Inter Speaker variability of labial coarticulation with the view of developing a formal coarticulation model for French

Vincent Robert, Brigitte Wrobel-Dautcourt, Yves Laprie, Anne Bonneau

Inter Speaker variability of labial coarticulation with the view of developing a formal coarticulation model for French

Vincent ROBERT, Brigitte WROBEL-DAUTCOURT, Yves LAPRIE, Anne BONNEAU

Speech Group, LORIA UMR 7503 BP239 –54506 Vandœuvre-lès-Nancy –FRANCE

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

ABSTRACT

Explaining the effects of labial coarticulation is a difficult problem that gave rise to many studies and models. Most of the time small corpora were exploited to design these models. In this paper we describe the realization and exploitation of a corpus with ten speakers. This corpus enables the most invariant labial features (protrusion, stretching and lip opening) to be established. Then we propose a formal prediction algorithm that relies on a standard phonetic description of French phonemes. We conducted a first evaluation of this algorithm that shows its relevancy.

1. INTRODUCTION

Our goal is to study inter speaker variation of labial coarticulation with the view of developing a talking head which would be understandable by lip readers especially deaf persons. In the long term, we want to determine a method of prediction of labial coarticulation adaptable to a virtual speaker. In this first study, we will try to find strategies which should then enable us to determine robust rules to design an algorithm to predict coarticulation.

Predicting labial coarticulation is a difficult problem that gave rise to many studies and models. To predict the anticipatory coarticulation gestures, three main models have been proposed: the look-ahead model (Henke[5], 1967, Öhman[8], 1967), the time-locked model (Bel-Berti and Harris[2], 1979) and the hybrid model (Perkell and Chiang[9], 1986).

These models were often compared on their performances in the case of the prediction of anticipation protrusion in VCV or VCCV sequences where the first vowel is unrounded, the consonants(s) is neutral with respect to labial articulation and the last vowel is rounded. There is no general agreement about the efficiency of these models. More recent models were developed, that of Abry and Lallouache[1] (1995) advocates for the theory of expansion movements: the movement tends to be anticipated when no phonological constraint is imposed on labiality. Löqvist[7] (1990), Cohen and Massaro[4] (1993) proposed dominance functions that require a substantial numerical training.

Let us also emphasize that most of these models derive from the observations of only a few number of speakers only.

We want to develop a more explicative model, i.e. essentially a phonetically based approach that tries to understand how speakers manage to control labial parameters from the sequence of phonemes to articulate.

2. DATA ACQUISITION AND MEASUREMENTS

We have developed a quite robust, accurate and inexpensive approach to track 3D face deformations. Our system (see paper of Wrobel et al. proposed in this conference [10]) uses painted markers and stereovision (Figure 1). Ten French native speakers (5 female and 5 male speakers) were recorded. Our corpus was made up of 5 isolated vowels (/i, y, a, o/), 6 consonants (/p, t, d, s, ʃ, f/) followed by schwa, 8 CV, 20 VCV, 18 VCCV and 2 phonetically balanced sentences. Unlike most of the previous studies we also include consonants with a primary or secondary labial articulation (/p, ʃ/) because we are interested in the general process of labial coarticulation.

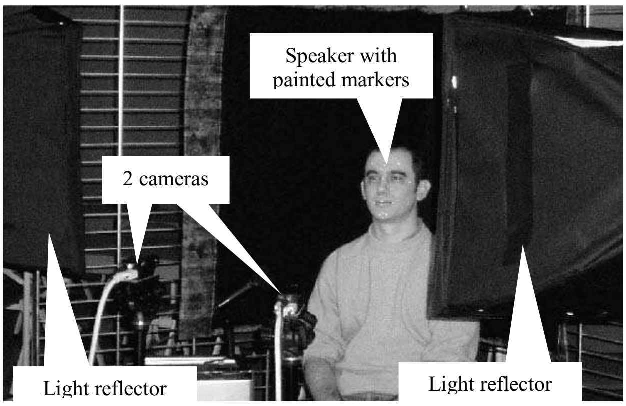

Figure 1. Data acquisition system

A carrier sentence as neutral as possible with respect to labial coarticulation and prosody was chosen.

We extracted three labial parameters: opening, stretching and protrusion. There are several ways of measuring lip protrusion: higher lip only, both lips together, both lips plus lip commissures. We thus conducted a preliminary experiment with a larger number of markers (210 on lips, jaw and cheeks) for one speaker and applied principal component analysis to find the most important factors and markers the most tightly associated to these factors. It turned out that lip commissures are also closely related to the protrusion movement. We thus designed the protrusion measure by taking into account the four markers *A*, *B*, *C* and *D* (see Figure 2). Protrusion is evaluated as follows. Firstly, we determine the average vector normal $\vec{v_n}$ to the plane formed by vectors $\vec{AB}$ and $\vec{CD}$ (on all the frames acquired for one speaker) after the overall head movements have been compensated for. Protrusion is given by the distance between the center of gravity of the four reference markers (*A*, *B*, *C*, *D*) and reference point *F* which is the projection of a fixed point of the head (for instance, a point on the top of the nose) on the normal vector $\vec{v_n}$. In addition to its relevancy with respect to protrusion movements, this measure turns out to give a smaller noise than isolated markers.

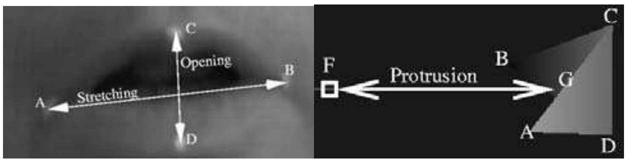

Figure 2. Measuring opening, protrusion and stretching

Despite, the very low level of noise we applied a slight smoothing through regularized splines to allow a relevant computation of protrusion velocity.

We did not use these measurements in millimeters to perform comparisons between speakers because the measures are directly influenced by the anatomic characteristics of speaker. We thus use centered and normalized values. In all the coarticulation figures, the y axis represents $\frac{X-\mu}{\sigma}$ where X is the parameter considered, μ the average value of this parameter over all the speech segments uttered by a subject and σ its standard deviation.

3. ANALYSIS OF THE RESULTS

3.1 Important inter-speaker variability

Results exhibit great inter-speaker differences particularly about the onset of the protrusion movement in /iCy/ utterances (where C is a non labial consonant). For some speakers, protrusion starts very early, sometimes even before the end of the vowel /i/, whereas for others, this event occurs much later. In the same way, most of the time the protrusion maximum is reached at the beginning of the vowel /y/ and sometimes, especially when the speech rate is high, at the end of this phoneme (see Figure 3). Nevertheless, none of the three main models of coarticulation, i.e. Look-ahead, Time-Locked or Hybrid, can account for this variability.

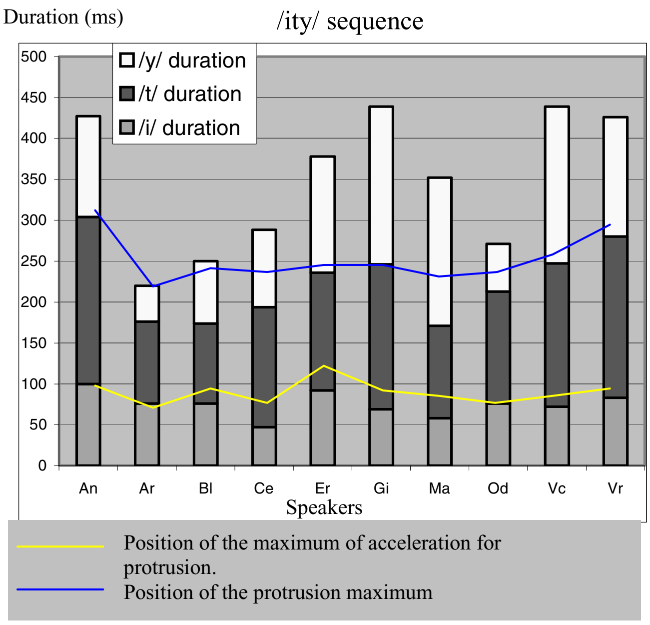

Figure 3. Variation of the protrusion onset (maximum of acceleration) and of the maximum of protrusion in the /ity/ sequence.

3.2 Small intra-speaker variability

In order to investigate intra-speaker variability we examined a series of pairs of utterances that should present similar or very similar coarticulation profiles. We took VCV or VCCV logatoms which only differ on one consonant, for instance /ity/ vs. /idy/.

The comparison of speakers Od. and Bl. for instance shows that Od. presents a strong anticipation in VCV or VCCV sequences unlike Bl. who starts protrusion later and also reaches the maximum of protrusion later (see Figure 4 and Figure 5).

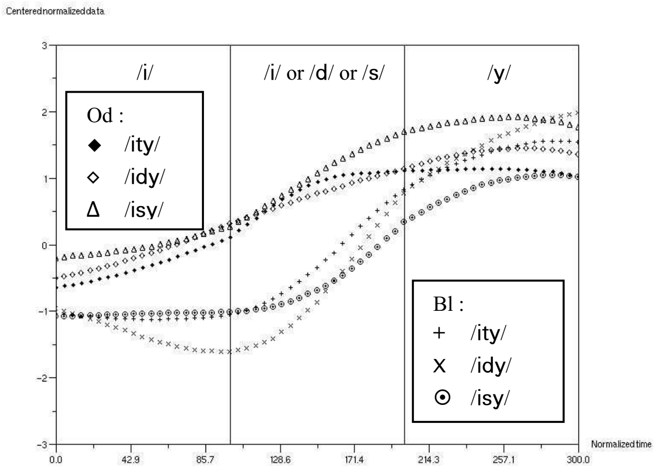

Figure 4. Protrusion variation for speakers Od. And Bl. in VCV sequences.

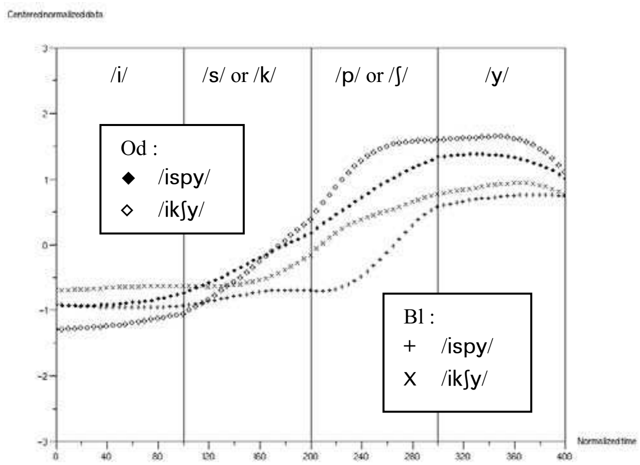

Figure 5. Protrusion variation for speakers Od and Bl in VCCV sequences.

3.3 Influence of the phonetic context

In addition to a strong anticipation, the starting point of coarticulation itself is strongly influenced by the phonetic context. Figure 6 shows lip shapes for /y/ (images were taken between the end the first third and the middle of the vowel). There is a clear influence of the subsequent consonant: /ʃ/ requires a small opening compared to that observed with /p/ to make the frication noise possible, /t/ does not impose any constraint on stretching so the anticipation of stretching of the vowel /i/ begins very early.

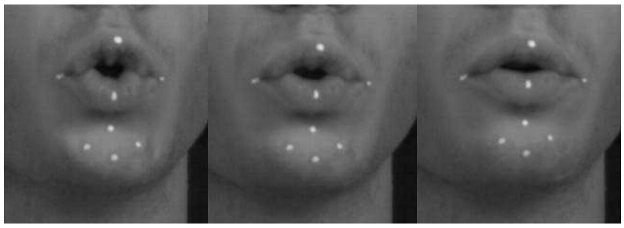

Figure 6. phoneme /y/ in three contexts (/ypi/, /yʃi/ and /yti/)

3.4 Invariants

Even if there are differences between speakers, we found strong common features that must be satisfied to build a generic coarticulation model.

- Anticipatory labial coarticulation occurs for all the speakers whatever the phonetic context. For V_1CCV_2 sequences where V_1 is unrounded and V_2 rounded, the maximum of protrusion is often reached at the beginning of V_2.
- Labial major contrasts between vowels are preserved whatever the context. Stretching, for instance, is always greater for /i/ and protrusion always greater for /y/ (see average values and standard deviations in Figure 7 for the phoneme /i/).
- There is a clear dependence between protrusion, opening and stretching for vowels and some consonants. This remark is the key of our coarticulation algorithm described bellow.

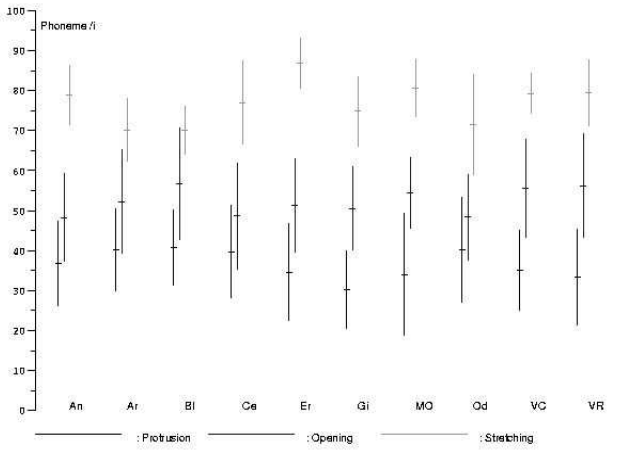

Figure 7. Variability of values for protrusion, opening and stretching for all speakers and for the phoneme /i/

Phoneme	Opening	Stretching	Protrusion
i	O_1	E_4	P_1
a	O_4	E_1	P_1
y	O_1	E_1	P_4
o	O_2	E_1	P_3
p	O_0		
t			
k	$O_{0.5}$		
f			
s			P_3
ʃ			
r			
ɹ			

Table 1. Extract of our phonetic features classification

4. DESIGN OF THE COARTICULATION ALGORITHM

Considering the multiple origins of intra and inter-speaker variability it is hard to imagine that a general model could be derived only by applying a numerical training procedure to a corpus of labial data. Indeed, this would require a huge corpus of data and the optimization procedure used to train the model could render minor labial coarticulation effects faithfully while forgetting more important labial characteristics. Beskow[3], for instance, compared several models of coarticulation and showed that the simulation may produce key parameters far from those expected for some phonemes. Furthermore, the evaluation takes into account numerical proximity between original and synthesized data and not perceptive aspects which are probably as important.

We thus worked from a different point of view which consists in designing an algorithm whose main objective is to respect key phonetic features.

4.1 Labial parameters for each phoneme

We have proposed an estimation of the three labial parameters for each phoneme. Table 1 shows the average degree of stretching, opening, and protrusion (rated from 0 to 4) for each phoneme of our corpus. This description is independent of the phonetic context.

As there is no systematic description of these features in the literature, data collected in this work will provide a good means to establish it by refining partial existing descriptions.

Whereas all vowels have typical values for the three labial parameters, we made the hypothesis that only four classes of consonants strongly influence lip shape: the bilabial consonants /p,b,m/ articulated with closed lips, the labial consonants /f,v/, articulated with slightly opened lips -the lower lip is very close to the upper front teeth- , the fricatives /ʃ,ʒ/ for which protrusion enhances their acoustic specificities, as well as the semi-vowels /j,w,ɥ/. One member of each of the first three classes is present in our corpus /p,f,ʃ/.

The values of protrusion, opening and stretching given in Table 1 are only rough estimations of these parameters for isolated vowels or consonants in neutral contexts. Of course, the values will vary according to the context. It will then be necessary during the application of our algorithm to quantify a degree of "resistance to the coarticulation" for the various phonemes. The first analyses show that this coefficient not only seems to depend on the phoneme, but also of the speaker.

4.2 Main rules

Stretching, opening and protrusion are interdependent parameters. A very strong link exists between protrusion and stretching which vary in opposite directions. A statistical study gave a correlation of -0.98 between these two parameters in our corpus. In addition, particularly for vowels and labial consonants, protrusion and opening vary in opposite directions.

The basic rules are described below (Table 2).

For all the phonemes	**P ↗ ⇔ E ↘**	**P ↘ ⇔ E ↗**
For vowels and labial consonants	P ↗ ⇔ O ↘ E ↗ ⇔ O ↘	P ↘ ⇔ O ↗ E ↘ ⇔ O ↗

Table 2 : Links between Protrusion, Opening and Stretching

The first stage of our algorithm consists in satisfying these rules by propagating values backwards, i.e. from the end of the utterance to the beginning, in order to take into account anticipation.

4.3 Details of the algorithm

Notation :

PR(Ph) : Protrusion of the phoneme Ph
OP(Ph) : Opening of the phoneme Ph
ST(Ph) : Stretching of the phoneme Ph
Ph_1 to Ph_n : Phonemes of the utterance

Initialization

for each phoneme, copy parameters from Table 1 if this parameter is known.

for $Ph_i \leftarrow Ph_1$ to Ph_n do
  if one parameter of Ph_i is unknown
    compute the evolution of this parameter (↗, ➔ or ↘)
  endif
endfor

Body

for $Ph_i \leftarrow Ph_n$ downto Ph_1 do
  if **PR**(Ph_i) > **PR**(Ph_{i-1}) then
    if **ST**(Ph_i) is known and **ST**(Ph_{i-1}) is unknown then
      if **ST**(Ph_{i-1}) = ↗ then *// incompatible with rules*
        ST(Ph_{i-1}) ← **ST**(Ph_i) *// stagnation*
      endif
    end

    if Ph_i = Vowel or Labial Consonant then
      if **OP**(Ph_i) is known and **OP**(Ph_{i-1}) is unknown then
        if **OP**((Ph_{i-1}) = ↗ then *// incompatible*
          OP(Ph_{i-1}) ← **OP**(Ph_i) *// stagnation*
        endif
      endif
    endif
  endif

//… same approach when **PR**(Ph_i) < **PR**(Ph_{i-1}) or
// **ST**(Ph_i) < ST(Ph_{i-1}) or **ST**(Ph_i) > **ST**(Ph_{i-1}) or
// **OP**(Ph_i) < OP(Ph_{i-1}) or **OP**(Ph_i) > **OP**(Ph_{i-1})

4.4 An example

Let us take for example the /ipʃy/ sequence.

Initialization

Phoneme	/i/	/p/	/ʃ/	/y/
O	1	0		1
E	4			1
P	1		3	4

⇒

Phoneme	/i/	/p/	/ʃ/	/y/
O	1	0	↗	1
E	4	↘	↘	1
P	1	↗	3	4

Table 3. Report of the phonetic classification and evolution of parameters

Body

Phoneme	/i/	/p/	/ʃ/	/y/
O	1	0	↗	1
E	4	↘	↘	1
P	1	↗	3	4

⇒

Phoneme	/i/	/p/	/ʃ/	/y/
O	1	0	1	1
E	4	↘	↘	1
P	1	↗	3	4

Not possible because protrusion is increasing

Table 4. First step of the algorithm

Phoneme	/i/	/p/	/ʃ/	/y/
O	1	0	1	1
E	4	↘	↘	1
P	1	↗	3	4

⇒

Phoneme	/i/	/p/	/ʃ/	/y/
O	1	0	1	1
E	4	↘	↘	1
P	1	3	3	4

Not possible because Opening is increasing

Table 5. Second step of the algorithm

At the end of the second step, rules were checked for the entire sequence.

Targets obtained previously give an idea of the evolution of three parameters, but do not set their values precisely. In order to obtain curves of simulation, we applied approximation splines between the targets. We chose approximation rather than interpolation because targets values give only rough information on the parameter.

4.5 Comparison between real data and simulated data

The prediction scheme described above will be refined in the near future to adjust target values with respect to specific speakers. Here, we thus report a preliminary evaluation.

Let us compare real data and simulated data for one speaker (see Figure 8) and for the /ipʃy/ sequence.

First, the overall coarticulation pattern has been correctly predicted. Protrusion, for instance, increases as expected, and more importantly the starting point is correctly predicted even if the real data exhibits a flatter shape for /ʃy/ than the predicted one. In addition, the interaction between /ʃ/ and /y/ is probably slightly more complex than it can be simulated by our algorithm. Indeed the influence of prosody has not been taken into account because few data are available. This remark probably also explains the differences observed for the stretching.

There are probably two origins in the differences observed for lip opening. In addition to the discrepancies due the prediction algorithm, the measurement itself introduces a slight bias. Indeed, the opening is measured through the distance between a marker on the lower lip and another on the higher lip. A side effect of protrusion is to increase this distance although the real opening remains constant.

Furthermore, parameter values given in Table 1 do no incorporate the influence of the phonetic context shown in section 3.3.

We obtained similar results on the set of logatoms studied in this corpus.

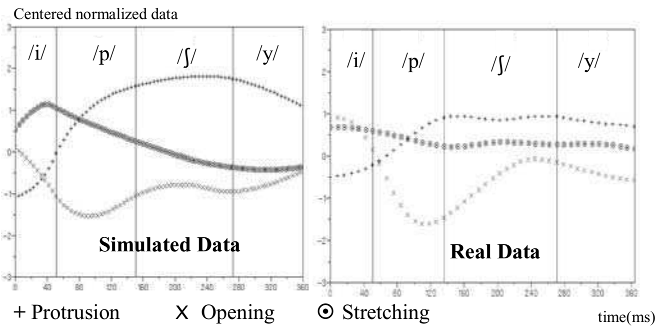

Figure 8. Real and simulated data for an /ipʃy/ sequence for one speaker

5. CONCLUSION

In spite of some differences between the real and simulated data, the general evolution of the parameters of opening, protrusion and stretching is respected. We are now working on two directions of research. Firstly, parameters values given in Table 1 are arbitrary values derived from standard phonetic knowledge. We will train these values from those measured from real data. Secondly, parameters values given by Table 1 should incorporate the influence of the phonetic context. We thus recorded a second larger corpus for one female speaker, which will enable the training of the phonetic influence.

Beside being useful to design a prediction algorithm the corpus used in this study shows that there is a wide inter-speaker variability. In addition to a prior rough evaluation by comparing measured and synthetic data the validation of a coarticulation strategy must thus essentially focus on perceptive tests.

6. REFERENCES

[1]. Abry, C., and Lallouache, T. "Le MEM: un modèle d'anticipation paramétrable par locuteur: Données sur l'arrondissement en français*", Bulletin de la communication parlée, 3, 85-89, 1995.*

[2]. Bell-Berti, F., and Harris, K.S. " A temporal model of speech production", *Phonetica, 38, 9-20, 1981.*

[3]. Beskow, J. "Trainable Articulatory Control Models for Visual Speech Synthesis", *International Journal of Speech Technology, 7, 335-349, 2004.*

[4]. Cohen, M.M., and Massaro, D.W. "Modeling coarticulation in synthetic visual speech", i*n N. Magnenat-Thalmann and D.Thalmann (Eds*). Models and Techniques in Computer animation. *Springer Verlag, Tokyo, 139-156.*

[5]. Henke, W. "Dynamic articulatory model of speech production using computer simulation", *Ph D dissertation*, MIT, 1966.

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

[7]. Löfqvist, A. "Speech as audible gestures", *In Hardcastle, W.J. and Marchal, A. (Eds),* Speech Production and Speech Modelling. *Dordrecht, Kluwer Academic Publishers, 289-322.*

[8]. Öhman, E.G. "Numerical model of coarticulation", *The Journal of the Acoustical Society of America, 39, 310-320, 1967.*

[9]. Perkell, J.S., and Chiang, C.M., "Preliminary support for a hybrid model of anticipatory coarticulation", *Proceeding of the XIIth International Congress of Acoustic, 1986.*

[10]. Wrobel-Dautcourt B., Berger, M.O., Potard, B., Laprie, and Y., Ouni, S., "A low-cost stereovision based system for acquisition of visible articulatory data", *Audio-Visual Speech Processing, Vancouver Island, BC, Canada, 2005.*