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

M.-A. Cathiard *, G. Tiberghien *, A. Cirot-Tseva **,
M.-l'. 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 identification of anticipatory 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; (iv) 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As an expert in visual speech perception, McGurk mentioned briefly an unpublished experiment [5], with a reaction-time paradigm: it would appear to demonstrate that this anticipatory gesture can be detected visually to identify CV syllables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 [zy] 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.
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].
2.1. Corpus

We collected data with [i] and [y] transitions which were embedded in a carrier sentence: "au dis: UHI lse?" [y i # i i y i z], "you say, ... " where UHI is, by convention, an "Indian name" and "ise" a third person present of a nonsense verb "ise". [y i # i i i z: is the control stimulus with IH as "Indian". Each transition had to be produced following two different pausing instructions, a short [i] and a long one [#: i]. Each sentence was repeated 10 times thus giving 40 utterances which were ordered 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 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 light of the computer screen. The task was to decide to which category (LA, TL, or H) a good auditory identification of [i]: 2], "you dis: tu dis: [?]", peak acceleration functions. What about the LA, TL nor H.

4. References