Triphone-based Coarticulation Model

Elisabetta Bevacqua

Department of Computer and System Science
University of Rome "La Sapienza"
elisabetta.bevacqua@libero.it

Catherine Pelachaud

LINC - Paragraphe
IUT of Montreuil - University of Paris 8
c.pelachaud@iut.univ-paris8.fr

Abstract

Our model of lip movements is based on real data (symmetric triphone 'VCV') from a speaker on which was applied passive markers. Target positions of vowels and consonants have been extracted from the data. Coarticulation is simulated by modifying the target points associated to consonants depending on the vocalic context using a logistic function. Correlation rules are then applied to each facial parameters to simulate muscular tension. Our model of lip movements is applied on a 3D facial model compliant with MPEG-4 standard.

1. Introduction

Our goal is to create a natural talking head with lip-readable movements. Based on real data, collected with an opto-electronic system that applies passive markers on the speaker's face [1], we have built a simple model of lips movement approximating the data and including coarticulation and correlation rules. Our 3D facial model is compliant with MPEG-4 standard [2]. We are using Festival [3] as speech synthesizer. Festival decomposes the text given in input into a sequence of phonemes with their duration. These temporal values are necessary to synchronize the lips movements with the audio stream. The system associates to each phoneme the corresponding viseme and then applies coarticulation rules. To reinforce labial tension effects, correlation rules are considered. Lip shapes are defined using labial parameters that have been determined to be phonetically relevant parameters [1]. The lip model presented here is embedded in a larger system that drive the animation of an Embodied Conversational Agent [15] that is being developed as part of the european project MagiCster[1]. After presenting the state of the art, we outline the architecture of our agent system 3. In section 4, we present our coarticulation model that is illustrated by an example. Correlation rules are explained in section 5. Section 6 provides a comparison study of our algorithm while section 7 concludes the paper.

[1]IST project IST-1999-29078, partners: University of Edinburgh, Division of Informatics; DFKI, Intelligent User Interfaces Department; Swedish Institute of Computer Science; University of Bari, Dipartimento di Informatica; University of Rome, Dipartimento di Informatica e Sistemistica; AvartarME.

2. State of the art

The approach of Pelachaud et al [4] implements the look-ahead model. Lip movements are viewed as a sequence of key positions (corresponding to phonemes belonging to non-deformable clusters) and transition positions (corresponding to phonemes belonging to deformable clusters). The model of coarticulation proposed by Cohen and Massaro [5] implements Löfqvist's gestural theory of speech production [6]. Their model has been evaluated using different tests of intelligibility [7]. Le Goff and Benoit [8] extended the formula developed by Cohen and Massaro [5] to get an n-continuous function and proposed a method for automatically extracting the parameters defining the dominance function from data measured on a speaker. Cosi and Perin's model [9] further improved Cohen and Massaro's work [5] introducing two new functions in order to obtain a better approximation of the curve: the temporal resistance and the shape. To overcome some difficulties such as those encountered in the realization of bilabial consonant stops for which labial closure is necessary but is not always maintained if one uses the dominance functions, Reveret et al [10] adapt Öhman's coarticulation model [11]. Recently a new version of Baldi [12] has been implemented. This talking head, named Baldini, speaks Italian.

3. System Presentation

Our model is based on a set of data that has been recorded at the Istituto di Fonetica e Dialettologia - C.N.R. of Padova - by means of ELITE. ELITE is an opto-electronic system that applies passive markers on the speaker's face [1].
Our model uses six parameters that have been found to be phonetically and phonologically relevant to describe visemes [1]: upper lip height (ULH), lower lip height (LLH), lip width (LW), upper lip protrusion (UP), lower lip protrusion (LP), and jaw (JAW). Our 3D facial model is MPEG-4 compliant [2]. Two sets of parameters describe and animate the model: facial animation parame-

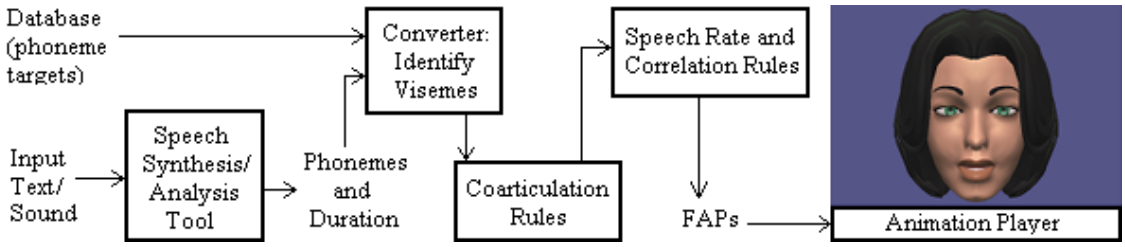
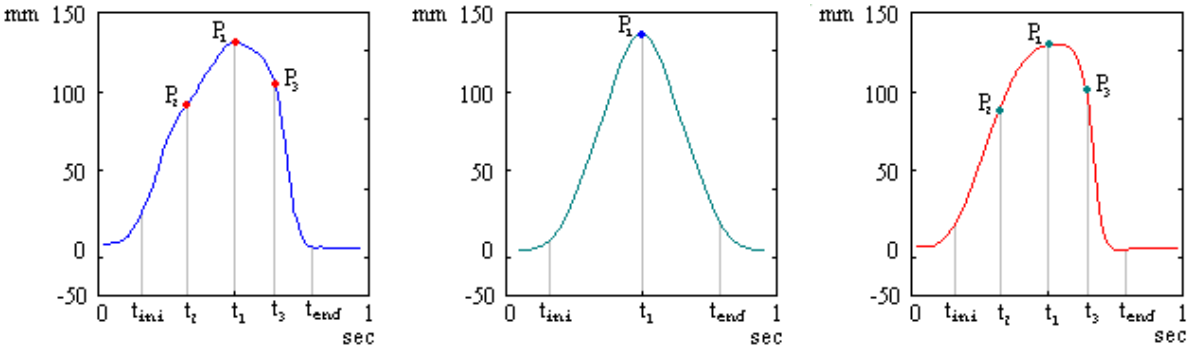


Figure 1: Architecture.



(a) Original curve sampled every $1/10^{th}$ sec. *(b)* B-spline curve generated using only one target point. *(c)* B-spline curve generated using three target points

Figure 2: Upper Lip Height of vowel /a/.

ters (FAPs) and facial definition parameters (FDPs). The FDPs define the shape of the model while FAPs define the facial actions. When the model has been characterized with FDPs, the animation is obtained by specifying for each frame the values of FAPs.

3.1. System modules

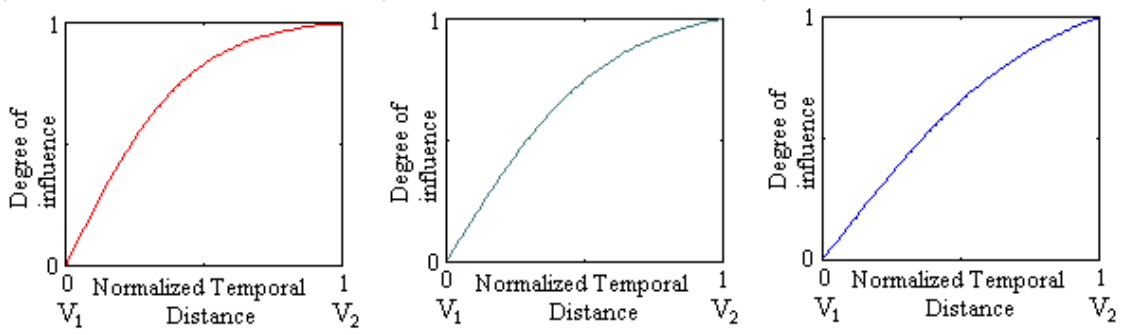
Our system takes as input a text file that is first decomposed into a list of phonemes with their duration. The decomposition may be done by Festival [3] when working with a speech synthesis or using analysis tools when using real speech. Our system includes two modules (see Fig. 1): the converter of the phonetically relevant parameters (PRPs) into facial animation parameters (FAPs) and the coarticulation rules. Within the module 'converter', a first step consists in defining fundamental values, called *target points*, for every parameter of each viseme associated with vowels and consonants; a target point corresponds to the target position the lips would assume at the apex of the viseme production. Such values, are extracted from the original data [1]. As noted in [1], the apex of a viseme production may not necessarily occur at the median of the viseme production; our definition of target points maintain such property. In Fig. 2(c) we can see that time t_1 is not exactly at the middle of the interval. Indeed the data represent values over time of six phonetically relevant parameters (LW, UP, LP, ULH, LLH, JAW) of a speaker saying 'VCV string where the first 'V is stressed while the second is unstressed.

The second module performs the computation of coarticulation effects which is done through the use of a mathematical function (called logistic function) to simulate the influence of vocalic contexts over consonants. Such an influence can involve a change of the consonantal target values.

Finally, we compute the interpolation between the target points; the resulting function represents the temporal evolution of each phonetically relevant parameter. The interpolation is obtained using interpolating B-spline which ensures the curves will go through the target points. Let us explain in further details the steps of the algorithm.

3.2. Targets for Vowels

The original curves corresponding to 'VCV have been sampled every $1/10^{th}$ sec; in average there are 150 sample points (corresponding to 1.5 sec) for each curve 'VCV. Since considering all the sample points to characterize each viseme would be too cumbersome, we decided to represent each curve by a few control points. As a first approach we chose as target point the maximum or the minimum of the curve that represents the vowel (see Fig. 2(a)). We notice that, considering only one target



(a) b=4.5 *(b)* b=3.5 *(c)* b=2.5

Figure 3: Example of influence functions varying the factor b in a sequence $V_1 C_1 \dots C_n V_2$ for Anticipatory Coarticulation.

point for each labial parameter (see Fig. 2(b)), the interpolating B-spline function intersects the target point without any notion of the shape of the original function and, as a result, of the breadth of the original curve around the target. Thus, vowels are not fully represented[2]. To get a better representation of the characteristics of a vowel, we consider two more points for every labial parameters (see Fig. 2(c)), one on the left of the target point and one on the right. Such values are not randomly chosen but individuated from the original data; let us call t_{ini} and t_{end} respectively the init and the end times of the vowel. t_1 represents the time of the vowel at its apex corresponding to the time of the target point P_1. We define two more time values, t_2 and t_3 as:

$$t_2 = \frac{t_1 - t_{ini}}{2},$$
$$t_3 = \frac{t_{end} - t_1}{2},$$

t_2 and t_3 are, respectively, the time of two new points P_2 and P_3, that are exactly at the median between t_{ini} and t_1 for t_2, and between t_1 and t_{end} for t_3. So each vowel is defined by three target points for each lip parameter and, since there are 6 labial parameters, vowels are determinate by 3x6=18 target points. As a result, the simulated curve better represents the shape of the original curve (Fig. 2(c)).

3.3. Targets for Consonants

Unlike for vowels, we only consider the target point that corresponds to the minimum or the maximum of the curve that represents a consonant. Often, visemes associated with consonants do not exhibit stable lip shape, rather they strongly depend on the vocalic context, that is on the adjacent vowels: this is one of the manifestations of the phenomenon of coarticulation. To take into account the vocalic context we gathered, for each labial parameter, all the targets of a consonant in every possible context (/a, e, i, o, u/) from the original 'VCV data[3]. For instance, for the consonant /b/, the targets points are extracted from the contexts /aba/, /ebe/, /ibi/, /obo/ and /ubu/. Since there are 5 possible contexts and 6 labial parameters, each consonant is defined by 6x5=30 target points.

4. Coarticulation

To be able to represent visemes associated to consonants, we need to consider the vocalic context surrounding the consonant. At first we determine which vowels will influence the consonants in $V_1 C_1 \dots C_n V_2$ sequence [13]. We use the properties established by phonetic studies [1, 14] that claim that vowels are linked by a hierarchical relation for their degree of influence over consonants:

$$u > o > i > e > a.$$

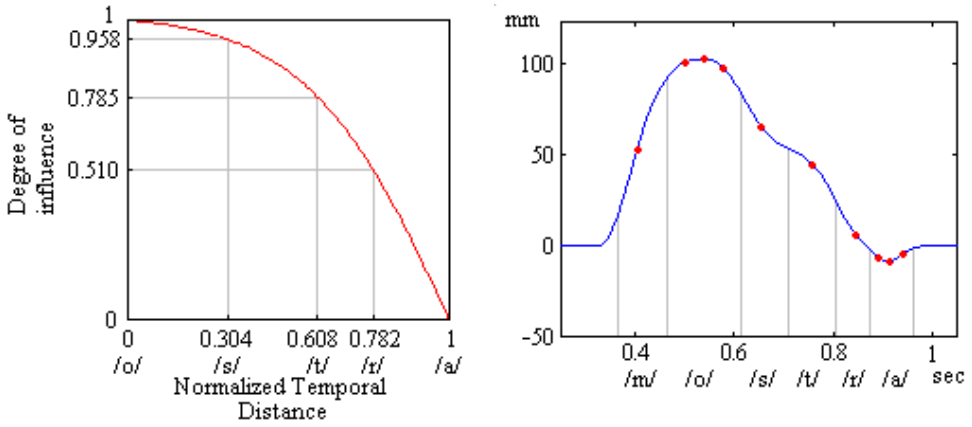
The influence that a vowel exerts on a consonant appears mainly on the labial parameters that characterise the vowel. For example for /u/ we consider the ULP and LLP parameters (Upper and Lower Lip Protrusion); while for /a/ we consider ULH and LLH (Upper and Lower Lip Height).
The influence of a vowel over adjacent consonants, on the labial parameters that characterise it, is determined through a mathematical function, called *logistic function*, whose analytic equation is:

$$f(t) = \frac{a}{1+e^{-bt}} + c,$$

where a, b, c are three constants. The parameter t corresponds to the temporal distance of a consonantal target from a vowel. The functions $f(t)$ allows us to obtain carry-over coarticulation (influence that a vowel exerts on the following consonants as /oo/ over /k/ in the word

[2] In all the figures shown in these sections, all the generated curves start with the lips at rest position, unlike the original curves. The rest position for the lips corresponds to the lips slightly closed.

[3] We consider only symmetric context as it is the only data available to us.



(a) Influence of /o/ on /s/, /t/ and /r/ in the word /mostra/. *(b)* Generated curve. ULP for the Italian word /mostra/.

Figure 4: Example.

/book/) and anticipatory coarticulation (influence that a vowel exerts on the previous consonants as /u/ over /strstr/ in the sequence /istrstru/ of 'sinistre structure'). The constants a and c force the function to be defined between 0 and 1, this ensures that the effects of the influence of a vowel do not modify the consonantal targets over a maximum value or below a minimum value. The constant b defines the steepness of the curve that represents different degrees of influence, by varing its value we may obtain different curves. We are using three curves to simulate several coarticulation effects that some vowels may have over consonants (see Fig. 3); for istance, b=4.5 defines a steeper curve which means that coarticulation effect is stronger just after in case of anticipatory coarticulation (or just before for carry-over coarticulation) the considered vowel.

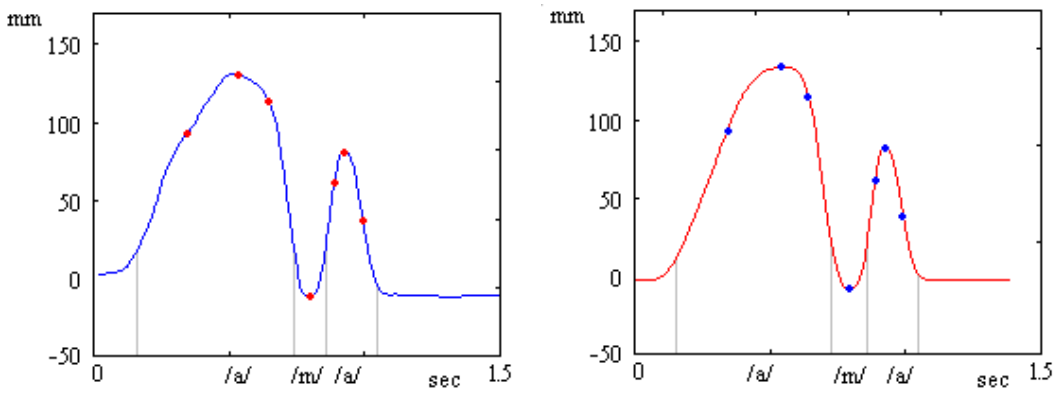
Once determined which vowel (V_1 or V_2) has the strongest influence over the consonants in a sequence $V_1C_1 \ldots C_nV_2$, the function f is applied to the labial parameters of the consonants that characterize best the vowel. To simplify the computation, the time interval between the vowels has been normalized. So time t=0 corresponds to the occurrence of V_1, and time t=1 corresponds to V_2. The consonants are placed on the abscissa depending on their temporal normalized distance from the vowels (see Fig. 4(a)). Let $t_1 \ldots t_n$ be their corresponding time. The level of influence of a vowel on a consonant C_i is given by the value of $f(t_i)$. The function f ensures that the first consonant after V_1 (or before V_2 for anticipatory coarticulation) is strongly influenced while the last consonant before V_2 (or after V_1) is minimally influenced. $f(t)$=0 means no influence is exercised on the consonant while $f(t)$=1 means the opposite, that is the considered labial parameters of the consonant will be set equal to the labial parameters of the vowel.

4.1. An example

To illustrate how our algorithm works, let us consider the sequence /ostra/ taken from the Italian word 'mostra' ('exhibition'). As we do not have data on asymetric sequence of the type /V_1CV_2/, the target of a consonant is calculated as a linear interpolation of the values for the consonant in the contexts /V_1CV_1/ and /V_2CV_2/. In our example, the targets for /s/, /r/ and /t/ in the context /oCa/ are obtained by linearly combining the targets defined in /aCa/ and /oCo/ contexts. The viseme associated with /o/ is mainly characterised by its lip protrusion parameters [13, 1], so we compute the influence of /o/ over the UP and LP parameters of the consonants /s, t, r/. Since such an influence is quite strong, the algorithm choses the influence function with b=4.5 that is the steeppest curve. The duration of each phoneme is provided by Festival. We normalized the temporal distance of each consonant from the vowel position /o/. The function f gives the value 0.958 for /s/ (Fig. 4(a)); which means that /s/ will be influenced 95.8% by the vowel /o/ and consequently that the target point of /s/ will be modified accordingly. We repeat this computation for /r,t/. Then we calculate the B-spline curve that goes through these points (see Fig. 4(b)).

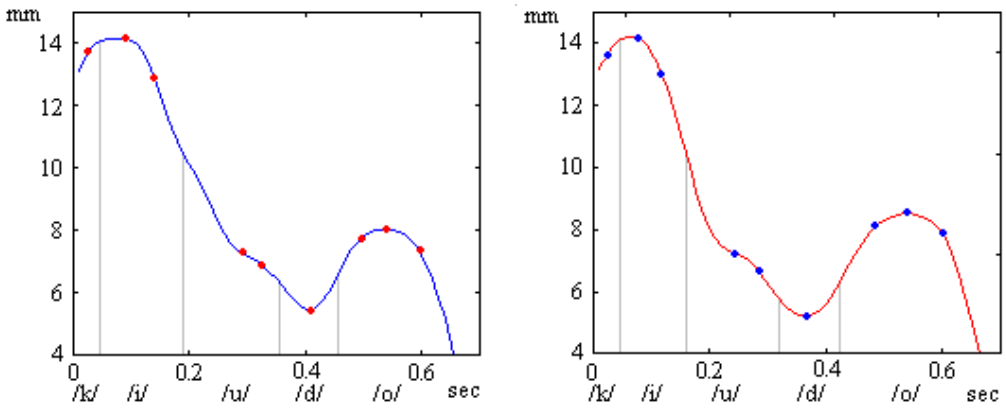
5. Correlation rules between FAPs

Since our 3D facial model is compliant with MPEG-4 standard, animation is obtained calculating the displacement corresponding to the FAPs. To simulate muscular tension, we have introduced some rules that take into account the correlation between the FAPs. For instance, when a bilabial consonant (as /b/) is uttered, lip compression must be simulated (specially at slow speech rate): if the distance between the FAPs on the upper internal lip and those on the lower internal lip decreases till a nega-



(a) Original curve. *(b)* Generated curve.

Figure 5: Upper Lip Height for the sequence /ama/.



(a) Original curve. *(b)* Generated curve.

Figure 6: Lower Lip Height for the word /chiudo/.

tive value, the FAPs on the external boundary are further lowered down, and the FAPs on the internal boundary are slightly moved inward. Similarly, when labial width increases, lips are stretched getting thinner, therefore a movement involving the FAPs acting on the lip corners, must involve FAPs on the external lip boundary that are lowered down to accentuate the effect of lip thinness during stretching.

5.1. Speech Rate

Labial animation strongly depends on the speech rate. At a slow speech rate lip movement is well pronounced reaching the maximum value of the targets for each viseme whereas, at a fast rate, the targets of every viseme are diminished forcing the lips into a shorter movement. For instance lip height is wider when an open vowel occurs and, when a bilabial consonant is uttered, lip compression is stronger. On the other hand, at a fast speech rate, the lip movement amplitude is reduced. That is the lips do not reach their target values but approximate them. To simulate the effects of fast speech rate, we recompute the value of the target points to be closer to the neutral position; whereas, for slow speech, the lips have time to fully reach their target values.

6. Comparison

As a first test, we compared the original curves of 'VCV data with those generated by the algorithm. Fig. 5 shows the original and the generated curves for the ULH (Upper Lip Height) parameter for the sequence /'ama/. Phonemes of the sequence 'V, C, V are shown by vertical lines. As already mentioned above, the lips start at the rest position for the generated curve while they might not do so for the original curve. We can notice that the original curve is quite well reproduced.
To evaluate the coarticulation model, we compare the original curves from real data representing the Italian

word /chiudo/ ('I close') with those generated by our algorithm. In Fig. 6 the original and the generated curves for the LLH (Lower Lip Height) parameter are shown; the generated curve has been obtained after the computation of the target points calculated by the coarticolation model. Phonemes segmentation are identified by vertical lines. We can notice that the duration of the phonemes in both, the original and the generated curves, differs. For the generated curve times are given by Festival, whereas, for the original one, times come from the analysis of real speech. We can see that both curves exhibit the same behaviors.
We are aware that such a comparison may not be a sufficient evaluation test to validate our lip model. We intend pursuing perceptual tests in the near future.

7. Conclusion and Future Development

We have presented a computation model of lip movements that is based on real data and that considered a coarticulation model. The targets associated to vowels and consonants were extracted from the real data [1]. We then modify the targets of consonants depending on the vocalic contexts to simulate the effect of coarticulation. We further apply correlation rules to simulate muscular tension. Movies of our lip model may be seen at the URL: http://www.iut.univ-paris8.fr/~pelachaud/avsp03.

8. Acknowledgements

We greatly thank E. Magno-Caldognetto for having providing us the 'VCV data, C. Zmarich for analysing the data and M. Bilvi, P. Cosi, F. Tesser for their precious help. We are specially grateful to L. Grippo for his kind supervision.

9. References

[1] Magno-Caldognetto, E., Zmarich, C. and Cosi, P., "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.

[2] Pelachaud, C., "Visual text-to-speech," in *MPEG4 Facial Animation - The standard, implementations and applications*, Igor S. Pandzic and Robert Forcheimer, Eds. John Wiley & Sons, to appear.

[3] Taylor, P., Black, A. and Caley, R., "The architecture of the Festival Speech Synthesis System," in *Proceedings of the Third ESCA Workshop on Speech Synthesis*, 1998, pp. 147–151.

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

[5] Cohen, M. M. and Massaro, D. W., "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.

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

[7] Massaro, D., *Perceiving Talking Faces : From Speech Perception to a Behavioral Principle*, Bradford Books Series in Cognitive Psychology. MIT Press, 1997.

[8] LeGoff, B., *Synthèse à partir du texte de visage 3D parlant français*, Ph.D. thesis, Institut National Polytechnique, Grenoble, France, 1997.

[9] Cosi, P., Magno Caldognetto, E., Perin, G. and Zmarich. C., "Labial coarticulation modeling for realistic facial animation," in *Proceedings of ICMI 2002*, Pittsburgh, PA, USA, October 14-16 2002.

[10] Reveret, L., Bailly, G. and Badin, P., "MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," in *Proceedings of ICSLP'00: International Conference on Spoken Language Processing*, X. Tang B. Yuan, T. Huang, Ed., Beijing, China, 1996, vol. II, pp. 755–758.

[11] Öhman, S.E.G., "Numerical model of coarticulation," *Journal of Acoustical Society of America*, vol. 41, no. 2, pp. 311–321, 1967.

[12] Cosi, P., Cohen, M. M. and Massaro, D. W., "Baldini: Baldi speaks Italian!," in *Proceedings of ICSLP 2002, 7th International Conference on Spoken Language Processing*, Denver, Colorado, September 16-20 2002.

[13] Benguerel, A. P. and Cowan, H. A., "Coarticulation of upper lip protrusion in french," *Phonetica*, vol. 30, pp. 40–51, 1974.

[14] Walther, E.F., *Lipreading*, Nelson-Hall, Chicago, 1982.

[15] Pelachaud, C., Carofiglio, V., De Carolis, B. and de Rosis, F., "Embodied Contextual Agent in Information Delivering Application," in *First International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS)*, Bologna, Italy, July 2002.