MODELLING 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 Italian 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. LITERATURE

The first facial model created by Parke [2] has been extended to consider other parameters specific to lip shape during speech (such as lip rounding and lip closure) [3, 4, 5]. 3D lip and jaw models have also been proposed [4] that are controlled by few labial parameters. EMG measurements of muscle contraction has been given as input to drive a physically-based facial model [6].

Video rewrite [7] 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 [8] 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. [9] implements the look-ahead model. On the other hand the approach proposed by Cohen and Massaro [3] implements Løfqvist's gestural theory of speech production [10]. The system uses overlapping dominance functions to specify how close the lips come to reaching their target value for each viseme. LeGoff [11] extended the formula developed by Cohen and Massaro to get a n-continuous function.

3. FACIAL MODEL

Our facial model is based on MPEG-4 standard [12, 13]. Two 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. As our goal is to compute lip movements from data, we do not consider the first FAP that defines visemes, rather we are proposing a method to define them as exposed in this paper. The FAP corresponding to expressions is not considered either, we also use here our own set of expressions [14]. But all other FAPS (the remaining 66) have been implemented.

The model uses a pseudo-muscular approach [15]. The muscle contractions are obtained through the deformation of the polygonal network around feature points. Each feature point corresponds to skin muscle attachment and follows MPEG-4 specifications.

---

1 We would like to thank Stefano Pasquariello for having implementing of the 3D facial model.
The model has been divided into regions defined around each feature point (see figure 1) and that correspond to muscle contraction major zone of influence [16]. Points within a single region may be modified by several FAPS, but they can re-act differently depending on the considered FAP (for example, given a region $r$ and two FAPS $FAP_i$ and $FAP_j$ that both act on $R$, $FAP_i$ may have a greater influence on each point of the region $R$ than $FAP_j$). Furthermore, the deformation due to a FAP is performed in a zone of influence that has an ellipsoid shape whose centroid is the feature point (see figure 2). 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 3 and 4). The displacement of a point depends also on which region it belongs to and how this region is affected by a given FAP. Let $W$ be the deformation function, $W'$ be the function defining the effect of a FAP on a region, and $FAP_i$ the value of the FAP. The displacement $\Delta P_j$ of a point $P_j$ that belongs to the area of influence of the $FAP_i$ and a region $r_k$ is given by:

$$\Delta P_j = F_i \ast W_j \ast W_{ki}$$

Where $F_i$ is the intensity of $FAP_i$, $W_j$ is the value of the deformation function at the point $P_j$. This value depends on the distance between $P_j$ and the feature point of the area of influence. Of course this value is equal to zero for all points outside this area of influence. This allows us to modify only the points belonging to a given area of influence of a FAP without modifying the other points of the facial model. On the other hand $W_{ki}$ represents the weight of deformation of the FAP $FAP_i$ over the region $R_k$. This factor specifies how the region $R_k$ is affected by the $FAP_i$. This factor can be set to zero if a region should not be affected by a given FAP. In figure 2 we can see the zone of influence of 3 FAPS (all have ellipsoid shape) and the 3 feature points where the FAPS are first applied. In figure 1 the regions over the same part of the face are shown. To be sure that under the action of the FAPS for the eyebrow, the points within the eyelid region will not be affected, all factors $W_{ki}$ between the eyelid region and the FAPS for the eyebrow are set to zero. Therefore the eyelid will have a null displacement under these particular 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 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 5, 7 and 8). Let $W_1$ the deformation function for a given FAP $FAP_1$ and $W_2$, the deformation function for the area of bulges.
This displacement $\Delta P_j$ of a point $P_j$ in the area of B of the bulges is computed as:

$$\Delta y_j = \Delta P_j \times K_i \times W_1 \times W_2$$

$W_1$ is the displacement function as defined in the equation 1 and depends on the distance between the point $P_j$ and the feature point of the area of influence; while $W_2$ is function of the distance between the point $P_j$ and the boundary between the area A and the area B (as defined in figure 5). $K_i$ is a constant that characterizes the bulge height. The course of the function $W_2$ is given in figure 6 and an example of bulges creation is given in figure 7.

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) [19]. Figure 9 and 10 represent the three-dimensional coordinates (LH, LW, LP) for the 7 Italian stressed, 5 unstressed and 3 cardinal vowels, and the 21 Italian consonants in the /aCa/ context, averaged along all the subjects' productions and 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 [19]. From the figure 10 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 consonantal articulatory target, ranges 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 of the 4 articulatory parameters over time.

Using a neural network model, we have written the curve as the weighted sum of radial basis functions $f_i$ of the form:

$$f_i(t) = \sum_{j=1}^{9} \lambda_j e^{-\frac{|t - \text{time}(t_j)|^2}{\sigma_j^2}}$$

Where $\lambda_j$ and $\sigma_j$ are the parameters that define the radial basis function. The approximation method tries to minimize the equation:

$$\min(f_i(t) - \text{Curve}_i(t))^2$$

that is we have to find the $\lambda_j$ and $\sigma_j$ that best verify this equation. For each VCV sequence we have 5 curves that corresponds to the 5 pronunciations by the same speaker of VCV. We are using these 5 examples giving us 5 Curve, (1 ≤ i ≥ 5) to be approximated by radial basis functions. Each radial basis function is characterized by 9 pairs $(\lambda_j, \sigma_j)$. We want to characterize the curves for the first V, the C and then the last V. For example when we want to characterize the curves for C, we define a single pair $(\lambda_c, \sigma_c)$ 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 $\lambda_c$ and $\sigma_c$ 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 $\lambda_j$ and $\sigma_j$:

$$\frac{\partial f_i(t)}{\partial \lambda_j} = \lambda_j e^{-\frac{|t - \text{time}(t_j)|^2}{\sigma_j^2}}$$

$$\frac{\partial f_i(t)}{\partial \sigma_j} = \lambda_j e^{-\frac{|t - \text{time}(t_j)|^2}{\sigma_j^2}} * 2 * \frac{|t - \text{time}(t_j)|^2}{\sigma_j^2}$$

Results of the approximation of the original curves for several lip parameters are shown in the figures 11, 12, 13 and 14.
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 for 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 figures 15 and 16).

For the moment we chose sequences of the type /aCa/ where C is one of the consonants /p, f, t, s, l, \( \lambda, l / , i.e. the most preferred consonants in the identification tests of the visible articulatory movements [19, 20]. 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 [19, 20]).

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 [21], produced by Festival [22], will be synchronized with articulatory movements. We are also planning to pursue perceptual study to evaluate the intelligibility of our lip model.

6. REFERENCES


