Labial Anticipation Behavior during Speech with and without Cued Speech

Marie-Agnès Cathiard, Virginie Attina and Delphine Alloatti
Institut de la Communication Parlée UMR CNRS No 5009 - Université Stendhal –
BP 25 38040 Grenoble Cedex 9
E-mail: cathiard@icp.inpg.fr

ABSTRACT

What about the labial anticipation phenomenon when the mouth is coordinated with hand gestures as it is the case in Cued Speech? We tested the rounding anticipation in French [iC$_n$y] sequences (C$_n$ = 0 to 6 consonants) realized with and without Cued Speech. We observed that: (i) anticipation increased with the consonantal interval with an expansion coefficient coherent to the Movement Expansion Model [6]; (ii) there was no difference in anticipation behavior with and without Cued Speech. So the hand action fits well into the mouth action. They complement each other in a pretty well coordinated fashion.

1. INTRODUCTION

It is well known that anticipation is a fundamental component of the coarticulated structure of speech. Our question in this paper is : What about this anticipation phenomenon when the mouth is coordinated with hand cued gestures ? We specifically question the preservation of the lip anticipation behavior during normal speech and speech with Cued Speech (CS), a manual code in complement of lipreading for deaf perceivers ([1]). When speaking, the Cued Speech talker places her/his hand near the face, the palm toward her/him so that the speechreader can see the back of the hand simultaneously with the lips. Placement of the hand around the face codes vowels, whereas handshapes or finger configurations distinguish among consonants. Thus, with the vision of handshape and handplacement, combined with the information visible on the lips, the deaf can identify a unique consonant-vowel syllable.

Several models were proposed to account for the anticipation of rounded vowels which is regularly present in the preceding consonants as in a sequence [iC$_n$u] or [iC$_n$y] (with C$_n$ = 0 to 6 consonants) : see [2] and [3] for a review of coarticulation models. We just refer to the "Movement Expansion Model" or M.E.M. [4, 5, 6], which has been tested for French, the language under study here. In this model, the global rounding gesture, measured on the superior lip evolution, is taken into account and does not depend on the acoustical offset or onset of vowels. In fact, there is a constant protrusion movement time which corresponds to a basic [iy] gesture (speaker-specific, around 140 ms). For [iC$_n$y] transitions, movement begins to expand with a one-consonant obstruence interval (OI, at about 100ms). Then it increases linearly depending on the effective duration (OI) of the string of consonants C$_n$. The slope of this law of expansion is speaker-dependent. So this model allows a parameterization by talker.

Complementing the protrusion aspect (protrusion M.E.M.), the MEM is the unique model which can also account for lip constriction data. The constriction M.E.M. ([6, 7]) is based on 2 events: (i) the duration of the time falling (TF) corresponding to the onset of the rounding gesture determined on a [iC$_n$y] transition from a 90% area onset value (reflecting the onset of the constriction towards [y]) down to 10% area onset value; and (ii) the duration of the hold phase determined from the 10% area onset value to 10% area offset value (Fig.1).

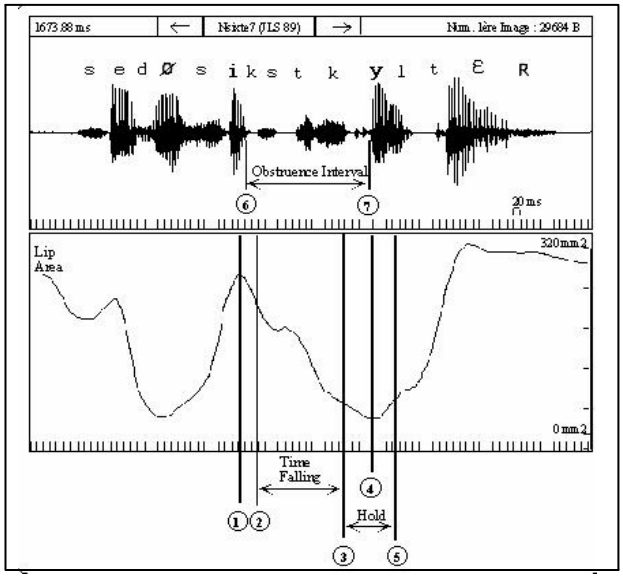

Figure 1: Acoustic signal (top) and temporal evolution of lip area (bottom) of the [sedøsikstkyltɛR] sequence [7] with the phases of « time-falling » and « hold ». Event (1) corresponds to [i] maximum 100% area, (2) is the 90% of lip area onset, (3) is the 10% of lip area onset (from the [i] maximum), (4) corresponds to the [y] target, (5) is the 10% of the lip area offset, (6) and (7) are the onset and offset of the obstruence interval.

As in the Protrusion M.E.M., the expansion of the constriction movement starts about 100 ms for a

ISBN 1-876346-48-5

one-consonant [iCy] obstruence interval and the movement expansion coefficient is also specific to each talker.

In our experiment, we only used the lip area parameter and we will compare our data to this constriction model.

2. METHOD

2.1. Corpus

We explored the lip constriction for French [iC$_n$y] sequences (from 0 up to 4 consonants) in sentences partly with nonsense words as : « Ces deux Scies utèrent » [sedøsiytɛR] (“These two saws [nonsense verb]?”) without consonant between vowels or as « Ces deux sixes scutèrent » [sedøsiksskytɛR] with 4 consonants, produced by a female talker without and with Cued Speech (see Fig.2 for an illustration of a Cued Speech sequence). So we obtained [iy], [iky], [isky], [iksky] and [ikssky] transitions.

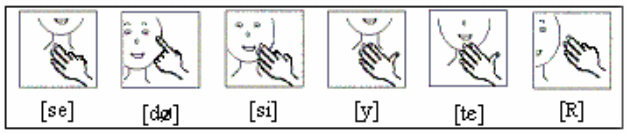

Figure 2 : [sedøsiytɛR] Cued Speech sequence

2.2. Talker

The Cued Speech talker is a 37 year-old French female. She was easy to lipread and had a good hand visibility and fluency during coding. She has been using Cued Speech at home with her hearing-impaired child for 9 years. She graduated in French Cued Speech in 1996 and regularly translates into CS code at school.

2.3. Audiovisual recording

The French Cued Speech talker was audiovisually recorded in a sound-proof booth, at 50 frames/second with a first camera in large focus for the hand and the face, and a second one in zoom mode dedicated to the lips and synchronized with the first one. The talker wore opaque goggles in order to protect her eyes against the halogen floodlight; moreover, a blue mark was placed on the left goggle, as a reference point for the lips and head measurements. Finally, her head was maintained fixed with a helmet to avoid movement. The talker's lips were made-up in blue to process automatically the lip contours. Colored marks were placed on the back of the hand to follow the displacement of the hand around the face (Fig.3).

The 2 cameras were connected to 2 different BetaCam video-tapes. At the beginning of the recording session a push button was activated thus switching on the set of LEDs (placed in the field of the two cameras) during the first A-frame instant of the video image. Thus correspondence between time-codes of the two cameras could be calculated. The audio signal was synchronically digitalised with the video image.

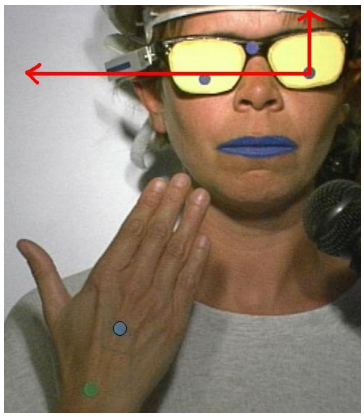

Figure 3: Image of the CS speaker with axes in superimposition used for landmarks localization.

The sentences were recorded with and without Cued Speech in two consecutive sessions. After elimination of the erroneous realizations, we disposed of 25 sentences with Cued Speech (9 for [i#y], 4 for [iky], 3 for [isky], 7 for [iksky], 2 for [ikssky]) and 15 without Cued Speech (respectively, 5, 3, 2, 3 and 2). For the [iy] sequences, we have asked the talker to vary the pause duration.

2.4.Data processing

The audio signal was digitalized and the between-lips area parameter was obtained by image processing [8, 9] every 20 ms. In synchrony with audio signal and lip area parameter, the x and y coordinates of the hand mark was extracted.

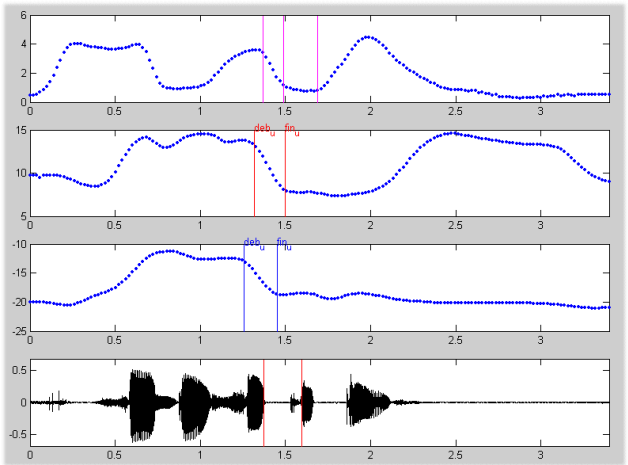

Figure 4: Example of a [sikytɛR] sequence. From top to bottom : 1. time trajectory of lip area S (cm^2); 2. & 3. x (cm) and y (cm) trajectories of the hand (50 Hz); 4. acoustic signal. On each signal, landmarks used for the analysis are visible (see text).

Different events were localized (see Figure 4 for an example from a [sikytɛR] sequence). (i) The audio signal was labeled, in particular the acoustic offset of the unrounded vowel [i] and the acoustic onset of the rounded vowel [y] (the time interval between these two events giving the obstruence or pausing interval OI). (ii) On the lip area evolution, the onset of lip rounding of the [y] vowel

ISBN 1-876346-48-5

(revealed by acceleration peak: called L1) and the onset of the rounding plateau (deceleration peak: L2) were labeled to determine the *time-falling rounding*; then the offset of the rounding plateau was labeled : the duration between the plateau onset and the plateau offset giving the extent of the *rounding plateau*. (iii) A similar labeling was conducted on the manual gesture : we labeled the onset (M1) and the offset (M2) of the manual gesture for the [y] vowel (from acceleration and deceleration peaks).

3. RESULTS

In this section we only present the results concerning the labial anticipation measured from three different cues. Work in progress is currently elaborating the relationships between lip and hand in a general coproduction framework.

3.1. Time-Falling phase

We calculated the Time-Falling phase (TF) duration in relation to the obstruence interval (OI) for the sequences with and without Cued speech (Fig.5). The regression analysis showed the TF and OI values are correlated for both data ($p<0.001$) with a slope of .19 for sequences with CS et .18 for sequences without CS. The comparison between the slopes of the correlation lines indicated that they were not different ($t=0.0655$ with $\nu=38$ degrees of freedom, the limit value being $t_{0.01}=2.71$).

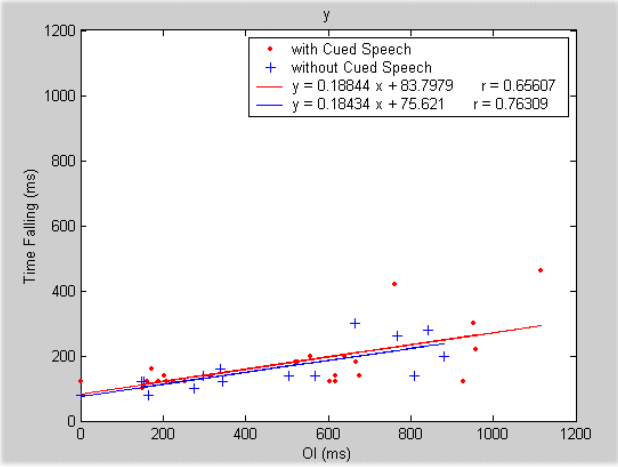

Figure 5 : Correlation between the obstruence interval and the time-falling phase duration for sequences with and without Cued Speech.

3.2.Global rounding gesture

We also calculated the global rounding duration (corresponding to the Time-Falling phase plus the Hold phase: TF+H) in relation to the obstruence interval (OI) for the sequences with and without Cued speech (Fig.6). The regression analysis showed the TF and OI values are correlated for both data ($p<0.001$) with a slope of .25 for sequences with CS et .21 for sequences without CS. The comparison between the slopes of the correlation lines indicated that they were not different ($t=0.5848$ with $\nu=38$ degrees of freedom, the limit value being $t_{0.01}=2.71$).

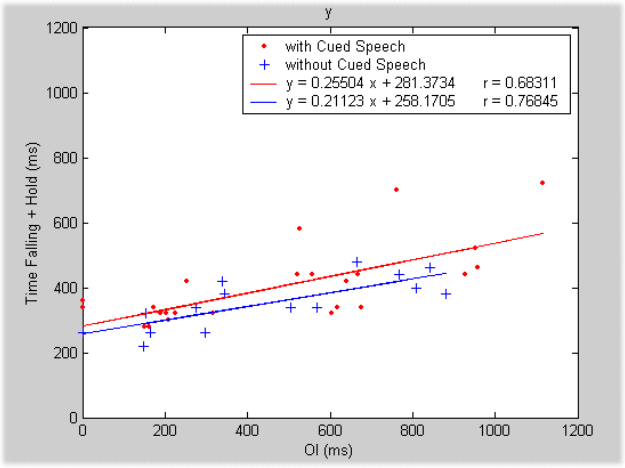

Figure 5 : Correlation between the obstruence interval and the TF+Hold phase duration for sequences with and without Cued Speech.

So we can conclude that labial anticipation increased with the obstruence interval, with expansion coefficients comparable to Abry et al. data ([6] Fig. 3) who obtained a slope of .15 (for the Time Falling phase of [i#y] sequences with pause # varying between 100 to 650 ms). On our data as on their data, we can observe a greater variation of TF and TF+Hold values when OI is long, revealing some different strategies probably in relation with prosodic realization of the string of consonants (when the string is long, pause insertion is more probable).

3.3. Global labial anticipation in relation to [y] acoustic onset.

We have also calculated for each sequence the duration between the onset of the labial constriction gesture (L1) and the acoustical onset of the [y] vowel. We obtained a mean labial anticipation (LA) for all sequences with Cued Speech of 257 ms before the [y] acoustical onset and of 212 ms for all sequences without Cued Speech. A Student test indicated no difference ($p<.01$).

From this set of results (on TF, TF+H and LA), we can conclude that our talker adopted a comparable labial anticipation behavior during speech with and without Cued Speech. So we now mix data with and without CS sequences.

3.4. Labial anticipation in relation to the pausing interval or the obstruence interval.

We separated the [iy] sequences realized with short pause instruction ([i#y] (with a pausing interval inferior to 255 ms) and long pause instruction ([i#:y] (with a pausing interval superior to 500 ms) and we calculated the labial anticipation value (LA) for each pause. We also calculated the LA value separately for sequences with 1, 2, 3 et 4 consonants (Table 1). We tested the significant differences between the different values of OI and constituted 3 classes : vocalic transitions with short pause or one-consonant, with

ISBN 1-876346-48-5

two-consonants and long pause and with 3 or 4 consonants. We tested in the same way the LA values and 4 classes emerged since LA was statistically different for 3 and 4 consonants.

	[i#y]	[iky]	[isky]	[i#:y]	[iksky]	[ikssky]
OI	142.1	175.96	509.4	535	729	845.6
LA	-131.8	-173.76	-232.1	-238	-291	-461.1

Table 1: Labial anticipation values in relation to the interval obstruence (or pausing interval, both called OI) values. The values in a same cell are not significantly different (t test, $p<.01$).

In a preceding experiment [10], we tested the constriction labial anticipation on [i#y] transitions, inserted in a carrier sentence : "T'as mis : UHI ise?" [tami#yiiz], with two pausing instructions. Sentences were audiovisually recorded by the same CS talker with all sentences realized with CS. The mean pausing interval obtained for the sequences realized with the short pausing instruction was 387 ms [#] and 1075 ms [#:] for long pausing instruction. The LA values, also measured from the labial constriction onset to the [y] acoustic onset, were respectively 280 ms and 419 ms.

If we now considered all values of the two experiments, we observed that : the more OI increased, the more AL is precocious. However, the speaker doesn't use all the time available to anticipate, particularly when the string of consonants is superior to one-consonant.

4. CONCLUSION

We observed that: (i) labial anticipation increased with the consonantal interval with an expansion coefficient similar to Abry et al.'s data [6]; (ii) there was no difference in anticipation behavior with and without cued speech. Recall that Cued Speech system is grounded on the ultimate CV syllabification of speech. The syllable string [C_ny] as we used in this experiment is necessary decomposed in CVs, with each consonant coded by the appropriate finger handshape *with the hand in side position*. It could have been possible that this successive coding of each consonant, not related with the specific [y] handposition, would block the natural labial anticipation time course. In this case, one could have observed labial anticipation only for the syllable containing the [y] vowel, i.e. the last consonant of the string. In fact, we observed no perturbation of the pattern of speech action necessary for the acoustic goal. On the contrary, the hand action fits well into the mouth action. They complement each other in a pretty well coordinated fashion.

Aknowledgments : This work was supported by a "programme Cognitique" (ACT1b) on Anticipation of the French Research Ministry, a "programme Cognitique" on Cued Speech of the French Research Ministry and a BDI grant from CNRS.

REFERENCES

[1] R.O. Cornett, "Cued Speech", *American Annals of the Deaf*, 112: 3-13, 1967.

[2] E. Farnetani and D. Recasens "Coarticulation models in recent speech production theories", in: *Coarticulation: theory, data and techniques*, W.J. Hardcastle and N. Hewlett (Eds.), Cambridge University Press, 31-65, 1999.

[3] E. Farnetani, "Labial coarticulation", in: *Coarticulation: theory, data and techniques*, W.J. Hardcastle and N. Hewlett (Eds.), Cambridge University Press, 144-163, 1999.

[4] C. Abry and M.-T. Lallouache, "Le M.E.M.: un modèle d'anticipation paramétrable par locuteur. Données sur l'arrondissement en français", *Bulletin de la Communication Parlée*, 3: 85-99, 1995.

[5] C. Abry, M.-T. Lallouache and M.-A Cathiard, "How can coarticulation models account for speech sensitivity to audio-visual desynchronization?", in D.G. Stork & M.E. Hennecke (Eds), *Speechreading by humans and machines: Models, Systems and applications*, NATO ASI Series F, 150: 211-219, 1996.

[6] C. Abry, M.-A. Cathiard, R. El Abed, M.-T. Lallouache, M.-C. Leroy, P. Perrier, F. Poveda and C. Savariaux, "Silent speech production: Anticipatory behaviour for 2 out of the 3 main vowel gestures/features while pausing", In *Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling & 4th Speech Production Seminar: Models and Data*, Autrans, France, 101-104, 1996.

[7] C. Abry, and M.-T. Lallouache, "Modeling lip constriction anticipatory behaviour for rounding in French with the MEM (Movement Expansion Model)", *Proceedings of the 13th International Congress of Phonetic Sciences*, 4: Stockholm, Sweden, 152-155, 1995.

[8] M.-T. Lallouache, "*Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres*", PhD. Thesis, Grenoble, Institut National Polytechnique, 1991.

[9] M. Audouy, "*Logiciel de traitement d'images video pour la determination de mouvements des lèvres*", Projet de fin d'études, option Génie Logiciel, ENSIMA, Grenoble, 2000.

[10] V. Attina, D. Beautemps and M.-A. Cathiard, " Attina, V., Cathiard, M.-A. & Beautemps, D., "Controling anticipatory behavior for rounding in french cued speech", International Conference on Spoken Language Processing (ICSLP 02), pp. 1949-1952, Denver, 2002.