An approach to an Italian Talking Head

C. Pelachaud

Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza", Rome, Italy
cath@dis.uniroma1.it

E. Magno-Caldognetto, C. Zmarich, P. Cosi

Istituto di Fonetica e Dialettologia
C.N.R. of Padova Padova, Italy
magno/zmarich/cosi@csrf.pd.cnr.it

Abstract

Our goal is to create a natural talking face with, in particular, lip-readable movements. Based on real data extracted from an Italian speaker with the ELITE system, we have approximated the data using radial basis functions. In this paper we present our 3D facial model based on MPEG-4 standard and our computational model of lip movements for Italian. Our experiment is based on some phonetic-phonological considerations on the parameters defining labial orifice, and on identification tests of visual articulatory movements.

1. Introduction

As computers are being more and more part of our world we feel the urgent need of proper user interface to interact with. The metaphor of face-to-face communication applied to human-computer interaction is receiving a lot of attention [1]. Humans are used since they are born to communicate with others. Seeing faces, interpreting their expression, understanding speech are all part of our development and growth. But face-to-face conversation is very complex phenomenon as it involved a huge number of factors. We speak with our voice, but also with our hand, eye, face and body. In this paper, we present our work on natural talking face. Our purpose is to build a 3D facial model that would have lip-readable movements, that is a face whose lips would be detailed enough to allow one to read from her lips. We first present our 3D facial model. Then we concentrate on the computation of lip movements.

2. Facial Model

Our facial model is based on MPEG-4 standard [2, 3]. The model uses a pseudo-muscular approach [4]. The muscle contractions are obtained through the deformation of the polygonal network around feature points. Each feature point corresponds to skin muscle attachment. The deformation is performed in a zone of influence that has an ellipsoid shape whose centroid is the feature point. The displacement of points within this area of influence obeys to a deformation function that is function of the distance between the points and the feature point (see figures 1 and 2). Tow sets of parameters describe and animate the 3D facial model: facial animation parameter set (FAPS) and facial definition parameter (FDP). The FDPs define the shape of the model while FAPS define the facial actions. When the model has been characterized with FDP, the animation is obtained by specifying for each frame the values of FAPS.

The facial model also includes particular features such as wrinkles and furrows to enhance its realism. In particular, bulges and furrows have been modeled using a specialized displacement function that move outward points within a specific

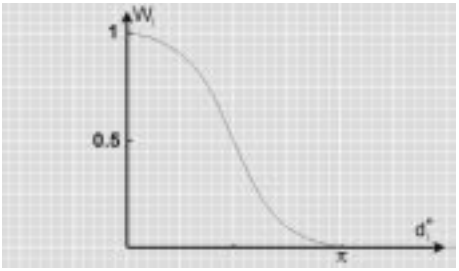

Figure 1: Deformation function

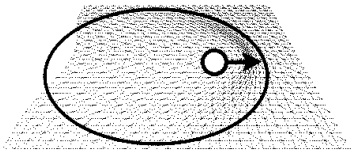

Figure 2: Skin deformation in the area of influence

area. The points of area A that are affected by muscular contraction will be deformed by the muscular displacement function, while the points of area B (area of the bulge / furrow) will be moved outward to simulate the skin accumulation and bulging (see figures 3 and 4).

3. Lip Movements

At the Istituto di Fonetica e Dialettologia-C.N.R. of Padova, the spatiotemporal characteristics of the 3D articulatory movements of the upper lip (UL), lower lip (LL) and jaw (J), together with

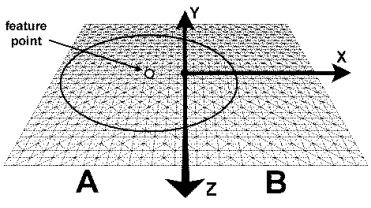

Figure 3: Within area of influence, the two zones A (muscular traction) and B (accumulation)

Figure 4: Simulation of the nasolabial furrow

the co produced acoustic signal, were recorded by means of ELITE, an optoelectronic system that applies passive markers on the speaker face [5, 6, 7]. The articulatory characteristics of Italian vowel and consonant targets in the /'VCV/ context were quantified from at least 4 subjects, repeating 5 times each item. These researches defined:

- the area of the labial orifice, by means of the following parameters, phonologically relevant: lip height (LH), lip width (LW), upper lip protrusion (UP) and lower lip protrusion (LP) [7]. Fig. 5 represents the three-dimensional coordinates (LH, LW, LP) for the 21 Italian consonants in the /'aCa/ context and averaged along all the subjects' productions, normalized by subtracting the values related to the position of the lips at rest. The parameter which best distinguishes, on statistical ground, the consonants from each other is LH [7]. From the Figure 5 it is evident that for LH, three consonants, /p, b, m/, present negative values determined by the compression of the lips performing the bilabial closure and that the minimum positive values were recorded for /f, v/. It is important to bear in mind that lip movements in Italian are phonologically critical in implementing distinctive characteristics of manner and place of articulation only for bilabial stops (/p, b, m/) and labiodental fricatives (/f, v/), whereas for the other consonants, for which the tongue is the primary articulator, lip movements are phonologically under-specified and determined mainly by the co-ordination with jaw closing movement and the coarticulation with contextual vowels.
- The displacement and duration of movement of the LH parameter for all the consonants.
- The relationship between the articulatory movements and the corresponding acoustic production. The analyses indicate that, for LH parameter and in almost all the consonants, the percentage value, representing the time interval between the acoustic onset of the consonant and the articulatory target range from 20% to 45% of the total acoustic duration of the consonant.

For the moment we have decided to concentrate on 4 parameters: LH, LW, UP and LP. These parameters have been found to be independent, as well as to be phonetically and phonologically relevant,. Our first step is to approximate the displacement curves representing by the 4 articulatory parameters over time. Our approach is to approximate each curve by a mathematically-described function. The original curves are read and stored in an array called $Curve_i(t)$. Using a neural network model, we have written the curve as the weighted sum

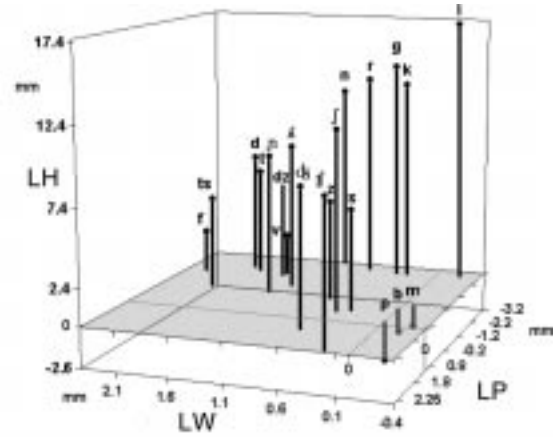

Figure 5: Spatial configurations of the labial orifice for the 21 Italian consonants in the context /'aCa/ based on LH, LW, and UP values (mm)

of radial basis functions $f_j(t)$ of the form:

$$f_j(t) = \sum_i \lambda_i e^{-\frac{|t - time(t_i)|^2}{\sigma_i^2}}$$

Where λ_i and σ_i are the parameters that define each radial basis function. Each curve has 3 peak values (maxima or minima points) corresponding to the production of V, C and V. For each of these targets within each curve, we look for the time of these peaks (see Figures 6, 7, 9, 10). We gather these temporal values in an array called 'time'. We can notice that we may encounter asynchronies of the labial target over the acoustic signal, according to the parameter and/or the phoneme. Further, the different ranges of extension for different parameters have to be stressed: for example, the UL and LL variations under 1 mm (cf. Figures 9, 10) are evidently not so much relevant. We want the curve to be particularly well approximated around these peak points. To ensure a better approximation, we consider 2 more points surrounding the peak: one point at time (time(t) - 1) and one point at time (time(t) + 1). The approximation method tries to minimize the equation:

$$min(f_i(t) - Curve_i(t))$$

that is we have to find the λ_i and σ_i that best verify this equation. We want to characterize the curves for the first V, the C and then the last V. For each 'VCV sequence we have 5 curves that corresponds to the 5 pronunciations by the same speaker of 'VCV. For example when we want to characterize the curves for C, we define a single pair (λ_c, σ_c) for each of the curves, that is this pair of parameters is common to each 5 curves, while the Vs will be characterized by distinct pairs of parameters. So we want to find the two parameters λ_c and σ_c that will best approximate all 5 curves around C. The same process is done to approximate the first V and the last V. We use unconstrained nonlinear optimization method as minimizing method using matlab. This approach uses a quasi-Newton algorithm and requires the gradients in λ and σ:

$$\frac{f_j(t)}{d\lambda_i} = e^{-\frac{|t - time(t_i)|^2}{\sigma_i^2}}$$

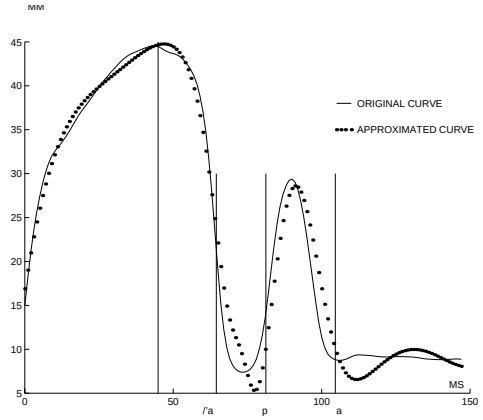

Figure 6: Lip height approximation of the sequence /'apa/; vertical lines defined the acoustic segmentation

Figure 8: Lip shape of /'a/ in /'apa/

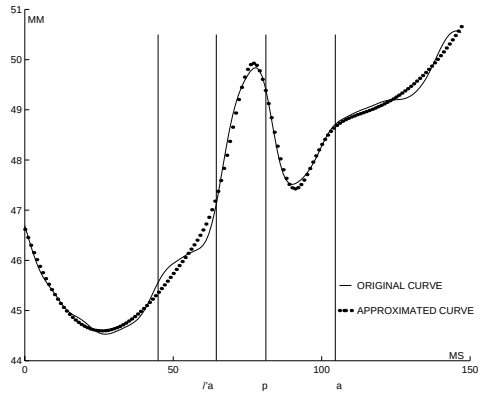

Figure 7: Lip width approximation of the sequence /'apa/

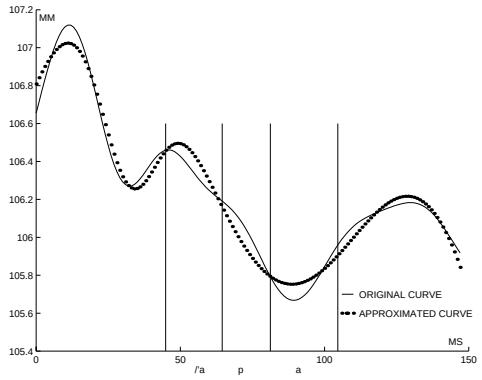

Figure 9: Upper lip protrusion approximation of the sequence /'apa/

$$\frac{f_j(t)}{d\sigma_i} = \lambda_i e^{-\frac{|t-time(t_i)|^2}{\sigma_i^2}} * 2\frac{|t-time(t_i)|^2}{\sigma_i^3}$$

Results of the approximation of the original curves for several lip parameters are shown in the figures 6, 7, 9, 10.

Having found the parameters that best described the curves 'VCV for V, C, and V, we are able to proceed to the first step toward animating the 3D facial model. The original curves are sampled every 1/10 of a second. For animating a 3D face we need a frame every 1/25 sec at a minimum. Having a mathematical representation of 'VCV curve for each 4 articulatory parameters, it is easy to get a value each 1/25 sec for these 4 parameters (lip height, lip with, upper and lower lip protrusion). Finally we need to convert these 4 parameters in parameters that drive the facial model, i.e. in FAPS (see as example Figure 8).

For the moment we chose sequences of the type /'aCa/ where C is one of the consonants /p, f, t, s, l, λ, $\int$/, i.e. the most preferred consonants in the identification tests of the visible articulatory movements [7, 8]. In fact, it is well known that the distinction, within homorganic consonants (as for instance /p, b, m/), between voiced and unvoiced consonants and between oral and nasal consonants, is not visually detectable, because vocal folds and velum movements are not visible. Assessment of the confusion errors so generated enables not only the identification of homophenous consonant groups (i.e. visemes, whose visible articulatory movements are considered as being similar and therefore transmit the same phonological information), but also the consonants acting as prototypes (for Italian, [7, 8].

4. Literature

The first facial model created by Parke [9] has been extended to consider other parameters specific to lip shape during speech (such as lip rounding and lip closure) [10, 11, 12]. 3D lip and jaw models have also been proposed [11] that are controlled by few labial parameters. EMG measurements of muscle contraction has been given as input to drive a physically-based facial model [13].

Video rewrite [14] uses real video footage of a speaker. Computer vision techniques are applied to tract points on the speaker's lips while morphing techniques are used to combine new sequences of mouth shapes. Voice Puppetry [15] does also use computer vision techniques but to learn a facial control model. The animation of the facial model is then driven by the audio.

The model of coarticulation used by Pelachaud et al. [16] implements the look-ahead model. On the other hand the approach proposed by Cohen and Massaro [10] implements

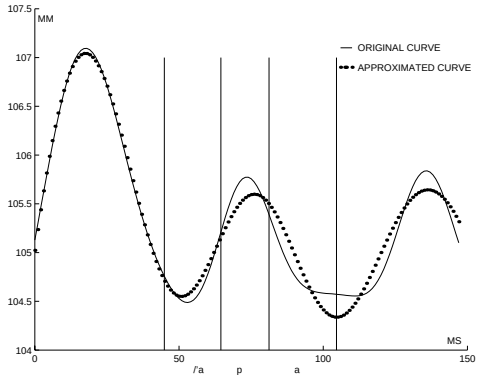

Figure 10: Lower lip protrusion approximation of the sequence /'apa/

Löfqvist's gestural theory of speech production [17]. The system uses overlapping dominance functions to specify how close the lips come to reaching their target value for each viseme. LeGoff [18] extended the formula developed by Cohen and Massaro to get a n-continuous function.

5. Future Developments

In the future we are going to process data from UL and LL movements separately. In fact, lips can displace in opposite directions and with different amplitude as for /p, b, m/: in this case lips change their shape because of compression. For the labiodental /f, v/ only LL behaves like an active articulator, while UL movement is due to a coarticulatory effect. Finally, for all the consonant targets, particular attention will be given to changes of LW (related to rounded/unrounded feature), and LP or UP (related to protruded/retracted feature), due to vocalic contexts. Synthesized Italian speech, produced by Festival [19], will be synchronized with articulatory movements. We are also planning to pursue perceptual study to evaluate the intelligibility of our lip model.

6. References

[1] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds., *Embodied Conversational Characters*, MIT Press, Cambridge, MA, 2000.

[2] P. Doenges, F. Lavagetto, J. Ostermann, I.S. Pandzic, and E. Petajan, "MPEG-4: Audio/video and synthetic graphics/audio for mixed media," *Image Communications Journal*, vol. 5, no. 4, May 1997.

[3] J. Ostermann, "Animation of synthetic faces in MPEG-4," in *Computer Animation'98*, Philadelphia, USA, June 1998, pp. 49–51.

[4] S. Pasquariello, "Modello per l'animazione facciale in MPEG-4," M.S. thesis, University of Rome, 2000.

[5] E. Magno-Caldognetto, K. Vagges, and C. Zmarich, "Visible articulatory characteristics of the Italian stressed and unstressed vowels," in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, 1995, vol. 1, pp. 366–369.

[6] E. Magno-Caldognetto, C. Zmarich, P. Cosi, and F. Ferrero, "Italian consonantal visemes: Relationships between spatial /temporal articulatory characteristics and coproduced acoustic signal," in *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, C. Benoit and R. Campbell, Eds., Rhodes, Greece, September 1997, pp. 5–8.

[7] E. Magno-Caldognetto, C. Zmarich, and P. Cosi, "Statistical definition of visual information for Italian vowels and consonants," in *International Conference on Auditory-Visual Speech Processing AVSP'98*, D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, Eds., Terrigal, Australia, 1998, pp. 135–140.

[8] E. Magno-Caldognetto and C. Zmarich, "L'intelligibilità dei movimenti articolatori visibili: caratteristiche degli stimoli vs. bias linguistici," in *Atti delle XI Giornate di Studio del G.F.S.*, P. Cosi and E. Magno-Caldognetto, Eds. UNIPRESS, Padova, Italy, in press.

[9] F.I. Parke, "Computer generated animation of faces," M.S. thesis, University of Utah, Salt Lake City, UT, June 1972, UTEC-CSc-72-120.

[10] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, M. Magnenat-Thalmann and D. Thalmann, Eds., Tokyo, 1993, pp. 139–156, Springer-Verlag.

[11] T. Guiard-Marigny, A. Adjoudani, and C. Benoit, "3D models of the lips and jaw for visual speech synthesis," in *Progress in Speech Synthesis*, J.P.H. van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, Eds. Springer-Verlag, 1996.

[12] J. Beskow, "Rule-based visual speech synthesis," in *ESCA - EUROSPEECH '95. 4th European Conference on Speech Communication and Technology*, Madrid, September 1995.

[13] E. Vatikiotis-Bateson, K.G. Munhall, M. Hirayama, Y.V. Lee, and D. Terzopoulos, "The dynamics of audiovisual behavior of speech," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, D.G. Stork and M.E. Hennecke, Eds., vol. 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*, pp. 221–232. Springer-Verlag, Berlin, 1996.

[14] C. Bregler, M. Covell, and M. Stanley, "Video rewrite: Driving visual speech with audio," in *Computer Graphics Proceedings, Annual Conference Series*. 1997, ACM SIGGRAPH.

[15] M. Brand, "Voice puppetry," in *Computer Graphics Proceedings, Annual Conference Series*. 1999, pp. 21–28, ACM SIGGRAPH.

[16] C. Pelachaud, N.I. Badler, and M. Steedman, "Generating facial expressions for speech," *Cognitive Science*, vol. 20, no. 1, pp. 1–46, January-March 1996.

[17] A. Lofqvist, "Speech as audible gestures," *Speech Production and Speech Modeling*, pp. 289–322, 1990.

[18] B. LeGoff and C. Benoit, "A French speaking synthetic head," in *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, C. Benoit and R. Campbell, Eds., Rhodes, Greece, September 1997.

[19] A.W. Black, P. Taylor, R. Caley, and R. Clark, "Festival," http://www.cstr.ed.ac.uk/projects/festival/.