Consequences on bimodal perception of the timing of the consonant and vowel audiovisual flows

Emilie Troille¹,², Marie-Agnès Cathiard² & Christian Abry²

¹ GIPSA-Lab-ICP, UMR 5216 CNRS-INPG-Université Stendhal
troille@icp.inpg.fr
² Université Stendhal - BP25 - 38040 GRENOBLE Cedex 9 FRANCE
marieagnes.cathiard@u-grenoble3.fr, christian.abry@u-grenoble3.fr

Abstract

It is now well known that speech can be seen before it is heard: this has been extensively shown for the vowel rounding anticipatory gesture leading the sound [1]. In this study, the perception of French vowel [y] anticipatory coarticulation was tested throughout a voiced fricative [ɔ] consonant with a gating paradigm. It was found that vowel auditory information, as carried by the noise of the fricative, was ahead of visual and even audiovisual information. Hence the time course of bimodal information in speech cannot be considered to display the same pattern whatever the timing of the coordination of speech gestures. As concerns vocal information only, consonantal coarticulation can carry earlier auditory information than the vowel itself, this depending of the structure of the stimulus. In our fricative-vowel case, it was obvious that the vowel building up movement was audible throughout the fricative noise, whereas the changes in formant grouping (indicating mainly the change from [i] to [y]) occurred later. As concerns the timing of audio and visual information, it becomes more and more obvious that the dates of the delivery of each component of the perceptual flow can change drastically the results obtained in the time course of brain activity for bimodal processing. Hence the flexibility of the phase relationships between the two flows urges to a better knowledge of the natural ecological variability in audiovisual signal production, a flexible coherence not to be violated on pain of breaking the Gestalt laws.

1. Introduction

Saying that speech is bimodal became quite a trite statement. But in spite of significant contributions accomplished for many years toward modelling audiovisual speech integration, the contributions of hearing and vision are far from being completely understood. Just start with two first questions, one basic about the input signals, and one still far from reach in the brain. (i) What about the respective time course of visible and audible speech flows, making seamless co-modality efficient? (ii) And what about the neural time course of integration for perceptual decisions? One third way to render “piecewisely” tractable the last question is to bind (i) and (ii) into (iii): then if “speech can be seen before it is heard” (what we illustrated some 10 years ago [1]), what about the benefit of this natural anticipatory timing for the brain? Recent studies in primatology brought a decisive outcome to this issue. According to Ghazanfar et al. [2]: “Coos are long-duration, tonal calls produced with the lips protruded […]. As in human speech [3], the onsets of mouth movements during production of rhesus monkey vocal signals predate the auditory component. [After processing of Local Field Potentials in the Superior Temporal Sulcus, and auditory Core Belt and Parabelt] enhanced responses were primarily seen when this time interval was <100 ms, and suppressed responses were primarily seen at intervals >200 ms.” (2), p. 5010). They added that the trespass of this Gestalt grouping window explained probably why other studies failed to evidence integration.

In order to disclose the perceptual unfolding of auditory and visual contributions, when presented alone or bimodally along time, we chose the vocalic rounding anticipation phenomenon, which offers a natural test case, in which one does not need to degrade audio by noise, while avoiding ceiling effects usually met in experiments with original signals.

Vocalic rounding anticipation was the topic of many studies as well in the field of speech production as in speech perception. It is well known that this vocalic gesture can begin several consonants before the rounded vowel (for an overview concerning English, see [4]; for French see [3] and [5]).

Visual identification of this rounding anticipation was extensively studied by Cathiard [6] along silent acoustic pauses. Sequences UH and IH were inserted in the sentence “Tu dis: UH ise?” [tydiʃiˈziz] ("Do you say: …") with a 160 short pause and 460 ms long one. Articulatory measurements showed that the rounding gesture began during the silent pause. In fact when visually tested, long pause and short pause anticipations, delivered categorical boundaries (50% [y] identification) from 90 to 210 ms before the acoustic onset of the [y] vowel. Which is exactly the range given by specialists of the “merging of the senses” (referred to by [2], just above).

As regards the perception of rounding anticipation through consonants, it was first tested auditorily in French by Benguerel and Adelman [7]. And recently by Hecker et al. [8], who showed again that anticipatory cues were carried in intervocalic consonants, naturally more by fricative long noises, than by plosive short bursts.

Few studies compared auditory and visual modalities in the consonant and vowel time flows. Escudier et al. [9] showed, with [iziy] sequences, that vision was in advance on hearing (a 40 to 60 ms lead). This result was strongly related to their stimulus, with an extremely early constriction gesture for [y], i.e. as soon the second half of the [i] vowel. In a more recent study, Roy [10] compared auditory and visual perception of normal-hearing and deaf subjects in [köy], sequences (n being a variable number of consonants). In audio condition, when Cn contained only plosive consonants, the rounded vowel was perceived at the release of the last consonant, close to the [y] vowel. But when Cn contained fricative consonants, [y] perception corresponded simply to the occurrence of a lowering of the lower limit of high energy, characteristic of the
friction noise. In this case, Roy noted that normal-hearing subjects were naturally more efficient to perceive the rounding anticipation in the auditory modality than in vision (but no quantitative estimation of this advance was given).

To our knowledge no study compared monomodal vs. bimodal CV flows. In this study, we propose to test the ongoing vocalic information through the consonant flow, under three conditions of presentation: auditory, visual and audiovisual.

2. Main experiment design

2.1. Recording

A French male speaker was audiovisually recorded. He produced in random order 12 repetitions of the 2 sequences “Ta[ as dit ZIZU ze!?” and “Ta[ as dit ZIZI ze!” (“Did you say…?”). These sentences were chosen to explore the whole transition, from the vowel [i] to the vowel [y], with an intervocalic [z] consonant. It is well-known indeed that this fricative is permeated by audible vocalic coarticulation effects (Whalen [11]).

The video recording was carried out in an anechoic booth, at 50 frames/second, with front and profile views, with the sound sampled at 22.05 kHz. A make-up in blue of the speaker lips allowed thereafter incrustation of a saturated black, using a Chroma-Key, in order to perform an accurate detection of labial parameters, notably small between-lip areas.

The ICP-Lip-Shape-Tracker (Lallouche [12]) via automatic image processing of the videos, provided a set of different lip parameters every 20 ms (frame by frame).

2.2. Data

Two articulatory parameters which characterize fairly well the rounding gesture, i.e. upper lip-protrusion (P1) and lip-area (S), will be considered.

The acoustical analysis consisted in the follow-up of formants (F1 to F4) for all sequences in order to choose one [zizi] realization and one [zizi] realization, which were the most similar as to the beginning of their initial [zizi] transition (fig. 1).

We also followed in the two stimuli the “formantic” resonance transitions along the frication noise of the intervocalic consonant. We used an LPC analysis, with a 0.04 s window appropriate (according to Munson [13]) for keeping track of the coarticulation effect (change in resonances) due to the following vowel. In the [zizi] transition, the falling movement of the main resonance, read from the frication noise of the intervocalic consonant [z], ran from 5653 Hz to 3293 Hz (fig. 1-2). Whereas frequencies remained stable in the [zizi] control transition, around a high value, with a mean 6067 Hz. The moving down of frequencies corresponded to the reduction in lip area along [z], due to the rounding anticipation of [y] (fig. 1-2). Hence it was like if the motion of the subject’s lips could be heard through the noise of the fricative. We also noticed throughout [z] in [zizi], a canonical falling trajectory of the third formant from [i] (3000 Hz) to [y] (2400 Hz). While it oscillated about 3000 Hz in [zizi].

![Figure 2: Follow-up of the high energy falling move in the fricative noise along the intervocalic consonant in [zizi], compared with change in lip area (S). Dates on the time scale are gating steps.](image)

2.3. Tests

We adopted a gating paradigm, currently used in speech perception studies, in order to test the auditory, visual and audiovisual identification of the rounded vowel through the consonant. Under the 3 conditions, sequences all started at the beginning of “Ta[ as dit” and finished at different dates in the gating domain with 20 ms steps. For [zizi], this domain started 40 ms before the acoustic offset of [i] (date:1560 ms) and finished at the end of [z] consonant (date: 1720 ms). For [zizi], 9 gated sequences were obtained and repeated 10 times. Three [zizi] gated sequences, used as control were repeated 5 times: they were introduced in the tests so that subjects could hear and see enough true [zizi] stimuli.

The audio gate sequences were realized with a Praat script. For the audiovisual test, images were added frame by frame to audio gated signal with Adobe Premiere Pro 1.5. For the visual test, we assembled the images, in front view, with only the onset of the audio sequence, i.e. just “Ta[ as dit”, used as an attentional starter for the subject. The sequences were inserted in Multimedia Toolbook software and presented in random order in the 3 tests (audio, visual and audiovisual).

2.4. Subjects

26 French participants, without auditory or visual deficiency were tested. None was familiar with lipreading. The group was
composed of 2 men and 24 women, from 20 to 25 years (mean 21 years and 3 months). Each subject began with the audiovisual test, then the audio test, and finished by the visual test. We chose first this order because we wished to collect uppermost the audiovisual identification (see section 4: Control experiment). The task was to decide if the final vowel of each gated sequence was [i] or [y].

3. Results

3.1. Results according to modality

For each modality, individual identification curves are presented together with the mean curve for all subjects. The identification functions – traced from [y] percent responses for each [zizy] gated sequence – have a classical S-shape. Of course [zizy] sequences are generally identified as [i].

(i) In the audio condition (fig. 3), the vowel [y] is identified at 1640 ms date with a 80% correct score, i.e. clearly in the [z] consonant (since the onset of this vowel is located at our last gating date, i.e. at 1720 ms). The individual identification curves are relatively grouped around the mean curve and the individual boundaries (at 50% [y] identification) varied from 1610 to 1640 ms dates.

(ii) The visual identification curves move very slowly (fig. 4). The mean identification curve reaches 80% correct only 20 ms before the acoustic onset of the vowel. A large variation span of the individual identification boundaries (at 50%) is observed from 1640 to 1710 ms dates, i.e. in a 70 ms span.

(iii) In the audiovisual test, the rounded vowel reaches more than 80% at 1680 ms (fig. 5). Dispersion of the individual curves around the mean curve is intermediate, in-between monomodal conditions. The individual 50% boundaries vary between 1620 and 1670 ms dates.

3.2. Comparison between modalities

Figure 6 presents the mean identification curves obtained in the 3 conditions (auditory A, audiovisual AV and visual V). The auditory identification of the vowel [y] is the earliest. It is followed by the audiovisual identification; then the visual identification. The analysis of the individual data indicates that this order is observed for 24 participants out of 26.

We carried out a Probit analysis (Finney [14]) on individual results to extract the boundaries and the slopes for each participant under the three conditions. We ran, on the boundaries as well as the slopes, an ANOVA (SPSS 13.0) with one intra-subject factor, i.e. the condition (A, AV or V).

(i) As regards boundaries (after a Greenhouse-Geisser correction of variance homogeneity), the condition of presentation has a significant effect: \( F(1,5,38) = 172.42; p=0.000 \), with \( A > AV > V \). Thus, the mean boundaries at 50% are: at 1629 ms for the audio, 1649 ms for audiovisual, and at 1673 ms in the visual condition. So the [y] vowel is respectively identified: 91 ms, 71 et 47 ms before its acoustic onset (located at 1720 ms).

(ii) ANOVA on the slopes also indicated an effect of the condition: \( F(2,50) = 7.451; p<0.001 \) (with no correction of the variance). There is no difference between the auditory and the audiovisual conditions (\( F<1 \)). When we group the audio and audiovisual data, and these new slopes are compared with the
visual ones, we obtain a significant difference: 
\( F(1,25)=11.18; p=0.003 \). The switch of identification is thus comparable in audio and audiovisual, but slower in visual identification.

### 3.3. Perception-production relationships

We compared mean identification curves to acoustic and articulatory parameters.

![Figure 7: Auditory and audiovisual identification curves (left scale) compared with high energy move in the frication noise (crosses; right scale) along the intervocalic [z] consonant in [zi zi]. Dates on the time scale are gating steps.](image)

(i) The comparison of the auditory and audiovisual identification functions, with the evolution of the [z] resonance trajectory into the frication noise (fig. 7), shows that the perception of the vowel [y] increases as soon as the movement in the frication noise starts to go down. We notice a 20 ms lag of the audiovisual identification after the earlier auditory identification, [y] being perceived in audio when the main frequency change of [z] goes under 4500 Hz (under 4000 Hz in the audiovisual).

![Figure 8: Visual and audiovisual identification curves (left scale) compared with lip area evolution (right scale, S inverted) along the intervocalic [z] consonant in [zi zi]. Dates on the time scale are gating steps.](image)

(ii) In front view, the relevant visual parameter is lip area (fig. 8). We compared lip-area functions with the visual and audiovisual [y] identifications. In the audiovisual condition, the vowel was correctly identified when the lip-area decreased below 1 cm²: the identification switched from 23% with 1.12 cm² (at date 1640 ms) to 76% with 0.85 cm² (1660 ms). On the contrary, the switch is later and slower for the visual identification: it is established at time 1680 ms, when the lip area is lower than 0.62 cm².

![Figure 9: Visual and audiovisual identification curves (left scale) compared with upper lip-protrusion evolution (P1 right scale) along the intervocalic [z] consonant in [zi zi]. Dates on the time scale are gating steps.](image)

(iii) In fact, in vision only, subjects seem rather to follow the evolution of the upper lip protrusion (fig. 9). In others terms, subjects probably waited to see a quite rounded labial form with lip pursed to be sure of their vowel identification.

In summary, in the auditory modality, the participants use as soon as possible the acoustic information to identify [y], following the lowering move of frication noise high energy in [z] intervocalic consonant, as a cue of the rounded vowel to come. In visual condition, the articulatory cues do not enable them to be so early. So it seems that a small audio movement in frication noise is enough for the subjects to start to identify [y], whereas they should wait a limit of constriction (below 1 cm²) to identify the vowel, visually. Under the audiovisual condition, it seems that the subjects were influenced by the two modalities at the same time.

### 4. Control experiment

Since we wanted to know if the order of presentation of the 3 conditions AV, A, and V, chosen for the main experiment, could have an effect on the pattern of results, we replicated this experiment using the same conditions, but in random order. Hence we tested 24 others subjects, with about the same age (mean 20 years and 9 months) and about the same proportion of women (21) and men (3).

![Figure 10: Comparison of the mean curves in auditory, visual and audiovisual modalities: A>AV>V. Dates on the time scale are gating steps.](image)

For 21 subjects out of 24 we found the same identification order A>AV>V with very similar curves: like these observed in our first experiment (fig. 10).

We carried out another Probit analysis to extract boundaries and slopes for each subject under the three conditions and we
ran an ANOVA with one intra-subject factor, i.e. the condition (A, AV or V). The condition of presentation had a significant effect: $F(2,46)=112.778$, $p<0.001$, with A>V>AV. The mean boundary (50%) being at 1627 ms date in audio, 1644 ms in audiovisual and 1672 ms in the visual condition.

ANOVA on slopes showed an effect of the condition: $F(2,46)=19.705$, $p=0.000$. There was no difference in slope between the auditory condition and the audiovisual condition ($F=1$). When we grouped the audio and audiovisual data, and these new slopes were compared with the visual ones, we obtained a significant difference: $F(1,23)=34.466$, $p=0.000$. The switch of identification is this time again comparable in audio and audiovisual conditions, but slower in visual identification.

Finally, to check the similarity between our first experiment and the control, we compared the results of these two groups of subjects. In the comparison of boundaries between groups 1 and 2, the group effect was non significant ($F=1$), just like group-condition interaction ($F<1$). The same for the slopes: no group effect nor group-condition interaction ($F=1$). Thus brings together the results of the two experiments, we obtained again a significant difference of the boundaries according to the condition of presentation: $F(2,98)=2.78$; $p=0.000$. We found that A and AV boundaries were different: $F(1,49)=151$; $p=0.000$; as well as the AV and V boundaries: $F(1,49)=177$; $p=0.000$. These 50% boundaries were at 1628 ms date in audio, 1647 ms in audiovisual and 1673 ms in visual, i.e. for A, AV and V conditions, respectively 92 ms, 73 ms and 47 ms before the acoustic onset of [y]. For the slopes, the difference was also significant, with $F(2,98)=23$; $p=0.000$. AV was different from V: $F(1,49)=31$; $p=0.000$. But A and AV were not different ($F=1$). Thus A and AV could be grouped together and compared with V: $F(1,49)=37$; $p=0.000$. This indicated that V slope was really slower than A and AV slopes. Finally, along these two experiments, we tested a population of 50 subjects finding the same pattern of results.

5. Conclusions and discussion

Our objective was to test monomodal and bimodal perception in the time course of the vocalic flow from [i] to [y], in the presence of an intervocalic [z] consonant consonant stream: that is why we chose a [zily] test stimulus.

(1) Results in audio (A) presentations evidenced an early perception of the rounding vowel into the preceding [z] consonant, as [y] perception started as soon as the end of the first quarter of the consonant: for a 120 ms [z] consonant, this occurred as early as 92 ms before its end. The audiovisual (AV) identification followed, with a delay of about 20 ms, while visual (V) perception was the last to come, showing a delay of 45 ms compared to the audio. Finally, the ranking of anticipation perception was 92 ms (A), 73 ms (AV) and 47 ms (V) before the acoustic onset of [y]. By comparison, Cathiard [7] found an advance of vision on hearing for a vocal produced word initially (without a consonant), and with a demarcating prosodic silent pause. In such a test case for the perception of the proper vocalic flow, it was chosen that the vowel information were not interacting with any particular consonantal stream. Hence it was predictable that vision would be earlier because the audio was naturally delivered later. In our [zily] case, articulatory information (the labial rounding constriction and protrusion) and acoustical information (move in the friction high energy noise) are delivered in synchrony: hence, the vocalic flow is available with the consonant flow. In such a case, the listener has to track the vowel identity as soon as possible throughout the consonant, using coarticulatory information, i.e. the evolution of zones of energy, from the first cue of a lowering move in the friction noise, until the third formant value of the [y] vowel is unambiguously available.

(2) We also showed that the audiovisual identification was temporally in-between the two monomodal presentations. We could compare this result to current interpretations of the McGurk illusion, which defend that an audiovisual percept is intermediate between the audio and visual stimuli. But we have to remind that we are not, in this experiment, in the case of conflictual information, but in the case of a congruent flow, which was proved to be recoverable earlier in a modality than in the other. We could think also that our result in the audiovisual modality contradicts the rule of the superiority of bimodal information on monomodal information: this rule is actually valid in a perturbed (noisy) situation, which is again not our case. We are thus in presence of a situation where, with a perfectly audible sound and a non contradictory visual flow, we observe nevertheless a 20 ms lag of audiovisual perception on auditory perception.

In conclusion, we have tested the building up of the perception of vocalic anticipation, when vocalic auditory information is carried by the noise of the preceding fricative. This coarticulatory audio information is perceived in advance on the visual one. And this audio advance is also effective, even when compared to bimodal audiovisual perception. Thus, the temporal evolution of bimodal information in speech strongly depends on the timing of the coordination of the articulatory gestures. With regard to vocalic information only, the consonant coarticulatoty audio flow can deliver the vowel identity, early in advance on the arrival of unambiguous formant cues for the vowel. In our fricative-vowel case, it was obvious that the vowel building up movement was audible throughout the fricative noise: so to say the lips could be heard along the noise. Whereas the changes in formant grouping, i.e. the main cue indicating the change from [i] to [y], occurred later (all that being readable on fig. 1).

Hence the time course of bimodal information in speech cannot be considered to display the same pattern whatever the timing of the coordination of speech gestures. As concerns vowel information only, consonantal coarticulation can carry earlier auditory information than the vowel itself, this depending on the structure of the stimulus. As concerns the timing of auditory and visual information, it becomes more and more obvious that the dates of the delivery of each component of the perceptual flow can change drastically the results obtained in the time course of brain activity for bimodal processing. Hence the flexibility of the phase relationships between the two flows urges to a better knowledge of the natural ecological variability in audiovisual signal production, a flexible coherence not to be violated on pain of breaking the Gestalt laws (see [2] again for such a claim).

6. Acknowledgements

We thank our speaker Jean-Luc Schwartz; and Alain Arnal, Hélène Loevenbruck, Solange Rossato and Christophe Savariaux for their technical assistance.

7. References


