Expressive Audio-Visual Speech

Elisabetta Bevacqua
Department of Computer and System Science
University of Rome “La Sapienza”
Email: elisabetta.bevacqua@libero.it

Catherine Pelachaud
LINC - Paragraphe
University of Paris 8
Email: c.pelachaud@iut.univ-paris8.fr

We aim at the realization of an Embodied Conversational Agent able to interact naturally and emotionally with user. In particular, the agent should behave expressively. Specifying for a given emotion, its corresponding facial expression will not produce the sensation of expressivity. To do so, one needs to specify parameters such as intensity, tension, movement property. Moreover, emotion affects also lip shapes during speech. Simply adding the facial expression of emotion to the lip shape does not produce lip readable movement. In this paper we present a model based on real data from a speaker on which was applied passive markers. The real data covers natural speech as well as emotional speech. We present an algorithm that determines the appropriate viseme and applies coarticulation and correlation rules to consider the vocalic and the consonantal contexts as well as muscular phenomena such as lip compression and lip stretching. Expressive qualifiers are then used to modulate the expressivity of lip movement. Our model of lip movement is applied on a 3D facial model compliant with MPEG-4 standard.

1 Introduction

With the development of 3D graphics, it is now possible to create Embodied Agents that have the ability to communicate verbally and nonverbally. Nonverbal communication is an important mean of communication. In particular, facial expressions can provide a lot of information. In particular they are a good window on our emotional state. Emotions are a fundamental aspect of human life influencing how we think and behave and how we interact with others. Facial expressions do improve communication; they can make clear what we are thinking, even without speaking. For example wrinkling our nose in front of something that we dislike communicates very clearly our disgust. Therefore, believable synthetic characters make the interaction between users and machine easier and more fulfilling providing a more human-like communication.

In the aim of creating a believable embodied conversational agent (ECA) able to serve as a bridge in the communication between humans and the machine, ECA ought to be empowered with human-like qualities. In particular, the agent should display 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 model of lip movement that includes coarticulation and correlation rules. Our model allows also the agent to communicate with a wide variety of expressions of emotion with the appropriate lip shape and expressivity. The system associates to each phoneme the corresponding viseme and then applies coarticulation rules. To reinforce labial tension effects, correlation and expressive rules are considered. Lip shapes are defined using labial parameters that have been determined to be phonetically relevant parameters [1].

In the next section we prevent an overview of the state of the art of audio-visual speech systems. We then describe the computational lip shape model. We follow by presenting the emotional lip model. Expressivity parameters for lip shape are presented in section 4. Our system is presented in section 5. Coarticulation model is explained in section 5.1 while correlation rules are described in section 5.3. We end the paper by describing some objective tests we perform in view of evaluating our model.

2 State of the art

The model of coarticulation proposed by Cohen and Massaro [2] implements Löfqvist's gestural theory of speech production [3]. The system uses overlapping dominance functions to specify how close the lips come to reaching their target value for each viseme. The final lip shape (target value) is obtained by taking the weighted average of all the co-occurring dominance functions. To model lip shapes during emotional speech, M. M. Cohen and D. W. Massaro add the corresponding lip shapes of emotion to viseme definition [4]. Le Goff and Benoit [5] extended the formula developed by Cohen and Massaro [2] 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 [6] further improved Cohen and Massaro's work [2] 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 [7] adapt Öhman's coarticulation model [8]. This model suggests that there is a dominant stream of vowel gestures on which are superimposed consonant gestures. Other rule-based systems include work by [9, 10, 11]. Several data-driven approaches have been proposed [12, 13, 14]. An example of this technique is Video-Rewrite developed by Bregler.et.al [12]. The animation of a new sequence of lip animation is done by concatenating data extracted from a large database of images. Models including emotional speech have been proposed. They are based on learning the mapping between expression and lip shapes [15, 16].

Our approach differs from other ones as we have added qualifiers parameters to simulate expressivity in lip movements. Our model belongs to the rule-based models. We drive our computational model of emotional lip shapes from real data for neutral and emotional speech [1].

3 Computational lip model

Our lip model is based on captured data of triphones of the type $'VCV$, collected with the optical-electronic system ELITE that applies passive markers on the speaker's face [1], where the first vowel is stressed whereas the second is unstressed. The data covers not only several vowels and consonants for the neutral expression but also different emotions, namely joy, anger, sadness, surprise, disgust and fear [1]. The original curves from the real speech data represent the variation over time of the 7 phonetically and phonologically relevant parameters that define lip shapes: upper lip height (ULH), lower lip height (LLH), lip width (LW), upper lip protrusion (UP), lower lip protrusion (LP), jaw (JAW) and lip corners (LC). Apart from the values of each labial parameter, the data contains the acoustic segmentation into V, C, V, marked in the figures by vertical lines. We can see that we may encounter asynchrony of the labial target with the acoustic signal, according to the parameter and/or the phoneme. That is, the apex of the labial parameters of a viseme may not always correspond to the apex of the phoneme production [1]. Rather than considering the whole list of sample points of a curve which would be too cumbersome for later manipulation of the coarticulation function, we decided to represent each curve by a few control points, that we call target points, and have a B-spline curve connecting these control points. A target point corresponds to the target position the lips would assume at the apex of the viseme. These targets have been collected analyzing the real data and are stored in a database.

As a first approach we chose as target point the maximum or the minimum of the curve that represents the vowel (see Figure 1(a)). We notice that, considering only one target point for each labial parameter (see Figure 1(b)), does not produce good result: 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[1]. To get a better representation of the characteristics of a vowel, we consider two more points for every labial parameters (see Figure 1(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_{init} 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, $t_2 = \frac{t_1 - t_{init}}{2}$ and $t_3 = \frac{t_{end} - t_1}{2}$. 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_{init} 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 7 labial parameters, vowels are determinate by 3x7=21 target points. As a result, the simulated curve follows tightly the shape of the original curve (see Figure 1(c)).

Unlike for vowels, we only consider the target point that corresponds to the minimum or the maximum of the curve that represents a consonant (see figure 1(c)). 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 vocalic contexts (namely here /a, e, i, o, u/) from the original 'VCV data[2]. 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 7 labial parameters, each consonant is defined by 7x5=30 target points. Figure 1(c) illustrates how the target representation we used does represent vowel and consonant with very good accuracy. We consider the

[1] 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

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

following consonants /b, tS, d, f, g, k, l, m, n, p, r, s, SS, t, th, v, w, z/; while for the vowels we have data for /a, e, i, o, u/ stressed and not stressed.

Vocalic and consonantal target data points are stored in a database. Besides targets values other information have been collected like the vowel or the consonant that defines the context, the duration of the phoneme and the time of the targets in this interval. A database is created for each of the six fundamental emotions and for the neutral expression as well.

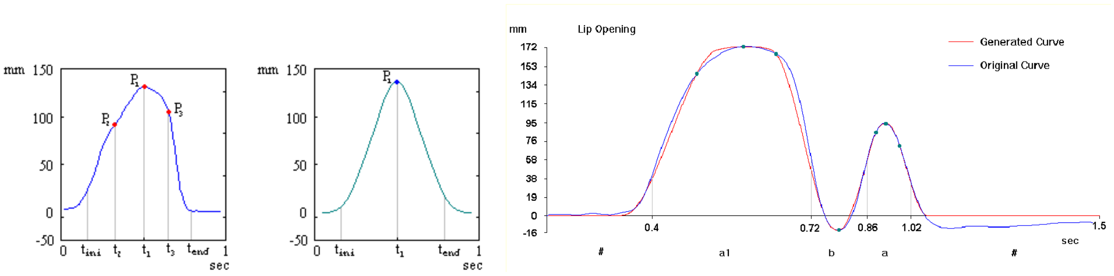

Figure 1: Target points characterization - for vowel: (a) original curves on which target points are extracted and (b) vowel characterization using only one target value; (c)Vowel and consonant target points for /aba/.

4 Emotion Characterization

Our computational model is based on real data that have been recorded for six emotions and for the neutral speech. Target values extracted from real speech characterized visemes for a given emotion. In order to allow the modelling of new emotion or emotion with different intensity, we need a way of specifying a new set of visemes. For this purpose we have developed a method that derives the target points that defined the visemes for the desired emotion as a weighted combination of existing data. The weights are associated to each labial parameter. Differentiated weights for each labial parameter refines the characterization of the new corresponding visemes set.

To do so, we describe each recorded emotion by a 7x7 matrix. The rows correspond to the seven recorded emotions, whereas the columns are the labial parameters. A value in the matrix represents the percentage of dependence that the corresponding lip parameter has on the corresponding emotion. Therefore, the value of the targets for each labial parameter will be a weighted combination of the targets in the emotions that have a value on the column different from zero.

	ULH	**LLH**	**JAW**	**LW**	**UP**	**LP**	**CR**
Neutral	0	0	0	0.2	0.1	0.1	0.2
Joy	1	1	1	0.8	0.9	0.9	0.8
Surprise	0	0	0	0	0	0	0
Fear	0	0	0	0	0	0	0
Anger	0	0	0	0	0	0	0
Disgust	0	0	0	0	0	0	0
Sadness	0	0	0	0	0	0	0

Table 1: Matrix of the emotion **Light-Joy**.

Let us consider the consonant /b/ in the triphone /aba/ uttered with the emotion ‘light-joy’. The matrix of this emotion is shown in Table 1. Now, let N_ab_a be the target of the lip width parameter of /b/ uttered in a neutral emotion and let J_ab_a be the target of the same lip parameter of /b/ uttered in the ‘joy’ emotion. The new value of this target LJ_ab_a in the light-joy emotion will be:

$$LJ_ab_a = 0.2 * N_ab_a + 0.8 * J_ab_a$$

As consequences, the lip width will be less wide in the ‘light-joy’ emotion than in the ‘joy’ emotion.

4.1 Expressivity qualifiers

We also defined two qualifiers to modulate the expressivity of a lip movement. The first one, *Tension Degree*, can be *Strong*, *Normal* and *Light* and it allows us to set different intensities of muscular strain. Strong tension may appear when the expressions of emotions like fear and anger are shown. Tension is modelled by varying lip thickness (defined as the distance between outer and inner lip contours) as well as ensuring less labial movements from one viseme to the next (see figure 2(a)). The second qualifier, *Articulation Degree*, can take the values *Hyper*, *Medium* and *Hypo*. During hypo articulation, it may happen that lip targets may not reach their apex. On the other hand during hyper articulation target overshoot may happen. The lip opening parameters has greater value (see figure 2(b)).

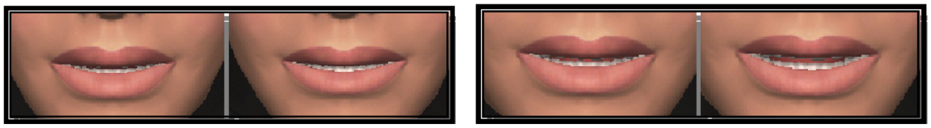

Figure 2: Examples of the expressivity qualifiers. (a) Tension degree variation: normal and strong values; (b) Articulation variation: normal and hyper articulations.

5 System Overview

Our system works as follow. It takes as input a text an agent should say. The text file is decomposed by the speech synthesizer, Festival [17] into a list of phonemes with their duration. The first step of our algorithm consists in looking up the values of the *target points* for every parameter of each viseme associated with the vowels and consonants that appear in the phoneme list. Then coarticulation rules are applied that modify the target values based on vocalic and consonant contexts. Speech rate rule may further modify these values. Once the targets are computed, the correlation rules are applied to modulate labial movement and add qualitative expressivity.

5.1 Coarticulation rules

Once all the necessary visemes have been loaded and modified according to the emotions, coarticulation rules must be applied. In fact, to be able to represent visemes associated to vowels and consonants, we need to consider the context surrounding them. Most of the time, the target point corresponding to a consonant gets modified by its surrounding vowel. But in some cases, target points of vowels may be modified also. Indeed, bilabial and labiodental consonants exert an influence on the lip opening parameter whereas the consonants /tS/, /SS/, /w/ and /q/ influence the LW, ULP and LLP parameters. Moreover some consonants ({/tS/, /SS/, /w/, /q/, /b/, /m/, /p/, /f/, /v/}) exert an influence on some labial parameters of adjacent consonants. To model coarticulation effects due to the vocalic context over consonant, we consider the following property of vowels: vowels are linked by a hierarchical relation for their degree of influence over consonants ($u > o > i > e > a$) [18]. At first we determine which vowels will affect the consonants in $V_1 C_1 \ldots C_n V_2$ sequence and which labial parameters will be modified by such an influence. Vowels act mainly over those lip parameters that characterize them. The new targets of the consonants for each lip parameter are computed through a linear interpolation between the consonantal targets ${}_{V_1}C_{i\,V_1}$ in the context deriving from the vowel V_1 and the consonantal targets ${}_{V_2}C_{i\,V_2}$ in the context of the vowel V_2:

$$ {}_{V_1}C_{i\,V_2} = k * {}_{V_1}C_{i\,V_1} + (1-k) * {}_{V_2}C_{i\,V_2} \quad (1) $$

The interpolation coefficient k is determined through a mathematical function, called *logistic function*, whose analytic equation is:

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

This function represents the influence of a vowel over adjacent consonants, on the labial parameters that characterize it, and allows us to obtain carry-over coarticulation and anticipatory coarticulation (see Figure 3). We determined these by trial and error. The parameter t corresponds to the temporal distance from a vowel. The constants a and c force the function to be defined between 0 and 1 both on the abscissa and the ordinate simplifying computation. The constant b defines the steepness of the curve that represents different degrees of influence. Time t=0 corresponds to the occurrence of V_1, and time t=1 corresponds to V_2. The consonants C_i are placed on the abscissa depending on their temporal normalized distance from the vowels.

The function f is applied to the labial parameters of the consonant between successive vowels. 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. Let $t_1 ... t_n$ be their corresponding time. The influence of a vowel on a consonant is given by the value of f (t_i). We can see that the function f will always vary between 0 and 1, ensuring that the first consonant after (or before for anticipatory coarticulation) will be strongly influenced while the last vowel will be minimally influenced. f(t) = 0 means no influence is exercised on the consonant while f(t) = 1 means the opposite; that is the labial parameters of the consonant will be equal to the labial parameters of the vowel. As already noted, vowels may show a distinctive behavior. Based on this, we applied the function of influence over a consonant only on the labial parameters that characterize the ‘strong vowel. Once all the necessary visemes have been calculated, lip movement over time is obtained interpolating the computed visemes using Hermite Interpolation.

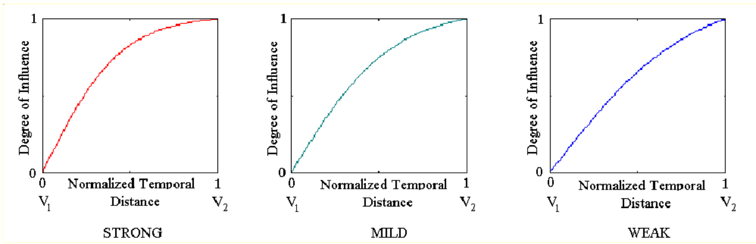

Figure 3: Logistic function with various degree of influence

5.1.1 Examples

For example, let us consider the sequence /ostra/ taken from the Italian word ’mostra’ (’exhibition’). Since the vowel /o/ is stronger than /a/, we select the target position of the consonants /s,t,r/ from the context /oCo/. Moreover, the viseme associated with /o/ is mainly characterized by its lip protrusion parameters. For example, we compute the influence of the UP parameter (Upper Lip Protrusion) of /o/ over the UP parameter of the consonants /s, t, r/. Let us consider the sequence /ostra/ taken from the Italian word ‘mostra (show). The targets for /s/, /r/ and /t/ in the contexts /o/ (i.e. /oCo/) and /a/ (i.e. /aCa/) are individuated. Thus, the consonantal targets of /s/, /t/ and /r/ in the context oCa must be calculated. The vowel /o/ exerts a strong influence over the following three consonants and the algorithm chooses the steepest influence function. Figure 4(a) shows how the logistic function is applied to define the interpolation coefficients and Figure 4(b) shows how the targets are modified using the logistic function. In the example ‘mostra’, we have an anticipatory coarticulation (see Figure 4(b)(a)). The function f(s)= I_s gives for /s/ the value 0.958; it means that /s/ will be influenced of 95.8% by the vowel /o/ and, consequently, /a/ exerts on the same consonant an influence of 4.2%, corresponding to 1-f(s)=0.42. If we indicate with ${}_oS_o$ the target point of the upper lip protrusion parameter in the context /oso/ and with ${}_aS_a$ the target value of the same parameter in the context /asa/, then the new target of /s/ in context /osa/ will be:

$${}_os_a = f *_o s_o + (1-f) *_a s_a = 0.95 *_o s_o + (1-0.95) *_a s_a$$

Because of the strong influence exerted on /s/ by /o/, the target ${}_os_a$ is almost the same as the target ${}_os_o$. We repeat this computation for /r, t/. Instead, since /r/ is further away from /o/ than /s/ is, the target ${}_or_a$ will be: ${}_os_a = 0.51 *_o r_o + (1-0.49) *_a r_a$. In the same way ${}_ot_a$ is calculated. Figure 4(b) shows the behavior of the labial parameter ULP over the sequence /ostra/. In the figure, the points indicate the targets (one per consonant, three per vowel) and the arrows indicate the displacement of the labial parameters due to the effect of the logistic function.

5.2 Speech Rate

Target points for vowels and consonants are first determined by looking up at their definition in the database: their values for each labial parameters may be found depending on the current emotion as well as the vocalic context. Coarticulation rules modify these values to

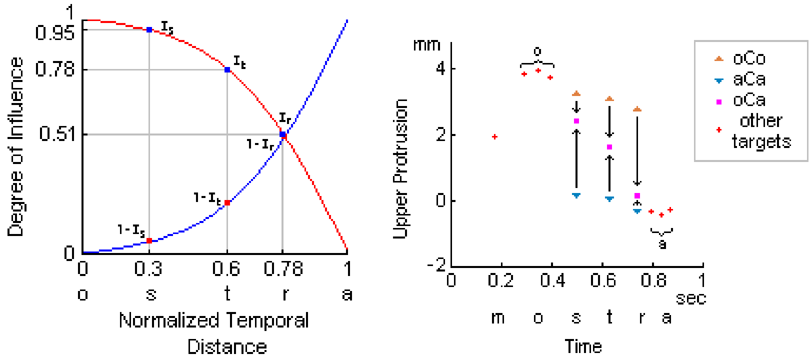

Figure 4: Coarticulation effects on consonants: (a) influence of /o/ on /s/, /t/ and /r/; (b) Alteration of UP targets for the Italian word /mostra/.

further consider the vocalic (and in some cases the consonant) context (see section 5.1). One more step is required to finalize the target points; namely the consideration of the speech rate. Indeed, speech rate strongly influences lip animation. For example, at fast speech rate, lip movement amplitude is reduced while at slow rate it is well pronounced (lip height is wider when an open vowel occurs or lip compression is stronger when a bilabial consonant is uttered). To simulate this effect, at a fast speech rate the value of targets points is modified and diminished in order to be closer to the rest position, while at a slow rate lips fully reach their targets.

5.3 Correlation rules

Once the target points are finally known, we compute the Hermite B-spline curve that goes through them. This is done for each labial parameters. Since our 3D facial model is compliant with MPEG-4 standard [19], animation is obtained calculating the displacement corresponding to the facial animation parameters (FAPs). Thus, the next step is to express each labial parameters as MPEG-4 FAPs. *Correlation rules* are applied modifying the value of the FAPs to simulate muscular tension for a particular viseme. For example, when a bilabial consonant (such as /b/) is uttered lip compression must appear: the facial parameters on the external boundary of the lips must be further lowered down. To simulate muscular tension, we have introduced some rules that take into account the correlation between the FAPs. **Rule 1**: when a bilabial consonant (such 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 negative value, the FAPs on the outer lip boundary are further lowered down, and the FAPs on the inner lip boundary are slightly moved inward. **Rule 2**: when labial width increases (under a movement involving the FAPs acting on the lip corners), lips must stretch getting thinner: FAPs acting on the outer lip boundary are lowered down to simulate this effect. **Rule 3**: when a bilabial consonant (as /b/) is uttered, lip protrusion must be less pronounced. **Rule 4**: when labial protrusion increases, the tip of the nose must be slightly lowered. **Rule 5**: when labial width increases, the nostril must be slightly spread apart.

6 Objective Evaluation Tests

To validate our method we perform some evaluation tests. We use an objective method based on the comparison between the original curves from the recorded data with those computed by our algorithm. As first example let us consider the triphone /aba/ uttered either in joy or in neutral expression. In Figure 5 the curves that represent the evolution of the labial parameter Lip Opening are shown. Since we split this original lip parameter in two, our Lip Opening curves are obtained as a sum of the parameters ULH and LLH. The generated curves are shown in Figure 5(a) whereas the original ones in Figure 5(b), in both figures the dotted line represents the lip movement in joy emotion, while the solid line describes the lip opening in neutral expression. One can notice that the joy emotion causes a reduction of the lip opening. The same behavior is also visible in the computed curves (see Figure 5(b)).

The second example is done on the Italian word ‘mamma’ (mom) and the considered emotions are disgust and neutral. Lip Opening curves are shown in Figure 5 where the dotted line represent the movement of the labial parameter lip opening in disgust emotion while the solid one is the evolution of this parameter in neutral expression. In Figure 5(c) we can see our computed curves whereas Figure 5(d) shows the original ones. The disgust emotion causes a diminution of lip opening and our model did predict this behavior. In the figure, phonemes segmentation is identified by vertical lines.

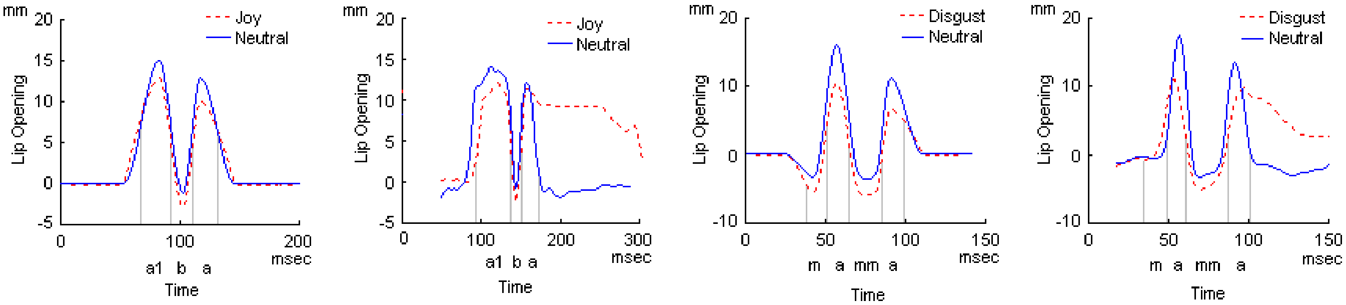

Figure 5: Lip Opening for: (a) Generated curves and (b) Original curves for the triphone /aba/ in joy and neutral expression; (c) Generated curves and (d) Original curves for the Italian word /mamma/ in disgust and neutral expression.

Figure 6(a) shows another example of comparison between original and computer curves for the Italian word ‘chiudo’ (which means ‘I close’) for the ‘neutral’ emotion. We have extended our algorithm for other languages. Figure 6(b) illustrates how our data may be adapted to other languages (for the words ‘good morning’. To do so we establish a new table of conversion between phonemes and visemes. In case a viseme may not be found in our database, we use the closest similar viseme.

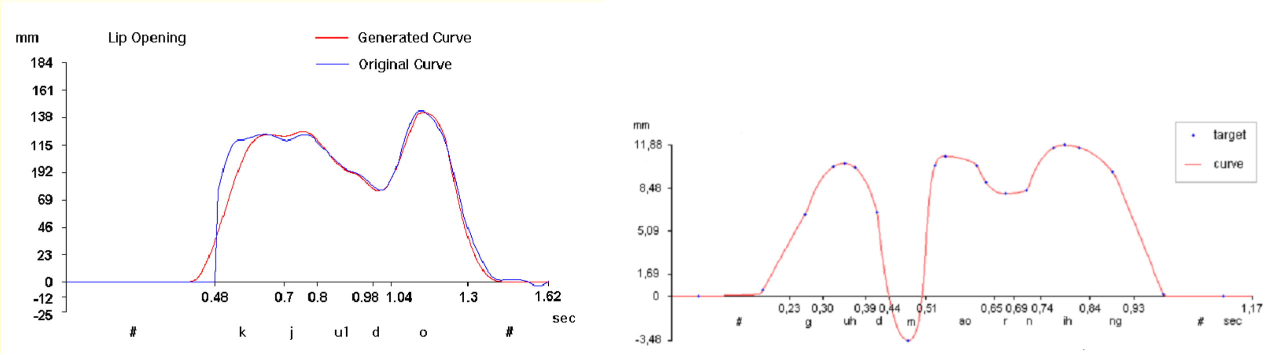

Figure 6: Lip opening parameter during neutral enunciation (a) on the Italian word ‘chiudo’ (I close) and (b) on the English words ‘good morning’.

A last specification is necessary. For the generated curves, phonemes duration is given by Festival, while for the original ones these values come from the analysis of real speech. Moreover in the example shown here, our model starts with the lips set at the neutral position; which may not be the case for real speech data. Those differences can make the calculated curves slightly different from the original ones.

7 Conclusion and Future Development

We have presented a computation model of emotional lip movement that is based on real data. The data covers neutral and emotional speech. Targets associated to vowels and consonants have been extracted from real data. The consonant targets are then modified depending on the vocalic contexts to simulate the effect of coarticulation. Coarticulation function is modelled using a logistic function that simulates the degree of influence vowels have on consonants. We further apply correlation rules to simulate muscular activity. Moreover, the simulation of specific muscular behavior (e.g. tense lips of anger) has been integrated to the model. Our next step is to perform some perceptual evaluation tests to check the feasibility of our models.

References

[1] E. Magno-Caldognetto, C. Zmarich, and P. Cosi. Coproduction of speech and emotion. In *ISCA Tutorial and Research Workshop on Audio Visual Speech Processing, AVSP'03*, St Jorioz, France, September 4th-7th 2003.

[2] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156, Tokyo, 1993. Springer-Verlag.

[3] A. Lofqvist. Speech as audible gestures. *Speech Production and Speech Modeling*, pages 289–322, 1990.

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

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

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

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

[8] S.E.G. Ohman. Numerical model of coarticulation. *Journal of Acoustical Society of America*, 41(2):311–321, 1967.

[9] C. Pelachaud, N.I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, January-March 1996.

[10] J. Beskow. *Talking Heads: Models and Applications for Multimodal Speech Synthesis*. PhD thesis, Centre for Speech Technology, KTH, Stockholm, Sweden, 2003.

[11] S.A. King, A. Knott, and B. McCane. Language-driven nonverbal communication in a bilingual conversational agent. In *Proceedings of CASA 2003*, pages 17 – 22, 2003.

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

[13] M. Brand. Voice puppetry. In *Computer Graphics Proceedings, Annual Conference Series*, pages 21–28. ACM SIGGRAPH, 1999.

[14] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *ACM Transaction on Graphics(Proceedings of ACM SIGGRAPH02)*, volume 21(3), pages 388–398, 2002.

[15] E.S. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *Proceedings of Pacific Graphics*, pages 68–76, 2002.

[16] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Computer Graphics Forum(Proceedings of Eurographics 2003)*, volume 22(3), 2003.

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

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

[19] C. Pelachaud. Visual text-to-speech. In Igor S. Pandzic and Robert Forchheimer, editors, *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, 2002.