LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model

Piero Cosi, Andrea Fusaro and Graziano Tisato

Istituto di Scienze e Tecnologie della Cognizione - C.N.R.
Sezione di Padova "Fonetica e Dialettologia"
Via G. Anghinoni, 10 - 35121 Padova ITALY
[cosi,fusaro,tisato]@csrf.pd.cnr.it

Abstract

LUCIA, a new Italian talking head based on a modified version of the Cohen-Massaro's labial coarticulation model is described. A semi-automatic minimization technique, working on real cinematic data, acquired by the ELITE opto-electronic system, was used to train the dynamic characteristics of the model. LUCIA is an MPEG-4 standard facial animation system working on standard FAP visual parameters and speaking with the Italian version of FESTIVAL TTS.

1. Introduction

There are many ways to control a synthetic talking face. Among them, geometric parameterization [1-2], morphing between target speech shapes [3], muscle and pseudo-muscle models [4-5], appear the most attractive.
Recently, growing interest have encountered text to audiovisual systems [6-7], in which acoustical signal is generated by a Text to Speech engine and the phoneme information extracted from input text is used to define the articulatory movements.
For generating realistic facial animation is necessary to reproduce the contextual variability due to the reciprocal influence of articulatory movements for the production of following phonemes. This phenomenon, defined coarticulation [8], is extremely complex and difficult to model. A variety of coarticulation strategies are possible and different strategies may be needed for different languages [9].
A modified version of the Cohen-Massaro coarticulation model [10] has been adopted for LUCIA and a semi-automatic minimization technique, working on real cinematic data acquired by the ELITE opto-electronic system [11], was used for training the dynamic characteristics of the model, in order to be more accurate in reproducing the true human lip movements[1].

2. Coarticulation Model

The coarticulation model proposed by Cohen and Massaro [12] implements Löfqvist's gestural theory of speech production [13], in its turn profoundly inspired by Browman and Goldstein work on articulatory phonology [14]. Each phoneme is specified in terms of speech control parameters (e.g. lip rounding, upper and lower lip displacement, lip protrusion) characterized by a target value and a dominance function.
Dominance functions of consecutive phonemes overlap in time and specify the degree of influence that a speech segment has over articulators in the production of preceding or following segments.
The final articulatory trajectory of a specific parameter is the weighted average of the sum of all dominances scaled by the magnitude of the associated targets. For a sequence of N phonemes, if T_i is the i'th target amplitude, t_i its time location and $D_i(t)$ its associated dominance, the final parameter function is given by:

$$F(t) = \frac{\sum_{i=1}^{N} T_i \cdot D_i(t - t_i)}{\sum_{i=1}^{N} D_i(t - t_i)} \tag{1}$$

where the dominance, having the form of a negative exponential function, is given by:

$$D(\tau) = \begin{cases} \alpha e^{-\theta_{\mathrm{bw}}|\tau|^c} & \text{if } \tau \le 0 \\ \alpha e^{-\theta_{\mathrm{fw}}|\tau|^c} & \text{if } \tau > 0 \end{cases} \tag{2}$$

where α indicates the magnitude of the dominance, θ_{bw} and θ_{fw} represent the rate of its backward (bw) and forward (fw) temporal extent and the power c influences its degree of activation (rise and fall off). The influence of a segment first increases then decreases, having maximal influence at the temporal location of the articulation target.

3. Modified Model

The method implemented by Cohen and Massaro can be improved in order to achieve an accurate description of the transitions between succeeding articulatory targets at various speech rates, and solve several difficulties in the production of bilabial and labiodental consonants. This objective is reached by adopting a new general version of the dominance functions and by adding temporal resistance and shape components to the original model. In fact, in the original model, the parameter c is set to a constant unit value but, in a

[1] Part of this work has been sponsored by MPIRO (Multilingual Personalized Information Objects, European Project IST-1999-10982, http://www.ltg.ed.ac.uk/mpiro/), TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali, joint "CNR - Provincia Autonoma Trentina" Project), and PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST-2001-37599, http://pfstar.itc.it/).

general context, it can be different for each phoneme and can also change for the backward and forward case (c_{bw} , c_{fw}). From this point it can be interpreted as the rate of activation and release of the phonemic articulatory gesture respectively. Variations of c_{bw} and c_{fw} generate different qualitative behavior of the dominances and consequently of the resulting control parameter that becomes particularly evident at growing speech rate [10, fig. 2].

As reported in [12], the use of a variable degree of dominance shares "the idea of a numerical coefficient for coarticulation resistance associated to some phonetic features in the theory of Bladon and Al-Bamerny" [9]. This is strictly related to different values of the dominance amplitude and reflects on how close the lips come to reach their target value. At a high speech rate dominances are close to each other and even if their amplitude value is high the final trajectory is far from target locations. This constitutes a problem if the target should be reached such as in the production of bilabial stops (/p, b, m/) and labiodental fricatives (/f, v/). To solve this problem the concept of coarticulation resistance has been applied to the temporal extent of dominances. A negative exponential $R(\tau)$, called temporal resistance function [10, formula (4)], has been associated to each dominance, whose main feature is that its backward and forward extent can change according to the *resistance coefficient* k_R of preceding and following phonemes. In other words, if the resistance of the following phoneme is maximum (i.e. $k_R = 1$), the forward temporal extent of the resistance function is equal to the temporal distance between the current phoneme target and the following one. In this way the combined function $D(\tau) \cdot R(\tau)$ falls to zero at the instant of the maximum of the dominance of the following phoneme and the corresponding articulatory target can be reached. For $k_R < 1$, the extent of $R(\tau)$ grows following a recursive procedure defined in [10].

A shape function [10, formula (5)] was also introduced in order to model the trajectory behavior in the proximity of the articulatory targets. This function is useful when we want to describe distinctive features like for example a slope next to the target or a transition characterized by an initial strong fall-off followed by a final low one.

In conclusion, the original parameter function (1) has been modified in order to include the new temporal resistance $R(\cdot)$ and shape $S(\cdot)$ functions previously defined, thus obtaining:

$$F_{new}(t) = \frac{\sum_{i=1}^{N} T_i \cdot S_i(t-t_i) \cdot R_i(t-t_i) \cdot D_i(t-t_i)}{\sum_{i=1}^{N} R_i(t-t_i) \cdot D_i(t-t_i)} \quad (3)$$

where the temporal resistance function was included in the denominator in accordance with its strict relation with the dominance.

4. Data Analysis

The values of the coefficients of the new model have been determined starting from a database of real labial movements of an Italian speaker pronouncing VCV symmetrical stimuli, where V is one of the vowels /a/, /i/ or /u/, and C is one of the Italian consonant phonemes. The database represents spatio-temporal trajectories of six parameters (upper lip opening, lower lip opening, upper lip protrusion, lower lip protrusion, lip rounding and jaw opening) recorded by the ELITE optoelectronic system [11].

The parameter estimation procedure is based on a least squared minimization of the error:

$$e(t) = \sum_{i=1}^{N} \left(Y(n) - F(n) \right)^2 \quad (4)$$

between real data Y(n) and modeled curves F(n) for 5 repetitions of the same sequence type.

An automatic optimization algorithm with a strong convergence property has been used [15]. Even if the number of parameters to be optimized is rather high, the size of the data corpus is large enough to allow a meaningful estimation, but, due to the presence of several local minima, the optimization process has to be manually controlled in order to assist the algorithm convergence. As illustrated in Figure 1, real and simulated curves, referring to the lip-opening parameter in three aCa examples, look quite similar.

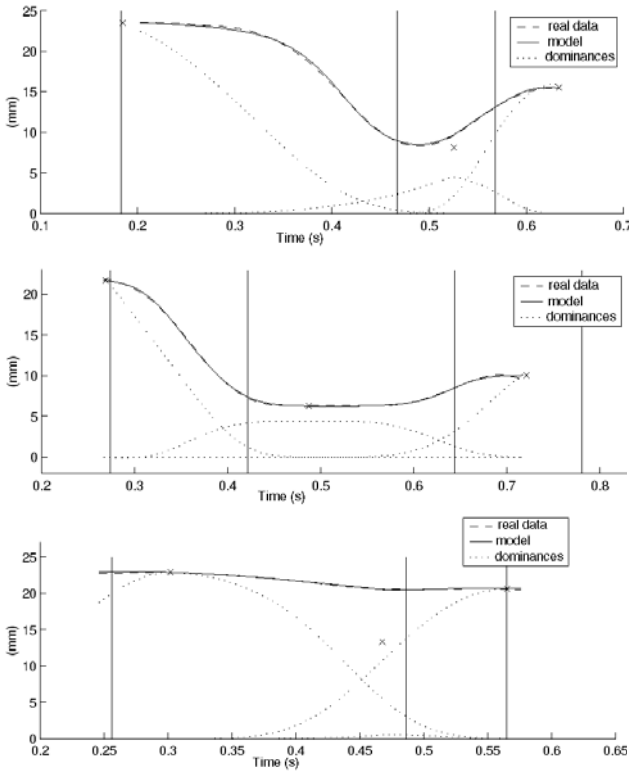

Figure 1. Example of the modeled trajectories of the lower lip opening parameter in the production of /'a d a/ (upper plot), /'a dz a/ (middle plot) and /'a l a/ lower plot) sequences. Dotted lines represent the dominance functions scaled by the target amplitudes.

The mean total error between real and simulated trajectories for the whole set of parameters is lower than 0.3 mm in the

case of bilabial and labiodental consonants in the /a/ and /i/ contexts [16, p. 63].

5. Lucia

The modified model has been applied to LUCIA, an Italian talking head based on the MPEG-4 standard [17], speaking with the Italian version of FESTIVAL TTS [18], as illustrated in the block diagram shown in Figure 2.
LUCIA is a graphic MPEG-4 compatible facial animation engine implementing a decoder compatible with the "Predictable Facial Animation Object Profile" [17].
MPEG4 specifies a set of Face Animation Parameters (FAPs), each corresponding to a particular facial action deforming a face model in its neutral state. A particular facial action sequence is generated by deforming the face model, in its neutral state, according to the specified FAP values, indicating the magnitude of the corresponding action, for the corresponding time instant. Then the model is rendered onto the screen.

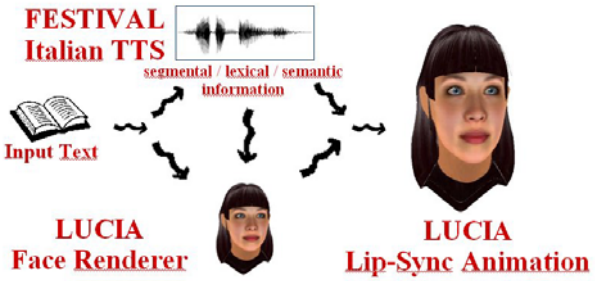

Figure 2: Lucia, a new Italian Talking head

LUCIA is able to generate a 3D mesh polygonal model by directly importing its structure from a VRML file [19] and to build its animation in real time.
At the current stage of development, as illustrated in Figure 3, LUCIA is a textured young female 3D face model built with 25423 polygons: 14116 belong to the skin, 4616 to the hair, 2688x2 to the eyes, 236 to the tongue and 1029 to the teeth respectively.
Currently the model is divided in two sub sets of fundamental polygons: the skin on one hand and the inner articulators, such as the tongue and the teeth, or the facial elements such as the eyes and the hair, on the other. This subdivision is quite useful when animation is running, because only the reticule of polygons corresponding to the skin is directly driven by the pseudo-muscles and it constitutes a continuous and unitary element, while the other anatomical components move themselves independently and in a rigid way, following translations and rotations (for example the eyes rotate around their center). According to this strategy the polygons are distributed in such a way that the resulting visual effect is quite smooth with no rigid "jumps" over all the 3D model.
LUCIA emulates the functionalities of the mimic muscles, by the use of specific "displacement functions" and of their following action on the skin of the face. The activation of such functions is determined by specific parameters that encode small muscular actions acting on the face, and these actions can be modified in time in order to generate the wished animation. Such parameters, in MPEG-4, take the name of Facial Animation Parameters and their role is fundamental for achieving a natural movement. The muscular action is made explicit by means of the deformation of a polygonal reticule built around some particular key points called "Facial Definition Parameters" (FDP) that correspond to the junction on the skin of the mimic muscles.
Moving only the FDPs is not sufficient to smoothly move the whole 3D model, thus, each "feature point" is related to a particular "influence zone" constituted by an ellipses that represents a zone of the reticule where the movement of the vertexes is strictly connected. Finally, after having established the relationship for the whole set of FDPs and the whole set of vertexes, all the points of the 3D model can be simultaneously moved with a graded strength following a raised-cosine function rule associated to each FDP.

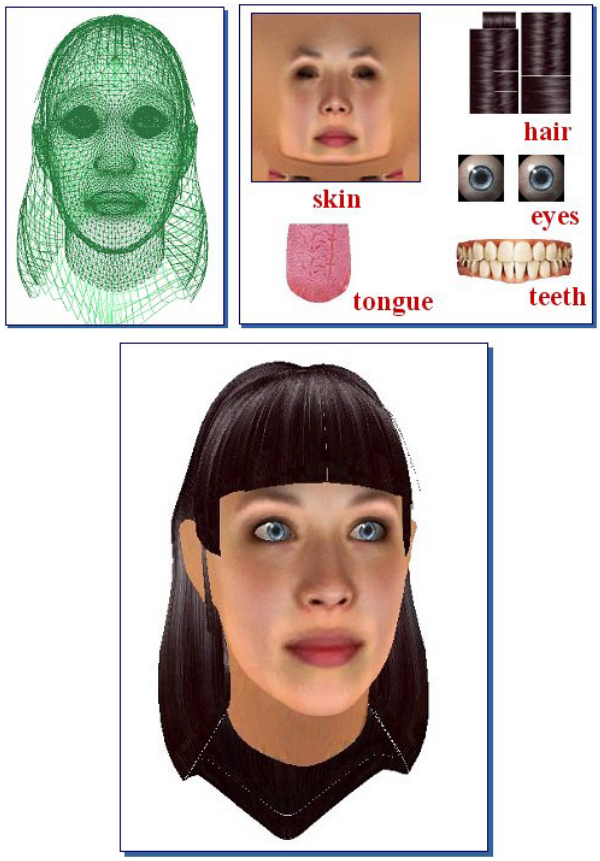

Figure 3: Lucia's wireframe and textures.

A sequence of snapshots of Lucia while producing the Italian sentence 'la gamba della mamma' ("the leg of the mam") /l a g a_1 m b a d e l: a m a_1 m: a/ is illustrated in Figure 4.

6. Concluding Remarks

The graphic engine of LUCIA is similar to others MPEG based projects that were previously realized, but the novelty is the high quality of the 3D model, and the very fine coarticulation model, which is automatically trained by real data, used to animate the face.
The modified coarticulatory model is able to reproduce quite precisely the true cinematic movements of the articulatory parameters. The mean error between real and simulated trajectories for the whole set of parameters is, in fact, lower than 0.3 mm.

Labial movements implemented with the new modified model are quite natural and convincing especially in the production of bilabials and labiodentals and remain coherent and robust to speech rate variations.

The overall quality and user acceptability of LUCIA talking head has to be perceptually evaluated [20] by a complete set of test experiments, and the new model has to be trained and validated in asymmetric contexts (V_1CV_2) too. Moreover, emotions and the behavior of other articulators, such as tongue for example, have to be analyzed and modeled for a better realistic implementation.

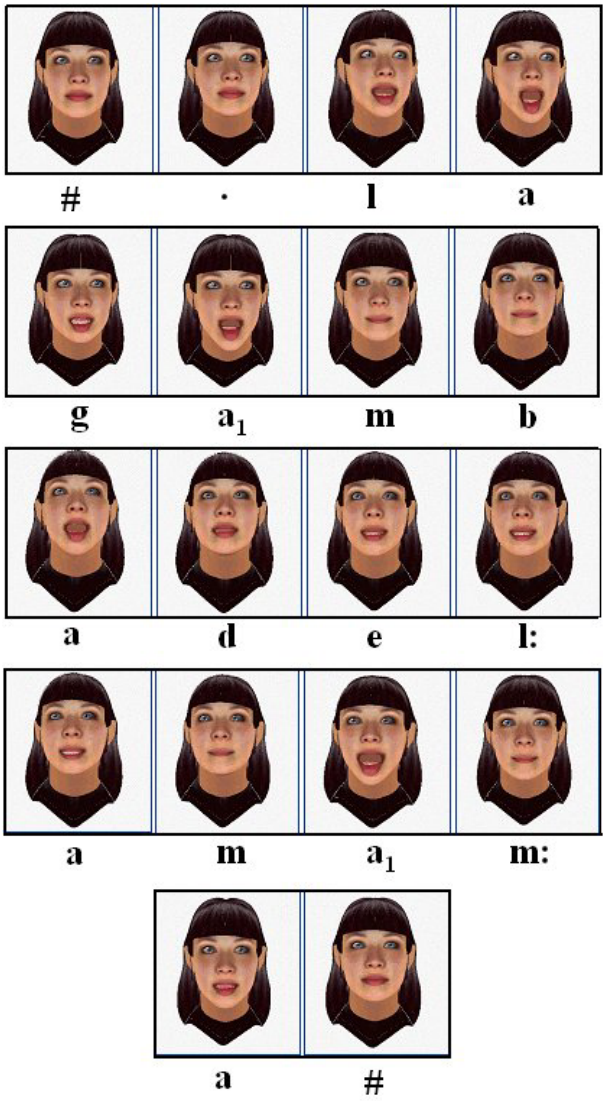

Figure 4: Snapshots of one sample animation of LUCIA pronouncing the Italian utterance 'la gamba della mamma' ("the leg of the mam") /l a g a_1 m b a d e l: a m a_1 m: a/.

7. References

[1] Massaro D.W., Cohen M.M., Beskow J., Cole R.A., "Developing and Evaluating Conversational Agents", in Cassell J., Sullivan J., Prevost S., Churchill E. (Editors), *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000, pp. 287-318.

[2] Le Goff, B. *Synthèse à partir du texte de visages 3D parlant français*, PhD thesis, Grenoble, France, October 1997.

[3] Bregler C., Covell M., Slaney M., "Video Rewrite: Driving Visual Speech with Audio", in *Proc. of SIGGRAPH '97*, 1997, pp. 353-360.

[4] Lee Y., Terzopoulos D., Waters K., "Realistic Face Modeling for Animation", in *Proc. of SIGGRAPH '95*, 1995, pp. 55–62.

[5] Vatikiotis-Bateson E., Munhall K.G., Hirayama M., Kasahara Y., Yehia H., "Physiology-Based Synthesis of Audiovisual Speech", in *Proc. of 4th Speech Production Seminar: Models and Data*, 1996, pp. 241-244.

[6] Beskow J., "Rule-Based Visual Speech Synthesis," in *Proc. of Eurospeech '95*, Madrid, 1995, pp.299–302.

[7] LeGoff B. and Benoit C., "A text-to-audiovisualspeech synthesizer for French", in *Proc. of the ICSLP '96*, Philadelphia, USA, pp. 2163-2166.

[8] Farnetani E., Recasens D., "Coarticulation Models in Recent Speech Production Theories", in Hardcastle W.J. (Editors), *Coarticulation in Speech Production*, Cambridge University Press, Cambridge, 1999.

[9] Bladon, R.A., Al-Bamerni, A., "Coarticulation resistance in English \l\", *Journal of Phonetics*, 4, 1976, pp. 135-150.

[10] Cosi P., Perin G., "Labial Coarticulation Modeling for Realistic Facial Animation", in *Proc. of ICMI '02*, Pittsburgh, PA, USA, 2002, pp. 505-510.

[11] Ferrigno G., Pedotti A., "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", in *IEEE Transactions on Biomedical Engineering*, BME-32, 1985, pp. 943-950.

[12] Cohen M., Massaro D., "Modeling Coarticulation in Synthetic Visual Speech", in Magnenat-Thalmann N., Thalmann D. (Editors), *Models and Techniques in Computer Animation*, Springer Verlag, Tokyo, 1993, pp. 139-156.

[13] Löfqvist, A. "Speech as Audible Gestures", in Hardcastle W.J., Marchal A. (Editors.), *Speech Production and Speech Modeling*, Dordrecht: Kluwer Academic Publishers, 1990, pp. 289-322.

[14] Browman, C. P., Goldstein, L., "Towards an articulatory phonology", *Phonology Yearbook*, 1986, 3, pp. 219-252.

[15] Schultz R., Schnabel B., Byrd M., "A Family of Trust-Region-Based Algorithms for Unconstrained Minimization with Strong Global Convergence Properties", *SIAM Journal on Numerical Analysis*, 22, 1985, pp. 47-67.

[16] Perin G., *Facce parlanti: sviluppo di un modello coarticolatorio labiale per un sistema di sintesi bimodale*, Master Thesis, Univ. of Padova, Italy, 2000-1.

[17] Mpeg-4 standard. Home page: http://mpeg.telecomitalialab.com/standards/mpeg4

[18] Cosi P., Tesser F., Gretter R., Avesani C., "Festival Speaks Italian!", in *Proc. of Eurospeech 2001*, Aalborg, Denmark, September 3-7 2001, pp. 509-512.

[19] Hartman J., Wernecke J., *The VRML Handbook*, Addison Wessley, 1996.

[20] Massaro D.W., *Perceiving Talking Faces: from Speech Perception to a Behavioral Principle*, Cambridge, MA, MIT Press, 1997.