CONFIGURATIONAL vs. TEMPORAL COHERENCE IN AUDIOVISUAL SPEECH PERCEPTION

M.-A. Cathiard*, M.T. Lallouache*, T. Mohamadi*, and C. Abry*
*Laboratoire de Psychologie Expérimentale, Université Mèdès-France, BP 47 F-38 040
Grenoble Cedex 9 /ICP URA CNRS n° 368 INPG/ENSERG Université Stendhal BP 25

ABSTRACT
This contribution evaluates the natural coherence of the building up of audiovisual information in the flow of speech that is provided on the face. Our conclusion is that the bimodal coherence of speech is above all configurational, and not essentially temporal, as it is the case in another intersensory system, such as auditory-visual localization in the ventrolateral effect [1, 2].

1. INTRODUCTION
The aim of this contribution is to evaluate the natural coherence of the building up of visual and audio information in the flow of speech, using a gating and desynchronization procedure.

It has been previously shown that, when optoelectrical information is naturally in advance on auditory — as in the case of a visible gesture like rounding —, vocalic/visual articulation can be identified only by eye several tens of milliseconds before any perceptible sound [3]. This bimodal information being in a sense naturally "desynchronized", an obvious experimental manipulation was to reduce the delay of auditory on vision, i.e. to make the speaker of visual signal in order to test the boundaries of such a bimodal temporal organization.

2. STIMULI
We used [#/y] and [#/i] (control) vocalic transitions embedded in a carrier sentence of the type: "Tas dit : UHU use?" ("Did you say : UHU ["Indian" name] wear out?"). Indian names were used in order to maximize pure vowel-to-vowel modulations of the output of the vocal tract, without intervening consonantal gestures. A French male talker was filmed at 25 frames/second. Stimuli were chosen with a 160 ms acoustic pause between vowels. For each digitized frame, articulatory parameters were automatically extracted by image processing [4].

For [#/i] transitions, upper lip protrusion PI starts at the end of [i] (Fig. 1), together with the constriction, i.e. lip area S begins to decrease. The domain we determined, for these two main components of the rounding gesture, stretches from 4 images before the pause, in order to allow a sufficient range in desynchronization, to 1 image after, when both components have reached their maximum.

3. VISUAL TEST
The visual information was explored by gating (40 ms steps): 10 gates of 1000 ms duration allowed to display in all cases the beginning of the carrier sentence "Tas dit... ." This visual [#/i] identification test (with 10 repetitions for each gate in random order) was performed by 10 naive subjects, with no deficit in vision and audition. The visual boundary in the [#/y] transition was measured on the mean curve of all subjects using Probit analysis. This boundary takes place 140 ms before the acoustic onset of [y], and less than 40 ms (a one image step) remains enough to switch from 80% [#/i] to 80% [#/y] (Fig. 1).

Anticipation is earlier (140 vs. 100) and category switching is steeper (40 vs. 80) than the one we obtained with another 160 ms paused signal, whose articulatory profile was actually slower, especially in the constriction building up [5]. Anyway, this 140 ms anticipation remains shorter than the maximum case we ever evidenced, in fact within a very long 460 ms pause: 210 ms [5].

Thus we confirm our previous findings on the natural advance of the eye relative to the ear.

4. AUDIO TEST
So what about the building up of audio information?

An audio [#/i] identification test, including the beginning of the stimuli "Tas dit...", stopped 2, 6 and 10 ms after the onset of [#/i] and [#/i], was performed by the same 10 subjects (10 repetitions by gate). Such a range from 2 to 10 ms has proven to be sufficient to scan properly the building up of featural acoustic information. Mean identification scores were: 58%, 95%, 99% for [#/i] (Fig. 2); and 100%, 99% et 99% for [#/i].

This supports the claim that often only one pulse (8 ms in the case of the vowel onsets of our talker) is sufficient to fairly identify the vowel (for French, see [6]).

The building up of visual information (40-80 ms) is thus slower than the audio one (10 ms). But this is fairly compensated by the visible anticipation on the sound, which is naturally displayed in speech (up to 200 ms) due to such a pervasive phenomenon as coarticulation.

5. AUDIOVISUAL TEST
But what are the audio/visual boundaries of this bimodal coherence?

The same visual stimuli were presented with the sound in synchrony or in advance. For each gate, for which we measured the time course of visual information, we tested the building up of the audio using the 3 steps previously determined (2, 6 et 10 ms), in order to obtain a desynchronization range from 0 to -360 ms, by 40 ms steps. The same 10 subjects where tested on the 10 steps for the 3 vowel onset durations.

Individual curves obtained for each acoustic duration show clearly different patterns. Since averaging was not representative, we grouped them according to similarity of their response profiles. Individual curves are either clearly S-shaped, or they show a first phase, before the visual [#/i] boundary, which is less regular and/or close to chance level (Figs. 3a-f). On the base of the scores for 10 ms vowel onsets (corresponding to high audio performances), we obtained two groups of 5 subjects. The first group (Figs. 3a-c) has, in the phase before the visual boundary, identification [#/i] scores below 20%; the second group has scores above 20% (Figs. 3d-f).

If you consider scores in the synchronous condition (plotted on gate n°10) for both groups, we see that, independently of vowel onset duration, individual [#/i] scores are generally at or above 90% (with one exception). Mean scores for 2, 6 et 10 ms durations are respectively: 96%, 98% et 99%.

Comparing audiovisual results obtained in the worst condition, 2 ms (96%), with the audio alone condition (58%), visual benefit reveals largely sufficient to disambiguate a poor audio signal. We thus join results in a more classical condition, namely speech in noise (for French, see [7]).

When desynchronization occurs, for these 2 ms vowels, we see that rounding information is an anticipating one in the original — can benefit up to -160 ms. One must recall that for this value, i.e. up to image n°6 (see Fig. 1), visual information alone reached 85% [#/i] responses, whereas just 40 ms before it scored 12%. In other terms, we were able to test step by step what phase of the anticipatory gesture could enhance ambiguous audio information. It comprises in fact all the phase "sheltered" by the gesture: after the visual boundary.

Let’s consider now desynchronization effect beyond this visual boundary, i.e. for the phase corresponding articulatorily to an [#/i]. For Group 1, we see that the duration of vowel [#/i] onsets comparatively to identification scores in the audio condition — does not seem to influence subjects’ behaviour. In fact, what is properly characteristic of this group is its high sensitivity to visible articulatory information. The curves we obtain for the three conditions display a clear S-shape, which looks strongly like the one (mean and individual) obtained for the visual alone condition: the identification boundary is located, for the three audiovisual conditions in the vicinity of the visual boundary. This similarity of the curves in the visual and audiovisual conditions indicates that, when desynchronization delivers images in advance of the sound — in this case an articulatory information specific of an [#/i] (in a desynchronization ranging from -200 to -360 ms, for this transition) —, then subjects identify the oncoming of an [#/i] vowel, in spite of the fact that they receive an audio information largely sufficient to recognize an [#/i]. Things are going as on if in the case of conflicting information — [#/i] being visible et [#/i] audible —, visible information was guiding perception.

Subjects from Group 2 are sensitive also to conflicting information. Whereas audio
is clearly identified by them as [y] — at least for 6 and 10 ms durations —, a visible [i] pushes their scores towards chance level.

To summarize: (i) desynchronization has the largest effect only when the visual boundary is crossed; (ii) beyond this boundary, no subject is insensitive to visual information, i.e. clearly no subject displays a steady 100% [y] along all desynchronization values. Moreover in (ii) the proportion of those who answer [i] for audio [y] is very close to the results found in a rounding judgment task for the same conflicting French vowels [8]. However, up to the present experiment, no such "McGurk effect" had been successfully obtained by a desynchronization procedure for vowels [9].

6. CONCLUSION

The natural delay of audio, relative to the visual signal, in speech coarticulatory anticipation, can be reduced without affecting intelligibility, as long as the configurational visual cues are in accordance with the sound. This hypothesis of a primacy of configurational over temporal coherence could be used to explain other results on desynchronization (reviewed in [10, 3, 11]) for detection tasks [12] as well as for intelligibility ones [13, 14].

Acknowledgements: This work was supported in part by a grant in Cognitive Science from the French Ministry of Research and Technology, a France Telecom-Université Stendhal contract n°9278032 and an Esprit Basic Research project n°6975 Speech Maps.

REFERENCES


Fig.1.– Mean visual identification results for [tad#y] superimposed on normalized articulatory parameters -5 and +P1 (scores are plotted on the last image of the gate).

Fig.2.– Mean auditory scores for [tad#y] with short [y] onsets. *Fig.3a-f.– Individual audiovisual scores for group 1 (left) and group 2 (right). N°10: scores in the synchronized condition with (from top to bottom) 2, 6 and 10 ms [y] onsets after image n°9. For scores on n°9 the sound is aligned on n°8, i.e. with a -40ms desynchronization; etc.; with a maximum -360 ms on n°1.*