce relatively simple gestures that can be referred to as *elementary articulators*. Speech can thus be thought to be generated by the carefully coordinated recruitment of such *articulators*. Thanks to the development and the availability of medical imaging techniques (e.g. Magnetic Resonance Imaging [MRI]) and of video processing systems, it has become possible to obtain 3D data on orofacial articulators (jaw, tongue, velum, lips, face,...) in appreciable amounts. The development of fairly realistic 3D orofacial models can thus be seriously envisaged today. The advantages of 3D orofacial articulatory models lie in their capability to produce area functions directly, to deal with lateral consonants and their lateral channels (which was not previously possible with traditional midsagittal models), and to display complete video-realistic faces, showing all articulators.

2.1 PRINCIPLES OF LINEAR ARTICULATORY MODELLING

Our approach to 3D articulatory modelling is described in detail in [5]. Each *elementary articulator* (a degree of freedom [DoF], in robotics terms) is defined by the simple set of movements that it can execute independently of the other articulators. Two basic assumptions underlie this approach: *linearity*, i.e. the shape of each speech organ is decomposed in a linear combination of elementary articulators, and *decorrelation*, i.e. weak or null cross correlations between the elementary articulators.

In our *data-driven*, *subject-specific* approach, we extract these DoFs from carefully designed corpora of articulatory data gathered on one subject using the same framework for tongue, lips and face, based on classical linear analysis techniques such as Principal Component Analysis (PCA) and linear regression analysis.

2.2 OROFACIAL DATA

As the elaboration of the linear components is entirely based on the data acquired from the subject, the corpus should constitute a representative sampling of the whole articulatory space of the subject. The corpus was thus made of the 10 French oral vowels, and of the artificially sustained consonants [p t k f s ʃ ʀ l] produced in three symmetric contexts [a i u]. The data needed for the tongue model were obtained by MRI, while the data for the lips / face model were acquired from videos of the subject's face.

For the tongue data, the subject was instructed to sustain each articulation for about 45 seconds, allowing thus the acquisition of sets of 53 MRI images orthogonal to the midsagittal plane. For each articulation, the tongue contours were then manually traced on each image in order to reconstruct an initial 3D shape of the tongue, and subsequently resampled along a dynamically adjustable semi-polar grid system that was fitted to the midsagittal tongue contour. This resulted in 22 plane contours, which constitute a structured 3D representation of tongue shape.

The lips / face data were extracted from simultaneous front and profile video recordings of the subject's face on which 168 flesh-points were marked by small plastic beads glued on the skin. In order ensure the maximum coherence between lips / face data and MRI data, the subject was instructed to produce, during the video recording, the artificially sustained articulations in much the same way as during the MRI recording session. Resulting video images were processed to extract three types of articulatory data: (1) the 3D coordinates of the 168 flesh points, reconstructed from the coordinates of the beads on both front and profile images; (2) the 3D coordinates of 30 points controlling a mesh that was manually adjusted to fit the lips shape optimally [6]; (3) the jaw position (coordinates of the lower incisors, inferred from chin beads coordinates).

2.3 ARTICULATORY MODELLING

The sets of 3D coordinates for the tongue and the lips / face data (respectively about 700 and 500 variables) were analysed and the main DoFs were extracted in order to built the linear articulatory models, as mentioned above (cf. also [5]).

The 3D tongue model is controlled by the five articulatory parameters *JH*, *TB*, *TD*, *TT*, and *TA*. The coordinates in the plane contours of the grid are computed as linear combinations of these five command parameters. The effects of some of these commands are demonstrated in Figure 1 which displays tongue shapes for two extreme values (–3 and +3) of one parameter, all other parameters being set to zero.

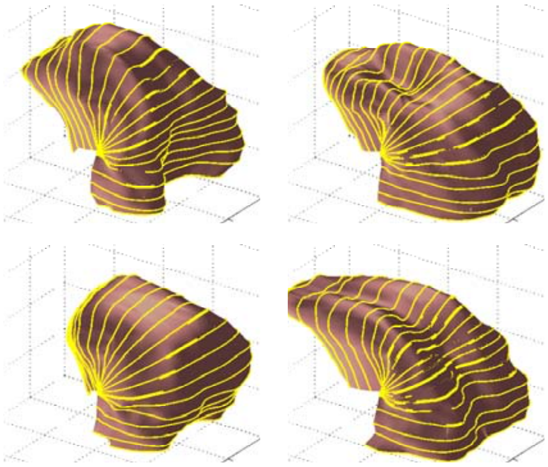

Figure 1: Nomograms of the tongue model for parameters *TB* (top) and *TD* (bottom) (left –3, right + 3).

JH controls the influence of jaw height on the tongue. The *front* / *back* displacement of the bulk of the tongue is associated with *TB*. Figure 1 illustrates the fact that much of the tongue groove characteristic of consonant [s] is achieved by this parameter. The *flat* / *arched* feature of tongue is taken into account by *TD*, and is also associated to some degree with tongue grooving (see Figure 1). The tongue tip is shaped by two parameters: (1) *TT* takes care of the global *up* / *down* movements of the last four sections of the tongue; *TT* is particularly active for [l^{a}] where the tongue body is lowered by the joint action of *JH* and *TB*, and the tongue tip / maxilla contact is ensured by the high value of *TT*; (2) parameter *TA* is associated with the advance of the tongue in the horizontal direction.

Interestingly, it was observed that the lateral consonant [l] seems mainly obtained by a depression of the tongue body achieved through a combination of jaw lowering, tongue body backing and tongue tip elevation: these movements that can be observed in the midsagittal plane are capable of creating the lateral channels characteristic of [l].

As for the tongue model, *JH* controls the first linear component of lips / face model. Three other parameters control components traditionally related to phonetic features of labiality: *LP* controls the protrusion – rounding gesture (cf. Figure 3); *LH* controls the aperture; *LV* controls the quasi-simultaneous vertical motion of both lips needed for the realization of labiodentals for this subject (and also for the open and protruded lips for consonants [ʃ ʒ]). Finally, a jaw advance parameter *JA* controls the effects of jaw advancing on lips and face.

In addition, a skin texture extracted from the subject's photos can be applied to the lips / face model to make it a video-realistic clone of the subject (cf. [7]). Moreover, the texture can be made semi- transparent so as to show the usually invisible articulators (see Figure 2).

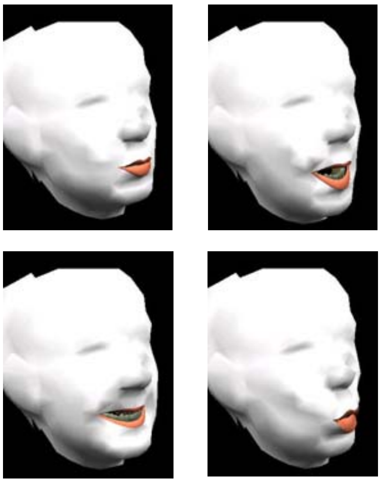

Figure 3: Nomogram of the lips / face model for parameters *JH* (top) and *LP* (bottom)(left –3, right + 3).

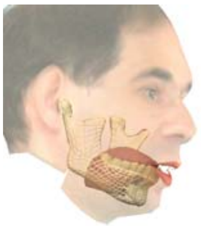

Figure 2: Texture mapped model with semi-transparent skin.

3. AEROACOUSTIC MODELLING

Once the articulators shapes can be coherently controlled by a small set of articulatory parameters, the vocal tract area function can be determined. The final stage of the complete *Virtual Talking Head* that represents all peripheral speech production processes is the aero-acoustic stage. The corresponding aerodynamic and acoustic models, inspired by Scully's work [8], are described in detail in [9].

3.1 AEROACOUSTIC MODELLING

The aerodynamic phenomena that convert the lung pressure into acoustic excitation sources for the vocal tract are taken into account by a simplified low frequency airflow model, associated with a two-mass model of the vocal folds for the voice source and with a functional model of the frication noise source. The airflow model is essential for ensuring: (1) the proper influence of the voice source upon the noise source (the modulation of the noise source amplitude induced by the vocal tract flow variations generated by the voice source), and (2) the influence of vocal tract oral constriction upon voicing (the reduction of voicing amplitude due to the increase of pressure drop at the oral constriction and the concomitant decrease of pressure drop at the glottis).

The voice source is a two-mass model of the vocal folds controlled by the *trans-glottal pressure*, the *rest glottis height* (*i.e.* the distance between vocal folds), and the *vocal fold length*. This model delivers the flow at the glottis, the derivative of which is used as the source of voicing in the acoustic model. The resulting glottal area signal is low-pass filtered in order to retain only the slow variations at the fundamental frequency and is re-injected in the airflow model. Note that the coefficients of this model have been tuned in order to be able to reproduce as faithfully as possible the reference subject's voice [10].

The frication noise source is modelled by a functional model, which predicts noise spectral characteristics from cross-sectional area and pressure drop at the oral constriction. This model is controlled by the low frequency component of the pressure drop across the oral constriction and by the main constriction area. This ensures that the low frequency flow fluctuations at the glottis are transmitted to the constriction pressure drop and used to modulate synchronously the frication source in the case of voiced fricatives.

To summarise, the whole aeroacoustic model is controlled, in addition to the oral constriction area, by three laryngeal articulatory-like parameters: *subglottal pressure*, *vocal folds length*, and *glottis rest height*.

The resulting synthetic speech sound is finally produced by a time-domain reflection-type line analogue, which takes into account the acoustic interaction between the vocal tract acoustic model and the two-mass model.

3.2 VOICING / FRICATION BALANCE

Due to the behaviour of the voice and frication sources, and to the aerodynamic interactions in the vocal tract, it appears that for a given subglottal pressure, increasing simultaneously voicing and frication amplitudes is contradictory (*cf. e.g.* [11]): the voicing / frication noise balance depends thus on a critical coordination between glottis and oral constriction coordination.

Simulations performed with our speech production model can exemplify this phenomenon. Figure 4 displays the difference between the amplitude L_v of the voice component and the amplitude L_f of the frication noise component produced with a simplified fricative articulation for a constant subglottal pressure, as a function of glottis area A_g and constriction area A_c. More precisely, the figure displays, in grey levels, the quantity $-20\ |\log (L_V/L_f)\ |$ that represents the balance between voice and noise levels: when both levels are identical, this quantity is zero, and thus displayed in white, whereas when the difference increases it is more and more negative, coded in darker and darker grey levels (in black below –12 dB). The narrow region, which is not black, represents thus the region of the control space that produces a signal for which the difference between the voice and noise components does not exceed 12 dB. It appears clearly that a rather strict coordination between the glottis and the oral constriction is needed to produce acceptable voiced fricatives [11]. This requirement of precise coordination may explain the much higher proportion of voiceless fricatives over voiced ones in the languages of the world (*cf. e.g.* [12]).

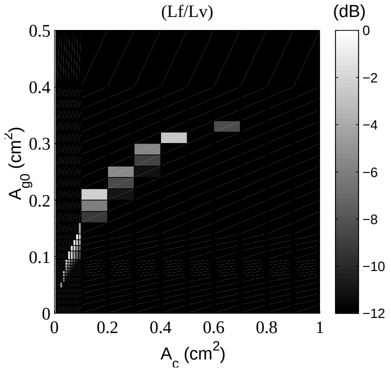

Figure 4: Difference between voicing and frication amplitudes (in grey levels) in the [A_g / A_c] control space.

4. CONTROL ISSUES

Integrating the models described above yields a video -realistic *Virtual Talking Head*, which represents the subject's audiovisual speech production capabilities. This talking head is controlled by two sets of articulatory parameters: supralaryngeal parameters (*i.e.* the command parameters of the articulatory model), and laryngeal parameters controlling the vocal folds (*glottal pressure*, *vocal folds length*, *glottis rest height*), that need to be carefully coordinated.

As mentioned earlier, an adequate balance between the contribution of voice and frication acoustic sources plays an important role in the quality and naturalness of the synthesised speech. We have shown that our complete aeroacoustic model allows us to use a simple strategy of control of the glottis gesture, in coordination with the oral constriction, to produce both voiced and voiceless fricatives [9]. The following strategy was used for voiced fricatives to determine the three parameters controlling acoustic excitation sources: the subglottal pressure was set at a constant value of 10 cm H_2O; the glottal rest height was set to about 0.03 cm; the vocal fold length was set to 1.6 cm. For voiceless fricatives, the glottal rest height was set to 0.1 cm at the instant of minimum sound power in the fricative segment, in order to ensure that voicing ceased during the consonant. A sigmoidal interpolation with the adjacent vowels ensured a smooth glottis opening / closing gesture during the fricative.

Different approaches can be envisaged to control the talking head in order to produce synthetic audiovisual speech. Mawass et al. [9] have shown that articulatory *inversion* from formants and lip aperture could be successfully employed to determine the articulatory control parameter trajectories from audiovisual recordings of the subject. Direct measurement methods such as electromagnetic articulometry coupled with tracking from video images [7] can be also employed, while dynamic MRI is becoming more and more a realistic alternative to traditional cineradiography.

A similar analysis-by-synthesis technique was also applied to recover facial movements from video recordings [7]. We have then a set of articulatory trajectories that can reproduce in a coherent way the visual and acoustic consequences of speech production for a training corpus.

Different techniques and control models can thus be further invoked and trained for generating these trajectories from phonetic input: (1) stylise them with a usual "target and transition functions" model, use a coarticulation model and follow the general scheme of rule-based synthesis (*cf. e.g.* [13], [14]); (2) store them integrally or in a compressed form into segments and use a standard concatenative synthesis scheme ([15], [16]); (3) use inversion techniques to drive the articulatory model from the acoustic output easily delivered by a more classical text-to-speech synthesizer ([17], [18]).

We tested these different approaches [19] for facial animation using a point-light technique [20]. While simple acoustic-to-articulatory inversion techniques were unable to provide adequate movements, a concatenation of diphone-like articulatory segments with a simple anticipatory smoothing strategy produces movements that are almost indistinguishable from the original movements.

In addition to the knowledge gained on speech production and on the degrees of freedom of the speech articulators, the present models open the way to audiovisual speech synthesis. A first text-to-speech audiovisual synthesis has been developed [21] and is available from the web[1]. A virtual tongue and a virtual hand have been recently added to evaluate the gain in intelligibility brought by augmented reality (seeing the tongue movements through a transparent skin) and manual cued speech (externalising hidden movements by hand gestures).

5. DISCUSSION AND PERSPECTIVES

The availability of the *Virtual Talking Head*, and of its associated articulatory and aeroacoustic models opened the way to a high quality articulatory synthesis of vowel-fricative-vowel sequences in French. The quality and realism of the Talking Head and of its control can still be greatly improved. More 3D data should help extend the model with a velum and appropriate nasal cavities, as well as better lips. Unexpectedly, our linear model was found to be able to account fairly well for the collision of the lateral sides of the tongue blade with the maxilla. For the tongue tip however, the question remains open whether non-linear modelling could reduce the need for an extra parameter that would deal more specifically with palatolingual contacts.

The present work is essentially based on one single subject. The problem of normalisation between speakers remains to be solved. Cloning more subjects, and finally developing meta-models flexible enough to be adapted to a range of different subjects is on the way.

The gains of intelligibility put forward by enhancing speech with seeing the movements that produced it is not a panacea. Intelligibility is the complex result of the intrinsic quality of the signals delivered to the listener and of the *a priori* knowledge the listener has and recruits when decoding. The level of recruitment of cognitive resources is of main interest if we aim at using such speech technologies in the every day-life as hearing aids as well as communicative virtual agents. A recent large-scale evaluation campaign has shown that the performance of model-based audiovisual synthesis [23] is still far from that of image-based synthesis that succeed in passing the Turing tests [24]. We should learn the lesson and put more and more emphasis on observing and modelling human beings in action.

ACKNOWLEGMENTS

This work has benefited, along years, from the participation of many people that we sincerely wish to acknowledge, among them: A. Arnal, M. Baciu, D. Beautemps, K. Mawass, X. Pelorson, C. Savariaux, & C. Segebarth. This work has been partly funded by the French National Network for Telecom Research (RNRT "TempoValse" and "Artus" projects).

REFERENCES

[1] C. H. Coker, "Synthesis by rule from articulatory parameters," presented at Proceedings of the 1967 Conference on Speech Communication Processes, pp. 52-53, 1967.

[2] P. E. Rubin, E. L. Saltzman, L. Goldstein, R. S. McGowan, M. K. Tiede, and C. P. Browman, "CASY and extensions to the task-dynamic model," presented at Proceedings of the 4th Speech Production Seminar - 1st ESCA Tutorial and Research Workshop on Speech Production Modeling: from Control Strategies to Acoustics, Autrans, France, pp. 125-128, 1996.

[3] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *Journal of the Acoustical Society of America*, vol. 98, pp. 1878-1889, 1995.

[4] T. Kaburagi and M. Honda, "Dynamic articulatory model based on multidimensional invariant-feature task representation," *Journal of the Acoustical Society of America*, vol. 110, pp. 441-452, 2001.

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

[6] L. Revéret and C. Benoît, "A new 3D lip model for analysis and synthesis of lip motion in speech production," presented at Proceedings of the International Conference on Auditory-Visual Speech

[1] See the audiovisual text-to-speech system using a the rule-based control model with Ohman's coarticulation model and the MEM anticipatory model [22]: `www.icp.inpg.fr/ICP/equipes/synthese/avtts/clone.fr.html`

Processing / Second ESCA ETRW on Auditory-Visual Speech, Terrigal-Sydney, Australia, pp. 207-212, 1998.

[7] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," presented at Proceedings of the Auditory-Visual Speech Processing Workshop, AVSP 2001, Scheelsminde, Denmark, pp. 90-97, 2001.

[8] C. Scully, "Speech production simulated with a functional model of the larynx and the vocal tract," *Journal of Phonetics*, vol. 14, pp. 407-413, 1986.

[9] K. Mawass, P. Badin, and G. Bailly, "Synthesis of French fricatives by audio-video to articulatory inversion," *Acta Acustica*, vol. 86, pp. 136-146, 2000.

[10] C. Vescovi, E. Castelli, and X. Pelorson, "Adaptation of a two-mass model of the vocal cords to a particular speaker," presented at Proceedings of the 4th EuroSpeech Conference, Madrid, Spain, vol. 3, pp. 1933-1936, 1995.

[11] C. Abry, P. Badin, K. Mawass, and X. Pelorson, "The Equilibrium Point Hypothesis and control space for relaxation movements or "When is movement actually needed to control movement ?", Commentary on target paper: P. Perrier, D.J. Ostry & R. Laboissière (1996), The Equilibrium Point Hypothesis and its application to speech motor control (*JSHR*, 39, 365-378)," *Les Cahiers de l'ICP, Bulletin de la Communication Parlée*, vol. 4, pp. 27-33, 1998.

[12] N. Vallée, L.-J. Boë, J.-L. Schwartz, P. Badin, and C. Abry, "The weight of phonetic substance in the structure of sound inventories," *ZAS Papers in Linguistics*, vol. 28, pp. 145-168, 2002.

[13] 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.

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

[15] S. Minnis and A. Breen, "Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis," presented at Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China, vol. II, pp. 759-762, 2000.

[16] T. Okadome, T. Kaburagi, and M. Honda, "Articulatory movement formation by kinematic triphone model," presented at IEEE International Conference on Systems Man and Cybernetics, Tokyo, Japan, vol. 2, pp. 469-474, 1999.

[17] H. Yehia, P. E. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23-43, 1998.

[18] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," presented at Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, Kloster Seeon, Germany, pp. 265-268, 2000.

[19] 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.

[20] L. D. Rosenblum, J. A. Johnson, and H. M. Saldaña, "Point-light facial displays enhance comprehension of speech in noise," *Journal of Speech and Hearing Research*, vol. 39, pp. 1159-1170, 1996.

[21] 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 Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China, vol. II, pp. 755-758, 2000.

[22] C. Abry and T. M. Lallouache, "Modeling lip constriction anticipatory behaviour for rounding in French with the MEM (Movement Expansion Model)," presented at Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden, vol. 4, pp. 152-155, 1995.

[23] I. Pandzic, J. Ostermann, and D. Millen, "Users evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, pp. 330-340, 1999.

[24] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Transactions on Graphics*, vol. 21, pp. 388-398, 2002.