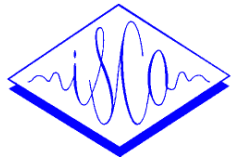
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

VISUAL PERCEPTION OF GLIDES *VERSUS* VOWELS: THE EFFECT OF DYNAMIC EXPECTANCY

Marie-Agnès Cathiard, Christian Abry & Jean-Luc Schwartz

Institut de la Communication Parlée, CNRS UPRESA 5009,
INPG-Université Stendhal, BP 25, F-38040, Grenoble Cedex 9
(e-mail: cathiard@icp.inpg.fr — abry@icp.inpg.fr — schwartz@icp.inpg.fr)

ABSTRACT

Does the visual perception of *glides* require a *dynamic* representation? In our previous experiments on the visual perception of rounding, we argued against the mandatory status of dynamic representations for visual *vowels*. In this paper, we focus on a specifically temporal contrast between the French vowel [y] and the corresponding [ÿ] glide. Results are twofold. First: for subjects who use this contrast, we demonstrate that the duration of the *static* phase of rounding provides a basic correlate of the vowel *vs.* glide identification. Second: a gating experiment indicates that the intrinsically dynamic nature of the glide is not exploited *until subject expectancy is oriented towards motion processing*. Again we cannot support an exclusive visual dynamic representation neither for vowels nor glides.

1. INTRODUCTION

In their chapter of *Hearing by Eye II*, Rosenblum & Saldaña [1, p. 76] repeatedly defended the idea that: «The recent evidence on speechreading, auditory speech and visual event perception — along with a new conception of speech events as gestural — supports time-varying information as primary». This is an ecological perception and «dynamical» stance originating in Strange's claims [2], addressed before exclusively in the field of *auditory vowel* perception. After an examination of controversial arguments, coming from Cathiard [3] and Campbell [4], talking against their preceding experiment [5], they finally pushed forward the following *visual* issue: «Still, the question remains of where the most salient visual information lies. According to the time-varying information thesis, the most coarticulated portions of the utterance should be most visually salient. Specific predictions could be made based on the auditory speech findings: for example, dynamic margins around a visible vowel should be more informative than the "steady-state" portions. Thus, examining the salience of coarticulated portions of visible speech should be a straightforward way to test time-varying information» [1, p. 76].

As a matter of fact, the answer lies explicitly in our own data on the *visual* perception of anticipation for the rounding gesture [3, 6]. (i) We showed that the «steady-state» portions, corresponding to the articulatory climaxes (or «targets») of the visible vowels, displayed obviously the best identification scores for rounding, whereas the «dynamic margins» gave scores that waxed and waned, like the articulatory gestures did [6, p. 215]. Just to deal with the case of anticipation, presentation of frozen images taken in the vicinity of the vowel acoustic onset delivered the same ceiling values (100%) than for a time-varying display. (ii) The only effect of this moving display concerned the location of the perceptual 50% rounding boundary, which took place during the «dynamic margin» of the vowel: it could occur sometimes 30 ms ahead of the curves recovered from a static presentation. Interestingly this dynamic benefit was not obtained for an optimal view, which in the case of the rounding gesture we demonstrated to be a profile one, but for front views, where protrusion in-depth can be recovered from *shading* (in frozen views) or *shading-from-motion* (in the movie). To account for this pattern of results, we proposed a *shape-from-shading-from-motion* approach, arguing that the *visual vowel shape* could be a possible representational format, whereas movement would be just a help for recovering shape, when this shape is undersampled (which is typically the case of point-light displays used by [5]), or not optimally projected (in our case, for the front-viewed waxing rounding gesture).

This visual phenomenon called *shape-from-motion* or *structure-from-motion*, in spite of recent advances (lately [7]), is still in need of an adequate brain functional parcellation [8, 9] between: (i) high-order motion processing, for point-light displays and revolving random-dot cylinders; *versus* (ii) low-level motion processing for kinetic boundaries. In addition, a recent PET imaging study [10] indicated that the McGurk illusion itself would possibly not rely on cortical

processing as reported by earlier studies, but on a subcortical network. In our opinion it could support an *auditory-visual motion processing* in the [ba]/[ga] dynamic opening phases, with a good candidate for intersensory and sensory-motor integration in audio-visual scene analysis, i.e. the superior colliculus [11]. We do not know if cross-talk between cortical (MT or V5) and subcortical networks could occur only during *experiential* illusions, such as Necker's depth-from-motion [7], binocular rivalry and… experiential McGurk. Anyway these motion processing systems are not needed for «steady-state» portions in vowels and consonants, which have to be crucially spotted in the speech flow, owing to an intact visual *shape* system [12]. And this appears especially in adverse conditions as revealed by the increase of mouth fixations for speech in noise, with a foveal eye region reputedly less sensitive to movement than the peripheral one [13, pp. 134-135].

As concerns now specifically the *audio-visual vowel*, one of the strongest natural counterevidence against the claim that «time-varying information is primary» comes from the very temporal organization of visible and audible vowel information in speech [6, p. 219, footnote 1]. What we observed in our articulatori-acoustic data, was that, in initiating an utterance, after a pause, typically with an initial vowel, the first glottal pulse occurred at or nearly the point where the articulatory setting of the desired vocalic configuration of the vocal tract was achieved. If the featural/gestural information on the oncoming vowel would have to take advantage of the dynamics of the gesture towards its target, the glottal excitation would have to be initiated as soon as possible during the transitional *gliding* phase, just in order for it to be heard. But this is clearly not what the speech temporal organization reveals (for a test of this configurational and temporal coherence see [14, 15]).

In concluding our paper coping with the representation of the *visual vowel* [6], we drew attention on the specific case of *glides*, which could differ from the vowels, since they could be typically made of «dynamic margins» only (traditionally called *on-glides* and *off-glides* in vowels and diphthongs), i.e. basically conceived as intrinsically time-varying in nature. The experiments presented here address two questions about this issue: (i) Is a "steady state" portion necessary for identifying a vowel *vs.* a glide ? (ii) Is the "dynamic margin" *per se* processed differently when subjects *expect* it is a glide or a vowel?

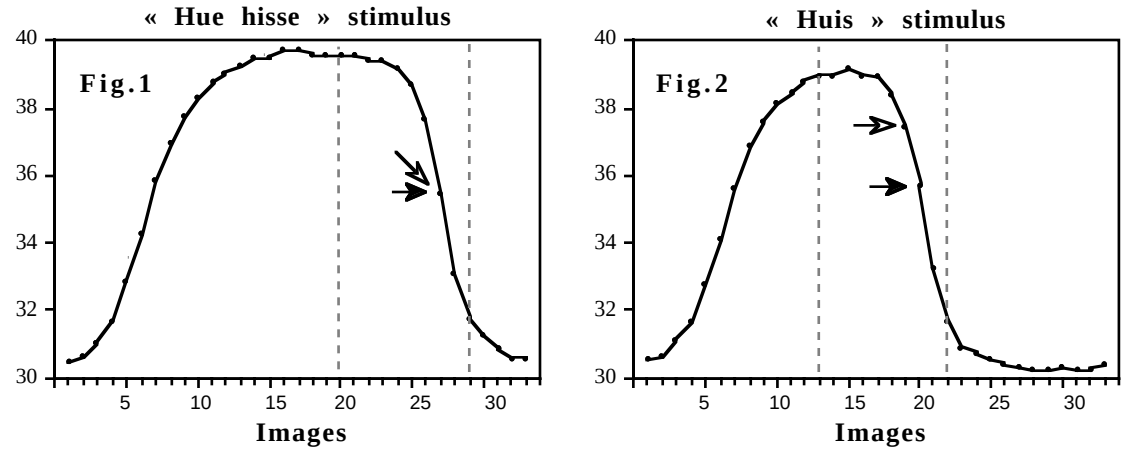
2. STIMULI AND ARTICULATORY ANALYSIS

We used [i#yi] and [i#ÿi] transitions which were embedded in the following carrier sentences, controlling a pause as: «T'as dis: Hue hisse?» [tadi#yis] («Did you say: …?») and «T'as dis: Huis?» [tadi#ÿis], where «Hue» and «Huis» are proper names and «hisse» a third person verb («to raise»). 10 repetitions of each of these two sentences were recorded audiovisually, at 25 images/second, by a French talker, with simultaneous front and profile views. After image processing, delivering, among other parameters, upper lip protrusion and area between lips, two utterances were selected. Their articulatory time course are very similar as concerns the «dynamic margins», i.e. the building-up phase of the rounding gesture and the retraction phase towards [i]. Thus they differ essentially by their «steady-state» portion, i.e. the *plateau* phase of protrusion, which is clearly longer for [y] in «T'as dis: Hue hisse?» (see Fig. 1 *vs.* 2). Note that these stimuli display close maximal and minimal values for lip area and upper lip protrusion, as well as very similar velocity and acceleration profiles in the transitional phases.

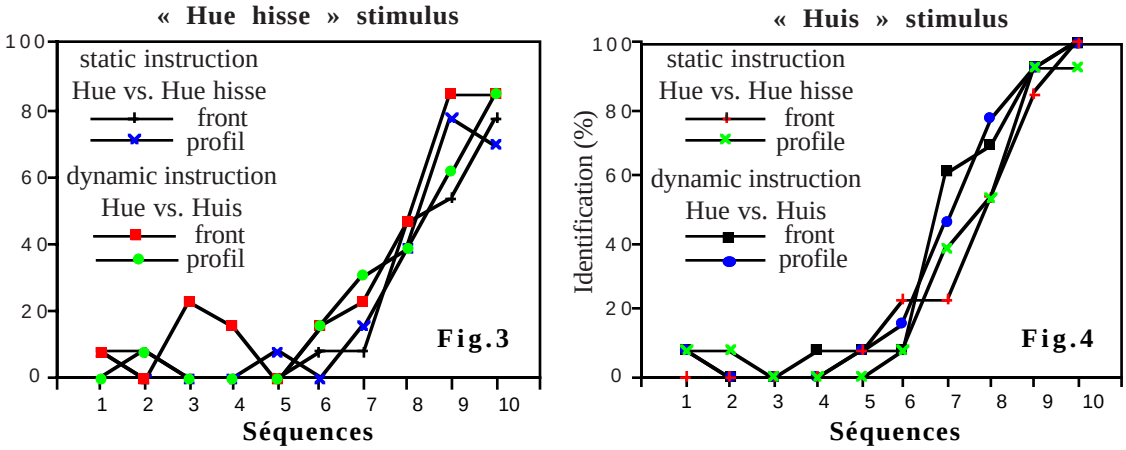
3. PRETEST: CONTRASTING [y] *vs.* [ÿ]

First, we realized a visual pretest in order to know if the two sequences «T'as dit: Hue hisse» and «T'as dit: Huis» could be identified with visual information only. These two sequences were presented in their full time course, 10 times each in a random order, under a front angle view. 27 French normally-hearing subjects, 18 to 20 years old, without visual deficit, participated in the test. Only 13 subjects obtained an identification score higher than or equal to 60% (60%=1 subject; 65%=1; 70%=3; 75%=3; 80%=2; 85%=1 and 90%=2).

This result indicates that the identification of the glide *vs.* its corresponding vowel, with the visual information only, is subject dependent. This corresponds to the phonological dialectal situation in the French speaking community, where e.g. «muet» («dumb») can be pronounced [mye] or [mÿe], depending on the degree of sensitivity of the subject to realizations with or without a glide. According to Klein's scale for Parisian French [16], «Hue hisse» has the lowest probability to be syllabified [ÿi], whereas «Huis» could hardly give rise to a two-syllabic [yi]. Consequently we can consider that our experiment corresponds to a kind of maximum possible visual difference.



Figures 1 and 2: Time course of upper lip protrusion for the vowel [y] in «Hue hisse» and for the corresponding [ÿ] glide in «Huis». The values of upper lip protrusion are given in mm, measured from a reference-point on a ruler fixed to the talker's goggles. On the horizontal axis, image number is indicated. The vertical broken lines specify the domain explored step by the last images of the 10 gated sequences in Exp. 2. The black arrows point to the identification boundaries obtained with the «static» instruction and the white arrows to boundaries from the «dynamic» instruction (see Text for Exp. 2).



Figures 3 and 4: Identification functions for «Hue hisse» and «Huis», obtained with «static» and «dynamic» instructions for front and profile angle views. For each stimulus: with «static» instruction, the curves display the « Hue hisse » identification percentages; and with «dynamic» instruction, the «Huis» identification scores.

4. EXP. 1: CONTRIBUTION OF THE *PLATEAU* DURATION IN ROUNDING

We wanted, in a first time, to test the sensitivity of the same 13 subjects (who succeeded in the identification pretest) to *plateau* duration in rounding. We prepared 3 stimuli with intermediate durations between «T'as dis: Huis?» and «T'as dis: Hue hisse?». First, stimulus Hh-1 was obtained by removing, from the stimulus «Hue hisse», one full image at the plateau centre: in fact this corresponds, in Fig. 1, to the suppression of 2 image numbers, or «fields», namely n°18 and n°19, since a video image is composed of two interlaced fields. Then, stimuli Hh-2 and Hh-3 were obtained by removing respectively 2 and 3 images at the plateau centre, i.e. 4 or 6 fields, respectively from n°16 to n°19 in Hh-2 and from n°16 to n°21 in Hh-3. Notice that the suppression of a fourth full image of the plateau in «T'as dis: Hue hisse?» would have resulted in a stimulus with a plateau duration as small as the plateau duration of «T'as dis: Huis?». Hence we do have at our disposal a complete continuum.

We presented the 5 stimuli («Huis», «Hue hisse» and the 3 intermediate stimuli «Hh-1», «Hh-2» et «Hh-3»), with a front angle view, 10 times each in a random order. Subjects could identify each of them as «Huis» or as «Hue hisse». The mean «Hue hisse» identification percentages are: for the stimulus «Huis» 18.5%, for «Hh-1» 28.5%, for

«Hh-2» 44.6%, for «Hh-3» 69.2% and for «Hue hisse» 80%. Hence increasing plateau duration results in increasing the number of «Hue hisse» identifications.

From this experiment, we can conclude that the plateau duration of the rounding gesture seems to be a relevant visual cue in order to identify the rounded [ÿ] glide *vs.* the corresponding [y] vowel.

Nevertheless, the presentation of the sentences *as a whole* does not allow us to know *at which moment* the subjects have sufficient information to decide in favour of the glide *vs.* the vowel.

5. EXP. 2: CONTRIBUTION OF THE TIMING OF THE *ONSET* OF THE RETRACTION PHASE

If plateau duration in rounding is a relevant cue for the realization of the task, it is likely that subjects must wait for the end of this plateau — i.e. wait for the *onset* of the retraction phase of rounding — in order to perceptually evaluate plateau duration. We decided to explore the timing of this retraction event by a *gating* technique, in order to obtain the evolution of identification scores step by step.

For each sequence «T'as dis: Huis?» and «T'as dis: Hue hisse?», we prepared 10 gates (each with a duration of 1200 ms), which always included the onset of the sentence and the middle of the rounding plateau. The gates glide, by 20 ms steps, along the retraction gesture (cf. Fig. 1 and 2; in this display a 20 ms step is feasible by stopping on one of the two fields of the video image or frame, from which the missing lines are restored by linear interpolation). Moreover we decided to test the identifications with front and profile angle views, the rounding gesture being better recovered through a profile view [6]. So we have a set of 4 tests: front «Hue hisse», profile «Hue hisse», front «Huis» and profile «Huis». For each test, the 10 gated sequences were presented in a random order.

In addition, within this gating paradigm, we manipulated subjects' *expectancy*. We assumed that, as long as the subject sees only the plateau phase, he/she expects more probably the vowel to be [y]; and it is only with the onset of the retraction gesture — which indicates the end of the plateau —, that the subject could decide in favour of [y] *vs*. [ÿ]. The manipulation of the instruction aims at modifying the subject's *expectancy of the moment where the retraction event is likely to appear*. Each test was presented two times to each subject. First we used a «static» instruction where the subject must identify the gated stimuli as «T'as dit: Hue?» or «T'as dit: Hue hisse?». Second we used a «dynamic» instruction where the subject must identify the gated stimuli as «T'as dit: Hue?» or «T'as dit: Huis?» (stimuli, instruction order and angle view were counterbalanced). The issue of this instruction manipulation is the following. Does the subject take more advantage of movement when he/she expects a rather dynamic lip configuration (as it is the case for the «dynamic» instruction, with which subjects are prepared to identify a *glide*), than when he/she expects a rather stable lip configuration (as it is the case with the «static» instruction, with which subjects are prepared to identify a *vowel*)?

Identification curves obtained by the 13 subjects are given for «Hue hisse» (Fig. 3) and «Huis» (Fig. 4), with front and profile views and with static and dynamic instructions. All identification curves have a classic S-shape. Our interest will be specifically centred on the boundaries at 50%.

Concerning the «Hue hisse» stimulus (Fig. 3), we can observe a stability in the boundaries, which take place, for the four curves, about the final image of sequence n°8: this corresponds, on the articulatory signal (see Fig. 1), to a moment where the retraction gesture is quite far on its way towards [i], nearly half of its time course.

Concerning the «Huis» stimulus (Fig. 4), we can observe, for the static instruction and for the two angle views, that the boundaries also take place about the final image of the same sequence n°8. For the dynamic instruction and for the two angle views, the boundaries occur about the final image of sequence n°7, hence with an advance of about 20 ms. In this condition, the subjects have taken the retraction into account more precociously, i.e. at a moment where the retraction gesture is at its onset (see Fig. 2).

We do not observe boundary differences depending upon angle view. This is not contradictory with our preceding results, since dynamic presentations of front views could reach *sometimes* scores for profile. What we showed at that time was that, contrary to front views, profile frozen displays were *always* as good as dynamic ones [3, 6]. So we cannot conclude on this issue, until we test frozen views of our present retraction gesture, which of course was not a priority in this experiment on glides. In addition, at that time, we did not manipulate expectancy in order to evidence a possible advantage.

This manipulation evidences now the following results. Boundaries for [y] in «Hue hisse», both for static and dynamic instructions, correspond to the boundaries of [ÿ] in «Huis», with static instruction. But the boundary is about 20 ms in advance for the [ÿ] glide with the dynamic instruction, in comparison with its boundary in the static one.

Consequently, subjects do not seem to process differently the time course of a [y] vowel *vs.* the time course of a [ÿ] glide, *unless their perceptual expectancy is particularly oriented towards the possible precocity of the onset of the retraction.*

Considering this pattern of results, we can set forth that, in order to take advantage of the more dynamic nature of a visual speech stimulus, it is necessary: (i) not only that the stimulus affords it (in our case when the subject sees a glide, and not a vowel); (ii) but also that subject's *expectancy* be guided (as it is the case with the dynamic instruction) towards the processing of events, be it for their kinematic or metrical (rhythmic) properties.

6. CONCLUSION

These results on the visual perception of the French rounded glide [ÿ] show that only half of our French subjects (13 on 27) are able to *visually* identify the glide *vs.* the vowel, which corresponds to the weak linguistic status of the contrast in the French-speaking community.

For these 13 subjects, it appears that the plateau duration of rounding — when experimentally shortened in the vowel — allows them to switch from the perception of [y] to the perception of the corresponding [ÿ] glide. We can compare this result to these of Greisbach [17] on the visual perception of geminate *vs.* non-geminate plosive consonants in German, which show that the duration of the occlusion, or «steady-state» phase, is relevant, especially for visible plosives as [p]. Such a «steady-state» portion seems thus necessary when a linguistic categorization calls for it, for a vowel as well as for a consonant.

Concerning the question of the relevance of the «dynamic margins» for the representational formats of vowels and glides, we must keep in mind that such «margins» occur both marginally for vowel on/off-glides and centrally in glides. It seems that subjects do not visually process differently this «dynamic portion» in a glide or a vowel. It is only when their perception is guided, by a manipulation of their expectancy, towards the more dynamic component of the glide itself, that movement information can be beneficial.

It remains to evaluate how these results obtained by a perception task oriented towards motion processing are compatible with *Expectancy Theory* proponed by Jones [18], which is still productive in musical rhythm and flow perception. The vocalic gestures, which are basically pervasive in the speech flow, carrying consonantal coarticulated gestures, do not seem to need a dynamic representational status. Since glides are also ubiquitous in the speech flow, appearing naturally as the unavoidable transitional portions between vowel «steady-state» phases, do they need a special status? It seems that is only when the transitional glide phase can be manipulated linguistically (as when the natural transitional consonant [p] in «Thompson» can be voluntarily lenghtened or reduced), that a dynamic feature can be evidenced by urging listeners who pratice this glide control to perceive the benefit of it. This was visually the case for the 13 concerned French subjects.

Again we cannot support an *exclusive* motion representational stance, this time for the *visual* perception of both *vowels* and *glides*, even if it remains possible that, *in acoustic and bimodal perception*, glides could need an intrinsically dynamic representation in order to distinguish them from their corresponding vowels.

7. REFERENCES

1. Rosenblum, L.D., and Saldaña, H.M. "Time-varying information for visual speech perception", in R. Campbell, B. Dodd and D. Burnham (Eds.), *Hearing by Eye II*, 61-81, Psychology Press, 1998.
2. Strange, W., Verbrugge, R., Shankweiler, D. and Edman, T., "Consonant environment specifies vowel identity", *Journal of the Acoustical Society of America*, Vol. 60: 213-224, 1976.
3. Cathiard, M.-A., *La perception visuelle de l'anticipation des gestes vocaliques: Cohérence des événements audibles et visibles dans le flux de la parole*, Thèse de Doctorat de Psychologie Cognitive, Université Grenoble 2, 1994.
4. Campbell, R., "Seeing brains reading speech: A review and speculations", in D. Stork and M. Hennecke (Eds.), *Speechreading by Humans and Machines*, NATO ASI Series F: Computer and Systems Sciences, Vol. 150: 115-133, Springer-Verlag, Berlin New York London Paris Tokyo, 1996.
5. Rosenblum, L.D., and Saldaña, H.M., "An audiovisual test of kinematic primitives for visual speech perception", *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 22(2): 318-331, 1996.

6. Cathiard, M.-A., Lallouache, M.-T., and Abry, C., "Does movement on the lips mean movement in the mind?", in D. Stork and M. Hennecke (Eds.), *Speechreading by Humans and Machines*, NATO ASI Series F, Vol. 150: 211-219. Springer-Verlag, Berlin New York London Paris Tokyo, 1996.
7. Bradley, D.C., Chang, G.C., and Andersen, R.A., "Encoding of three-dimensional structure-from-motion by primate area MT neurons", *Nature*, Vol. 392: 714-717, 1998.
8. Vaina, L.M., Lemay, M., Bienfang, D.C., Choi, A.Y., and Nakayama, K., "Intact 'biological motion' and 'structure from motion' perception in a patient with impaired motion mechanisms: A case study", *Visual Neuroscience*, Vol. 5: 353-369, 1990.
9. Marcar, V.L., Zihl, J., and Cowey, A., "Comparing the visual deficits of a motion blind patient with the visual deficits of monkeys with area MT removed", *Neuropsychologia*, Vol. 35(11): 1459-1465, 1997.
10. Okada, K., Kawashima, R., Fukuda H., Mori, K., Imaizumi, S., Kiritani, S., and Ogawa, A., "A PET study of the McGurk effect", *NeuroImage*, Vol.7(4), May, Parts 2 of 3 parts, 1998, p. S163.
11. Stein, B.E., and Meredith, M.A., *The merging of the senses*, MIT Press, 1993.
12. Campbell, R., "Seeing speech in space and time: Psychological and neurological findings", in *Proceedings of the Fourth International Conference on Spoken Language Processing*, 1493-1496,, 1996.
13. Munhall, K.G., and Vatikiotis-Bateson, E., "The moving face during speech communication", in R. Campbell, B. Dodd and D. Burnham (Eds.), *Hearing by Eye II*, 123-139, Psychology Press, 1998.
14. Cathiard, M.-A., Lallouache, M.-T., Mohamadi, T., and Abry, C., "Configurational *vs.* temporal coherence in audivisual speech perception", in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Vol. 3: 218-221, 1995.
15. Abry, C., Cathiard, M.-A., and Lallouache, M.-T., "How can coarticulation models account for speech sensitivity to audio-visual desynchronization?", in D. Stork and M. Hennecke (Eds.), *Speechreading by Humans and Machines*, NATO ASI Series F, Vol. 150: 247-255, Springer-Verlag, Berlin New York London Paris Tokyo, 1996.
16. Klein, M., *Vers une approche substantielle et dynamique de la constituance syllabique. Le cas des semi-voyelles et des voyelles hautes dans les usages parisiens*. Thèse de Doctorat de Linguistique, Université Paris 8, 1991.
17. Greisbach, R., "Parametrization and integration of facial speech gestures", *Speech Maps (Mapping of Action and Perception in Speech)*, EEC Esprit/BR Project N°6975, year 3, Vol. IV et V, Deliverable 33, RP2: 23-38, 1995.
18. Jones, M.R., and Boltz, M., "Dynamic attending and responses to time", *Psychological Review*, Vol. 3: 459-491, 1989.

ACKNOWLEDGMENTS

To Tahar Lallouache with whom we prepared the experiments reported here. To Ruth Campbell, Jean Decéty and Julie Grèzes for their references. This study was supported by a Région Rhônes-Alpes project (directed by G. Tiberghien, Institut des Sciences Cognitives, Lyon, France).