VISUAL PERCEPTION OF ANTICIPATORY ROUNDING DURING ACOUSTIC PAUSES : A CROSS-LANGUAGE STUDY

M.-A. Cathiard *, G. Tiberghien *, A. Cirot-Tseva **,
M.-T. Lallouache **, P. Escudier **

* Laboratoire de Psychologie Expérimentale, CNRS URA 665
** Institut de la Communication Parlée, CNRS URA 368
Grenoble, France

ABSTRACT

This paper deals with visual perception of anticipatory rounding in French vowel-to-vowel gestures during acoustic pauses. Visual identification was studied for French and Greek subjects. Our results show that : (i) rounding anticipation can be identified *only by eye* several centiseconds before any perceivable sound; (ii) when the pause tripled, visual anticipation doubled, i.e. temporal *positions* of phonemic visual boundaries were dependent upon the extent of articulatory anticipation; (iii) but the boundaries *steepness* (switching time) was not; (iiii) the comparison between French and Greek subjects did not revealed significant differences in rounding anticipation capture.

1.INTRODUCTION

Several studies in speech production have investigated anticipatory vowel rounding (of which, [1] is the most outstanding for French), particularly through consonant clusters, in order to investigate a major motoric issue, serial ordering.

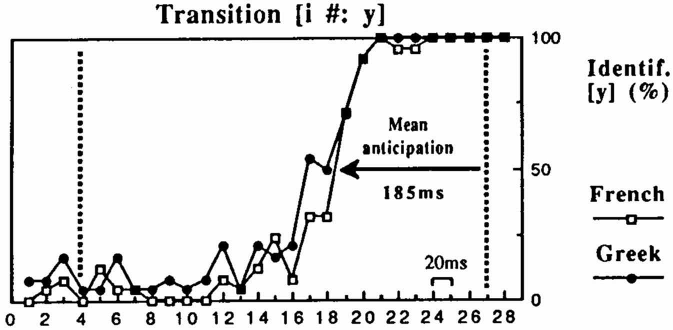
As an expert in visual speech perception, McGurk mentioned briefly an unpublished experiment [5], with a reaction-time paradigm : it would appear to demonstrate that this anticipatory gesture can be detected visually to identify CV syllables from lip movements, *prior to their being perceived auditorily.* More recently [2] found, for French [zizy] syllables, that the anticipation of the rounding gesture was perceived visually by the subjects who were able to identify the [y] vowel before the end of the [i], whereas it was not detected auditorily as early.

We studied, for French stimuli, visual perception of such an anticipation in vowel-to-vowel gestures *without intermediate consonants,* using natural productions of *acoustically silent pauses* between the vowels. Such pauses have, of course, a prosodic signalling function. So it is not the prosodic stream which is *acoustically* (if not visually) interrupted, but segmental information, here rounding. Consequently the general issue to be tackled is : *can this segmental flow be tracked from the optic signal only, when the acoustics are disrupted?*

In this paper, two specific questions are focused on : (i) is there visual information capture of the second vowel stimulus, prior to its acoustic onset, and, if so, how long before?; (ii) is there a shift in the visual boundary for speakers of Greek – who do not have the [y] vowel in their phonological inventory – by comparison with native French subjects?

For lack of models strictly dedicated to the audio-visual perception of speech *anticipation* (in spite of [6]), we will use here the predictions of three current articulatory models [7] and transpose them to the visual level, in order to evaluate which processing the "eyes" perform on speaker's labial gestures :
(i) the *look-ahead* model [LA] predicts a maximal anticipatory span, i. e. as soon as the rounding movement is possible; (ii) for the *time-locked* model [TL], movement onset occurs at a fixed time before the acoustic onset of the rounded vowel; (iii) the *two-stage* or *hybrid* model [H] allows to describe lip protrusion gestures with two components, a gradual initial phase, which begins as soon as possible in a *look-ahead* fashion, and a more rapid second phase (its onset is a peak in acceleration), which is *time-locked.*

Fig.1



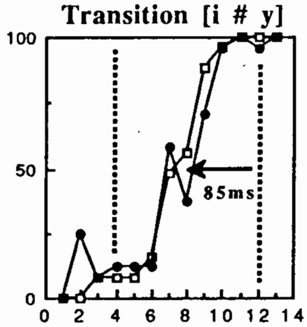
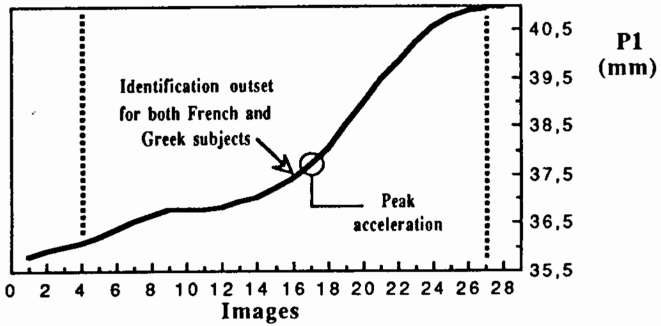


Fig.2

Figures 1-4. – Above : identification functions of [i -> y] transitions for 25 French and 24 Greek subjects.
– Below : corresponding protrusion gesture for the upper lip (P1).
The left dotted line indicates the acoustic offset of the [i] and the right one the acoustic onset of the [y].

Fig.3



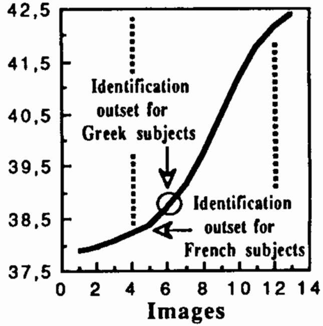


Fig.4

We will try to test these models by analysing *articulatory and visual* data.

2.METHOD

2.1.Corpus

We used [i # y] transitions which were embedded in a carrier sentence : "tu dis : UHI ise?" [t y d i # y i i: z], "you say : ...", where UHI is, by convention, an "Indian name" and "ise" a third person present of a nonsense verb "iser". [t y d i # i i i: z] is the control stimulus with IHI as "Indian". Each transition had to be produced following two different pausing instructions, a short [#] and a long one [#:]. Each sentence was repeated 10 times thus giving 40 utterances which were recorded in random order.

2.2.Video recording

A French male talker was filmed, at 50 frames/second, with simultaneous face and profile views, in a sound-proof booth. Talker's lips were made-up in blue : a Chroma-key was connected to the output of the front camera so that the blue color was transformed to saturated black in real time in order to realize a maximal outlines detection of the lip slit. The subject wore black sunlight goggles in order to protect his eyes against the 1000 W halogen floodlight; a slide rule was fixed on the right side of the goggles to ensure adequate profile articulatory measurements [4].

2.3.Selection of visual stimuli

2.3.1.Acoustic measurements

Four utterances were selected among 40 after duration measurements of all intervocalic pauses. They were chosen as representative of mean durations for the short pause (# = 160 ms) and the long one (#: = 460 ms).

2.3.2.Articulatory processing

For each digitized frame (512 x 512 pixels), eight articulatory parameters, describing front slit and lateral protrusion characteristics, were automatically extracted by image processing [4] and kinematics (velocity and acceleration) were obtained by a cubic spline smoothing of position functions. Examination of traces of upper lip protrusion (P1) vs. time (one of the usually available parameter in others studies), for [i # y] and [i #: y] trajectories, revealed movements profiles with two components, i.e. *hybrid profiles.* Nevertheless (as in [7]), peak acceleration was not time-locked, occuring about 120 ms before the acoustic onset of the [y] in [i # y] versus 200 ms in [i #: y]. Movement onset was neither time-locked (as in [7]), since it occured 260 ms before the acoustic onset of the [y] in [i # y] versus 560 ms in [i #: y] (i.e. the protrusion gesture began 100 ms into the [i] vowel). Finally our articulatory stimuli correspond better to a LA model, with respect to *dates* of onsets, but they display rather H *profiles* (fig. 3 & 4).

2.4.Test procedure

We selected 13 images for short transitions and 28 images for the long ones, with 3 images before pause onset and 1 after pause termination. We thus obtained a total of 82 stimuli which were presented in random order, with a shift of 5 images between each subject. At the beginning of the test, 4 extra images were proposed to familiarize subjects with the task. The stimuli were displayed individually to each subject on a high resolution computer screen. The task was to decide whether the speaker was uttering [i] or [y]. Subjects were encouraged to answer rapidly (within a few seconds) via a computer mouse.

2.5.Subjects

25 French and 24 Greek normal-hearing native speakers served as naive subjects (their hearing and vision acuities were checked). A good auditory identification of the [i] vs. [y] contrast was confirmed for all Greek subjects (mean score : 93.5%).

3.RESULTS

The identification functions – traced from [y] percent responses for each image – have a classical S-shape (fig. 1 & 2). Of course control transitions displayed steady state profiles, since [i –> i] images were generally identified as [i] (above 80%). Subjects were able to identify correctly (at 100%) "targets" of the presented vowels, i.e. images corresponding to the non silent outsets of [y]. Moreover, they were clearly able to capture anticipated segmental information on rounding (95% correct) *up to 120 ms before the acoustic onset of the vowel, be they French or Greek.*

3.1.Differences and similarities in visual boundaries

A quantitative comparison between identification functions was achieved by Probit Analysis [3]. First, this method allowed us to *date* the position of visual

boundaries (corresponding to 50% [y] responses) with regard to the acoustic onsets, and to test the significance of time differences. In addition, it allowed us to test the parallelism between functions, thus delivering information on the possible similarity in *steepness* between the boundaries.

For [i # y] : boundaries took place 90 ms before the acoustic onset of [y] for French subjects, and 80 ms for Greek.

For [i #: y] : boundaries anticipated of 180 ms, for French, and 190 ms, for Greek.

There was a reliable difference (at p<0.01) between the two conditions [i # y] and [i #: y], within each language group : i.e. *when the pause tripled, visual anticipation doubled.* But while the temporal *positions* of phonemic visual boundaries were dependent upon the *extent* of anticipation in protrusion, on the other hand, the temporal *accuracy* of these boundaries (i.e. their *steepness* estimated by functions gradients) did not depend on anticipation : *80 to 110 ms were sufficient to switch from [i] to [y] in all cases.*

On these two points, there were no significant differences (p<0.01) between French and Greek subjects. Notice that the Greek had a rather fair competence in *auditory identification* of [i] vs. [y] (but their [y] *productions* were usually biased toward [i]). The other way round they could have read the "U" choice as [u]. In both cases ([y] or [u]) however, they did not capture significantly less *rounding* anticipation than French did.

3.2.Visual perception of anticipation and articulatory models.

The observed significant shifts in boundaries could by themselves discredit the prediction of a time-locked visual anticipation. In fact, our perceptual as our articulatory (cf. 2.3.2.) data allow us to reject strong versions of both TL and H models : neither *onsets* nor *peak accelerations* are time-locked on our temporal functions. What about the LA model? It can be rejected on the basis of our *visual data only* : while the anticipatory gesture begins as early as possible, the subjects ignore *visually* this change until it is clearly *accelerated* (fig. 3 & 4). More precisely, it is the position of the visual identification *outset* (detected as the first peak of the second derivative of the smoothed function) which reveals itself synchronous with the *acceleration peak* of the protrusion gesture (with a limit discrepancy of 1 image [20 ms] between these two events).

4.CONCLUSION

Rounding anticipation in vowel production has proved to be reliably identifiable *only by eye* several centiseconds before any perceivable sound (up to 120 ms). These results are at least valuable for stopped images. They need additional research on movement processing in speech (especially for acceleration detection) and further elaboration of appropriate models : neither LA, TL nor H.

The cross-language comparison did not revealed significant discrepancies in visuo-temporal boundaries, whether the rounding dimension was bound to the front/back contrast, as in Greek, or whether it was free, as in French [i] vs. [y]. Whether this result argues for a universal lipreading skill, remains of course an open quest.

* Many thanks to J.-L. Schwartz and W. Serniclaes for their advices in Probit Analysis and to T. Brennen for improving our English.

5.REFERENCES

[1] BENGUÉREL, A.P. & COWAN, H.A. (1974), "Coarticulation of upper lip protrusion in French", *Phonetica, 30*, 41-55.

[2] ESCUDIER, P., BENOIT, C. & LALLOUACHE M.-T. (1990), "Visual perception of anticipatory rounding gestures", *JASA, Suppl.* 1, *87*, S126.

[3] FINNEY, D.J. (1971), *Probit analysis,* Cambridge University Press.

[4] LALLOUACHE, M.-T. (1990), "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux", *Actes des XVIIIèmes J.E.P. (28-31 Mai)*, Montréal, Canada, pp. 282-291.

[5] McGURK, H. (1981), "Listening with eye and ear" (paper discussion), in T. Myers, J. Laver & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 394-397), Amsterdam, North-Holland.

[6] MASSARO D.W. (1987), *Speech perception by ear and eye : a paradigm for psychological inquiry,* Lawrence Erlbaum Associates.

[7] PERKELL, J.S. (1990), "Testing theories of speech production : implications of some detailed analyses of variable articulatory data", in W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, (pp. 263-288), Kluwer Academic Publishers, Dordrecht, Boston, London.