A phonetic concatenative approach of labial coarticulation

Vincent Robert, Yves Laprie, Anne Bonneau

Speech Group, LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
FRANCE

http://parole.loria.fr
email : vrobert@loria.fr

Abstract

Predicting the effects of labial coarticulation is an important aspect with a view to developing an artificial talking head. This paper describes a concatenation approach that uses sigmoids to represent the evolution of labial parameters. Labial parameters considered are lip aperture, protrusion, stretching and jaw aperture. A first formal algorithm determines the relevant transitions, i.e. those corresponding to phonemes imposing constraints on one of the labial parameters. Then relevant transitions are either retrieved or interpolated from a set of reference sigmoids which have been trained on a speaker specific corpus. This labial corpus is made up of isolated vowels, CV, VCV, VCCV and 100 sentences. A final stage consists in improving the overall syntagmatic consistency of the concatenation.

1. Introduction

One of the most crucial aspects of the development of a talking head is the synthesis of labial movements from the knowledge of the sequence of phonemes to articulate. A good modelling of coarticulation phenomena is important to make lip reading possible.

A first approach consists in realizing the visual synthesis by manipulating video images [1, 2, 3]. Reference images correspond to monophones, diphones or triphones. In the first two cases, the effects of coarticulation cannot be captured and then rendered during synthesis. In the case of triphones, on the one hand the corpus necessary becomes huge to cover all the existing triphones, and on the other hand, more complex coarticulation effects involved in VCCV or VCCCV cannot be captured correctly. Moreover, these methods are often available for 2D images even if there are solutions to reconstruct pseudo 3D images, and they provide a talking head linked to a given speaker only.

Our objective is closer to works concerning 3D artificial talking heads [4, 5, 6, 7]. These works comprise the determination of control parameters as a function of the phone sequence to pronounce, as well the reconstruction of the head as a function of these parameters. This paper only concerns the determination of the labial control parameters. Previous phonetic results, as well as more recent studies (that of Meada [8] on two speakers, and our investigations on a ten speaker labial database[9] by means of a principal component analysis) show that the main modes of labial deformations are lip protrusion, lip stretching, and jaw opening.

The model proposed by Cohen & Massaro [5] relies on the gesture theory of Löfqvist [10] which associates one target vector to each segment. Dominance functions are used to weight target values as a function of time. Dominance functions are made up of two exponential terms, one for anticipation and the other for retention. As underlined by Beskow [11] Cohen & Massaro's model can be considered as a "time locked" model since the duration of these exponentials are context independent. The "time locked" model stipulates that the gesture starts at a constant period of time before the segment considered. This model is not well adapted to the bilabials /p,b,m/ and labial fricatives /f,v/ because dominance functions cannot capture the resistance to coarticulation. Indeed, the complete closure of lips cannot be reached through the mixture of dominance functions. Cosi [12] thus refined Cohen & Massaro's model by adding a resistance term that forces labial parameters when needed. However, as stressed by Reveret [7] coarticulation cannot be easily rendered by the overlapping of several gestures weighted by dominance functions. He thus proposed a model based on Öhman's theory [13] designed to model tongue movements. According to this theory tongue movements are piloted by vowels, to which consonant gestures are superimposed. This theory belongs to the family of the "look-ahead" models because coarticulation can start as soon as possible provided that there is no antagonism. Beskow [11] compared Cohen & Massaro and Öhman models on the same corpus and found out that Cohen & Massaro model performs better if the evaluation focuses on the deviation between predicted and real data. On the other hand, perception experiments realized with a sample of 25 subjects showed the inverse result. One explanation is that the perceptive weights of labial features are not equivalent and also depend on the phonetic context. For instance, the perceptive consequence of an erroneous lip aperture prediction for /t/ is probably less sensitive than for /p/. This explains that a rule based algorithm performs slightly better because the key points of labial coarticulation are explicitly given, even if the general result is slightly less realistic. One of these algorithms [4] uses target values specific to every phonemes except some, which are not given any value so that labial coarticulation can be determined by neighbour phonemes. Keating [14] proposed a similar approach by defining the degree of freedom for coarticulation in the form of an authorized window. In some sense, these windows correspond to the definition of resistance to coarticulation [15].

2. Our approach

Taking into account these results, we accepted a rule based method where rules are automatically learnt from a corpus. One expected advantage is the possibility of controlling the dynamics of each parameter, and more importantly the control strategy

itself, at the phoneme level.

In a previous work [9] conducted on a corpus of /iCy/ articulated in a carrier sentence, and where C is a non labial consonant, we showed that there exists a wide range of anticipation strategies according to the speaker. Figure 1 shows that some speakers, like Bl, present a strategy close to a time locked model whereas others, like Od, present a strategy closer to the look-ahead model or even to the time expansionist model [16] which stipulates that the gesture duration can be substantially lengthened but not shortened.

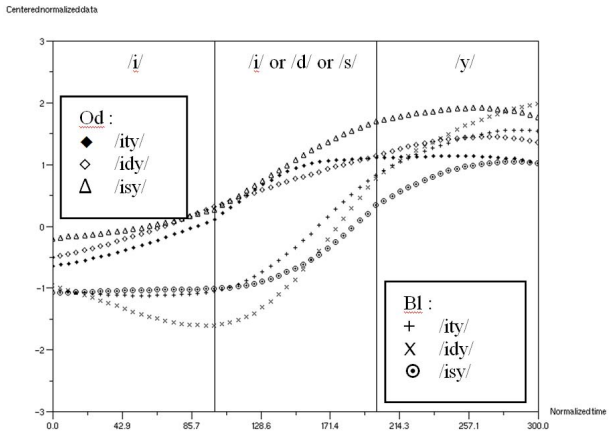

Figure 1: *Comparaison between two coarticulation stategies.*

Given this wide range of coarticulation dynamics we retained sigmoid curves to represent the time evolution of labial parameters (Equation 1). V_i and V_f represent initial and final values of the sigmoid. c monitors the speed transition and t_0 corresponds to the time of the sigmoid center. Indeed, as shown in Figure 2, sigmoid curves can easily accommodate for very different time evolution profiles: in this example, a value of 0.05 for c corresponds to the "look-ahead" model, whereas a value of c greater than 0.10 together with an appropriate choice of t_0 gives rise to a "time-locked" shape. It is also possible to approximate the "expansionist" model. In addition, this kind of curve easily enables the speech rate to be changed and the hyper-articulation to be simulated by modifying Vi and Vt. Note that the curves generated by Cohen & Massaro's model are close to sigmoids.

$$Sig(t) = V_i + \frac{(V_f - V_i)}{1 + e^{-c.(t-t_0)}} \quad (1)$$

We thus built a set a reference sigmoids from the corpus recorded by an experienced talker being used to speak to hard of hearing children, and thus probably easily lip readable. This corpus dedicated to French is made of all the vowels, CV sequences, systematic VCV for a reduced set of consonants and vowels, the most frequent VCCV in French, and 100 phonetically balanced sentences [17]. This corpus was recorded by using a stereovision system [18] that enables the 3D tracking of markers painted onto the speaker's face at the rate of 120 Hz and the recording of the corresponding speech signal. Protrusion, stretching, lip and jaw aperture deformation modes have been determined by calculating the geometrical deviation of well chosen points with respect to their average location (see [18] for further details). The time evolution of these parameters has been slightly smoothed by means of regularizing splines.

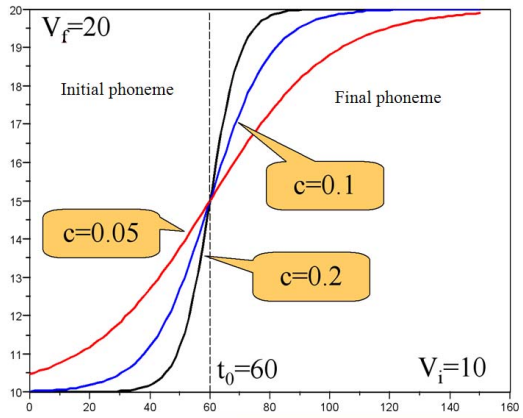

Figure 2: *Influence of parameter c on the shape of sigmoids.*

Only relevant sigmoids, i.e. those corresponding to characteristic dynamics of labial parameters, have been estimated. For instance, unlike /p/ which imposes a very clear time evolution for lip aperture, /k/ does not impose any clear constraint on this parameter, and consequently no corresponding sigmoid was defined. A previous study [9] has been dedicated to the design of a coarticulation algorithm that finds out which are the relevant labial parameters from a formal description and from the quantification of phonemes in terms of labial parameters. This algorithm also exploits interdependencies between labial parameters (for instance lip stretching decreases when protrusion increases).

Once the algorithm has been applied to a sequence of phonemes, a labial parameter may remain indefinite over a sub-sequence because corresponding phonemes are neutral for this parameter. For instance, protrusion is indefinite for the phoneme /k/ of the sequence /iky/ and only one sigmoid, that for the transition /iy/, will be considered. Unlike other algorithms the mutual influence between labial parameters is taken into account.

3. Training and reconstruction process

The whole corpus comprises isolated vowels, systematic CV and VCV for /i,a,u,y/, 70 VCCV or VCCCV, and 100 phonetically balanced sentences. All the logatoms and 70 sentences, i.e. 85% of the corpus recorded, have been used to extract sigmoid parameters as well as minimum, maximum and mean values of labial parameters for every phoneme. Firstly, standard values of opening, stretching and protrusion parameters for vowels articulated by our subject were determined from the isolated vowels recorded. Then, the formal coarticulation algorithm was applied to find out which are the relevant sigmoids that need to be learnt for the CV, VCV, VCCV sequences and 70 phonetically balanced sentences of the training corpus.

The remaining corpus, i.e. 30 phonetically balanced sentences, was used to evaluate quality of the coarticulation synthesized.

The synthesis comprises the following stages. The first consists in segmenting the sentence to be synthesized into consecutive overlapping CV, VCV, VCCV sequences. These sequences are then recovered from the training corpus if they have been recorded, or reconstructed from existing sequences otherwise.

3.1. Completion of missing phonemes

3.1.1. Completion of missing vowels

Since it was not possible to record all the VCV, $VCCV$, $VCCCV$ encountered in French, only /a, i, u, y/ vowels were considered. Other sequences are thus linearly interpolated from /a, i, u, y/ sequences. Weights are the barycentric coordinates of the unknown vowel with respect to /a, i, u, y/ . For example, rebuilding the sequence $/\varepsilon tu/$ requires the calculation of the barycentric coefficients α_i for $/\varepsilon/$ compared to /a, i, u, y/ using data obtained during the recording of the isolated vowels. So, sigmoid parameters are estimated by $\alpha_1.f([atu]) + \alpha_2.f([itu]) + \alpha_3.f([utu]) + \alpha_4.f([ytu])$.

In the case of a $V_1C...CV_2$ sequence where neither of the two vowels belongs to /a, i, u, y/, /$aC..CV_2$/, /$iC..CV_2$/, /$uC..CV_2$/, /$yC..CV_2$/ are first reconstructed by using the technique exposed above. Then , $V_1C...V_2$ is linearly interpolated from /$aC..CV_2$/, /$iC..CV_2$/, /$uC..CV_2$/, /$yC..CV_2$/.

3.1.2. Completion of missing consonants

All the CV and VCV with V in /a, i, u, y/ were recorded, but only 90 $VCCV$ were retained to limit the size of the corpus. They were selected from a statistical analysis of a large French corpus made up to maximize the phonetic coverage of French [19].

VC_1C_2V not in the training corpus are rebuilt by overlapping VC_1V and VC_2V sequences. This technique separates the role of vowels and consonants in agreement with Öhman's Theory [13] which stipulates that the time varying shape of the vocal tract gradually changes from V_1 to V_2, onto which a consonant gesture is superimposed.

The sequence /ikty/, for instance, does not belong to the training corpus, and is thus reconstructed by using /iky/ and /ity/ sequences which, on the contrary, belong to the corpus. Even if protrusion increases from /i/ to /y/ in both sequences, /ity/ and /iky/ present different protrusion profiles. Fig. 3 shows how this VCCV is reconstructed from these two sequences.

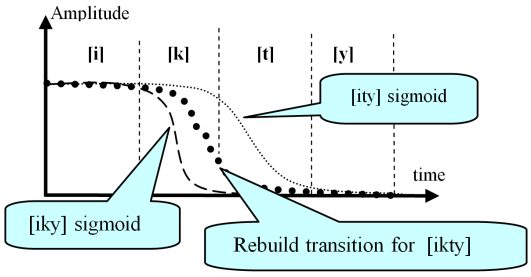

Figure 3: *A VCCV reconstruction of the stretching profile of /ikty/ using consonant completion*

3.2. Time and amplitude sigmoid adaptation

3.2.1. Temporal adaptation

Durations of phonemes recorded in the corpus are not necessarily identical to their target values in the sentence to be synthesized. It is thus necessary to adapt sigmoid durations. For each sigmoid, the relative position of the sigmoid center (t_0) is kept invariant according to the end of the first phoneme and the onset of the last phoneme of the sequence approximated by this sigmoid.

3.2.2. Amplitude adaptation

The reconstruction of labial parameters of a sentence depends on sigmoids concatenated. It is important to guarantee the syntagmatic coherence, i.e. to keep distinctive contrasts between the sounds of the sentence to be synthesized, as well as paradigmatic coherence, i.e. the intrinsic characteristics of the sounds. During the reconstruction process, it is thus necessary to adjust sigmoids to one another. For that purpose, a multidimensional minimization method (Powell type) is applied to determine the optimal shift to apply to each sigmoid so that the deviation to the neighbours is minimal (syntagmatic axis). In parallel, in order to take into account the paradigmatic axis, minimization integrates the deviation of vowel labial parameters with respect to their values for the corresponding isolated vowel. Equation 2 details the expression to be minimized where $Sig_i(t_{min})$ and $Sig_i(t_{max})$ are the initial and final values of the sigmoid number i. The fist part of the equation represents the difference between the final value of one sigmoid and the average of the onset values of the sigmoids immediately following it (the starting phoneme of these sigmoids thus corresponds to the final phoneme of the current sigmoid). It thus corresponds to the syntagmatic axis. The second and third terms of the equation are the differences between sigmoid extremities (corresponding to vowels) and the mean values of isolated vowels. These terms are intended to capture the paradigmatic axis.

We only apply this minimization to phonemes having a certain degree of resistance to the coarticulation, i.e. those presenting characteristic values for labial parameters. α_i are factors weighting the paradigmatic axis versus the syntagmatic axis and are phoneme dependent. α_2 is intended to represent the paradigmatic axis and is given a non zero value only for vowels. After these corrections, a spline curve is generated to approximate parameters trajectories as shown by Figure 4.

$$\begin{aligned}\alpha_1 \sum_{i=1}^{n-1} \left| Sig_i(t_{max}) - \frac{\sum_{j=1/Sig_j \in Next(i)}^{k} Sig_j(t_{min})}{k} \right| \\ +\alpha_2 \sum_{i=1}^{n} \left| Sig_i(t_{min}) - \overline{IsolatedVowel(t_{min})} \right| \\ +\alpha_2 \sum_{i=1}^{n} \left| Sig_i(t_{max}) - \overline{IsolatedVowel(t_{max})} \right|\end{aligned} \quad (2)$$

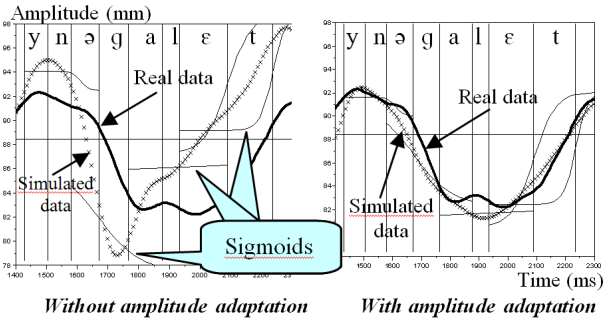

Figure 4: *Example of an amplitude adaptation for protrusion*

4. Concluding remarks

Two measurements are used to provide an objective evaluation of our results: RMSE (root mean squared error) and correlation

between observed and predicted data. RMSE values were calculated as a percentage of the full range of the target parameter. RMSE found on the test corpus is 11.66% and correlation 0.67, i.e. slightly not as good as those found in [11] by comparing several coarticulation methods.

However, a further examination of the results shows that our algorithm, unlike other coarticulation algorithms, primarily focuses on phonetically relevant features. For example, curves presented in [11] show that the jaw opening is almost always fairly underestimated for open vowels. On the contrary, in our case, the first step of the sigmoid construction searches for the relevant labial parameters. This means that the opening of open vowels, or any other labial salient characteristic, will receive a specific attention during the training. This thus guarantees that the characteristic labial parameters are well rendered by our approach. On the other hand, this is achieved to the detriment of the overall fitting between predicted and observed labial profiles. A perceptive experiment will thus be conducted in a near future to evaluate the benefit of preserving the most salient labial characteristics.

The other very strong point of our approach is that syntagmatic and paradigmatic axes are explicitly taken into account through an optimization stage covering the whole sentence to be synthesized. This allows the most salient labial values to be preserved while removing the shifts between sigmoids concatenated, which originate in the discrepancies between phonetic and prosodic contexts encountered in the training corpus.

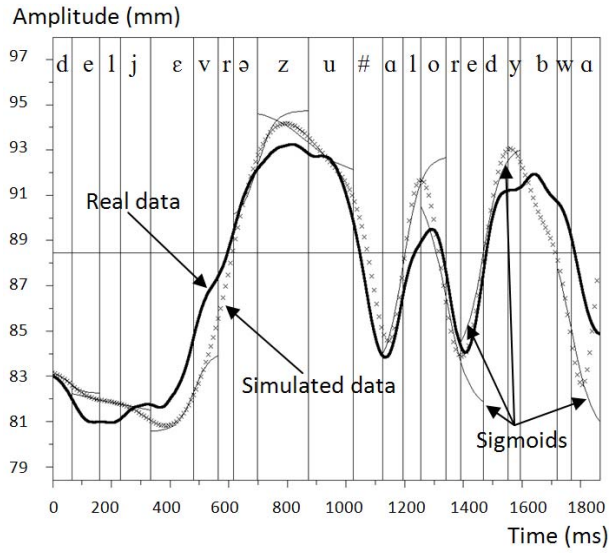

Figure 5: *Comparison between real data and simulated data for protrusion (The sequence is "Des lièvres jouent à l'orée du bois")*

5. References

[1] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of ACM SIGGRAPH*, 1997, pp. 353–360.

[2] N. Brooke and D. Scott, "Two- and three-dimensional audio-visual speech synthesis," in *Auditory Visual Speech Processing*, Terrigal, Australia, 1998, pp. 213–218.

[3] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of ACM SIGGRAPH*, San Antonio, Texas, 2002, pp. 388–398.

[4] J. Beskow, "Rule-based visual speech synthesis," in *European conference on Speech Communication and Technology*, Madrid, Spain, 1995, pp. 299–302.

[5] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," *Models and Techniques in Computer Animation*, pp. 139–156, 1993.

[6] C. Pelachaud, "Visual text-to-speech," in *MPEG-4 Facial Animation - the Standard, implementation en Applications*, I. Pandzic and R. Forchheimer, Eds. John Wiley and Sons, 2002, pp. 125–140.

[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 *Proceedings of the 6th International Conference on spoken Language Processing*, Bejing , China, 2000, pp. 755–788.

[8] S. Maeda and al., "Functional modeling of the face during speech production," *XXIVème Journées d'Etudes sur la Parole*, pp. 341–344, 2002.

[9] V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau, "Inter speaker variability of labial coarticulation with the view of developing a formal coarticulation model for french," *Auditory Visual Speech Processing*, pp. 1021–1024, 2005.

[10] A. Löfqvist, "Speech as audible gestures," in *Speech Production and Speech Modelling*. Hardcastle, W.J. and Marchal, A. (eds). Dordrecht: Kluwer Academic Publishers, 1990, pp. 289–322.

[11] J. Beskow, "Trainable articulatory control models for visual speech synthesis," *International Journal Of Speech Technology 7*, pp. 335–349, 2004.

[12] P. Cosi and al., "Labial coarticulation modeling for realistic facial animation," in *Proceedings of ICMI'02, 4th International Conference on Mutlimodal Interfaces*, Pittsburgh, PA, USA, 2002, pp. 505–510.

[13] S. Öhman, "Numerical model of coarticulation," *The Journal of the Acoustical Society of America*, vol. 39, pp. 310–320, 1967.

[14] P. A. Keating, "The window model of coarticulation," *UCLA Working Papers in Phonetics 69: 3-29*, 1988.

[15] R. A. Bladon and A. Al-Bamerni, "One stage and two-stage temporal patterns of velar coarticulation," *The Journal of the Acoustical Society of America*, vol. 72, 1982.

[16] C. Abry and T. Lallouache, "Le MEM: un modèle d'anticipation paramétrable par locuteur: Données sur l'arrondissement en français," *Bulletin de la communication parlée, 3*, pp. 85–99, 1995.

[17] P. Combescure, "20 listes de dix phrases phonétiquement équilibrées," *Revue d'acoustique 56*, 1981.

[18] B. Wrobel-Dautcourt, M.-O. Berger, B. Potard, Y. Laprie, and S. Ouni, "A low-cost stereovision based system for acquisition of visible articulatory data," in *Proceedings of the 5th Conference on Auditory-Visual Speech Processing, Vancouver Island, BC, Canada*, 2005, submitted.

[19] L. Lamel, J.-L. Gauvain, and M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," in *Proceedings of European Conference on Speech Technology*, Genova, Italy, September, 1991, pp. 505–508.