Audio-visual speech scene analysis: Characterization of the dynamics of unbinding and rebinding the McGurk effect

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz

Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect

Olha Nahorna, Frédéric Berthommier & Jean-Luc Schwartz[(1)]

GIPSA-Lab, Speech and Cognition Department,

UMR 5216, CNRS – Grenoble University – France

Suggested running title: Binding dynamics in the McGurk effect

Abstract

While audiovisual interactions in speech perception have long been considered as automatic, recent data suggest that this is not the case. In a previous study, Nahorna et al. (2012) [J. Acoust. Soc. Am, **132**, 1061-1077] showed that the McGurk effect is reduced by a previous incoherent audiovisual context. This was interpreted as showing the existence of an audiovisual binding stage controlling the fusion process. Incoherence would produce unbinding and decrease the weight of the visual input in fusion. The present paper explores the audiovisual binding system to characterize its dynamics. A first experiment assesses the dynamics of unbinding, and shows that it is rapid: an incoherent context less than 0.5s long (typically one syllable) suffices to produce a maximal reduction in the McGurk effect. A second experiment tests the rebinding process, by presenting a short period of either coherent material or silence after the incoherent unbinding context. Coherence provides rebinding, with a recovery of the McGurk effect, while silence provides no rebinding and hence freezes the unbinding process. These experiments are interpreted in the framework of an audiovisual speech scene analysis process assessing the perceptual organization of an audiovisual speech input before decision takes place at a higher processing stage.

Suggested PACS Classification numbers

Main section: 43.71

Detailed classification: 43.71.An, 43.71.Es

Keywords**:** audiovisual speech perception; multisensory coherence; conditional binding; attentional mechanisms; audiovisual fusion

I. Introduction

A. The standard model of audiovisual fusion in speech perception

Audiovisual interactions in speech perception are generally described as an unconditional fusion process in the sense that (1) visual and auditory modalities would be translated into a common format and/or converge towards a given representational stage, where the entries would be merged in a way still to define, and (2) this merging process would be automatic, depending neither on the input stimuli nor on the context and in particular not on possible attentional effects. In other words, if I_A and I_V are respectively the auditory and visual inputs at time t, audiovisual perception would be described by the following process:

$$P_{AV}(t) = F(I_A, I_V) \quad \text{(Eq. 1)}$$

where $P_{AV}(t)$ is the percept at time t, and F is a fusion function whose output exclusively depends on inputs I_A and I_V.

This framework provided the basis for explaining the results of the two main paradigms for the study of audiovisual interactions: speech perception in noisy conditions, in which the visual input enhances the intelligibility of auditory input degraded by acoustic noise (Sumby and Pollack, 1954; Erber, 1969; Benoît et al. 1994); and the McGurk effect, in which two conflicting inputs (typically an audio “b” and a video “g”) are combined into a specific fused percept, typically "th" or "d" (McGurk and MacDonald, 1976) .

The literature in the 80s and 90s was mainly focused on specifying the nature of the F operator in (Eq. 1), and in particular on the two components of this operator: (1) the nature of the common representation towards which the auditory and visual inputs would converge before fusion, and (2) the mathematical content of the fusion operator.

The first question involved assumptions about auditory vs. motor recoding and the issue about early fusion (combination of sensory inputs recoded into a common pre-phonological format before decision occurs) vs. late fusion (separate classification of sensory inputs followed by a decision fusion process, operating in a common space of phonetic or phonological features): see reviews in Summerfield (1987) and Schwartz et al. (1998). Concerning the second question, Massaro's group extensively studied the fusion operator content. They proposed the Fuzzy-Logical Model of Perception (FLMP) and presented systematic comparison of possible operators competing with the optimal fusion operator realized by a multiplicative process in the FLMP (Massaro and Cohen, 1983; Massaro, 1987, 1989).

B. Non-automaticity of the fusion process

While the fusion process has long been considered as automatic (Massaro, 1987; Soto-Faraco et al., 2004), works in the 90s and 2000s displayed various departures from this hypothesis in several directions.

This began with the issue whether the fusion process might depend on the subject and especially her/his culture and language. The pioneer experiments by Sekiyama and Tohkura (1991, 1993) displayed lesser McGurk effect in Japanese compared to American English and generated many studies and much debate in the 90s (e.g. Massaro et al., 1993; Furster-Duran, 1996). It has however been obscured by methodological problems associated with model comparison in an audiovisual perception experiment, since it is difficult to disentangle what comes from unisensory perception (i.e. how subjects perceive each input independently of the other) and what is actually due to fusion. We recently showed how the use of a rigorous methodological framework for comparing models

(Schwartz, 2006) enables to confirm the existence of differences between subjects, some subjects giving more weight to one or the other modality independently on the input content (Schwartz, 2010). We can summarize this first point by assuming that the fusion process is actually of the form:

$$P_{AV}(t) = F(I_A, I_V, S) \quad \text{(Eq. 2)}$$

where S represents the subject with her/his own specificities, both individual ("auditory" vs. "visual subjects") and possibly cultural or linguistic (Sekiyama and Burnham, 2008).

The second direction was provided in the 2000s by experiments showing the potential role of attentional effects. In the "face-leaf" study by Tiippana et al. (2004), a visual distractor (a transparent leaf gently moving on the speaking face) superimposed on a conflicting audiovisual stimulus (such as seeing the face of a female speaker uttering "k", superimposed on a "p" sound) decreased the McGurk effect (with fewer fusion responses "t" and more auditory responses "p"). The authors' interpretation was that the participants attributed less weight to the visual modality in the fusion process because the leaf distracted their visual attention (see also Andersen et al., 2001). Once again, the use of a rigorous mathematical framework enabled to confirm this interpretation (Schwartz et al, 2010) by introducing an attentional factor in the fusion process. This could be formalized by the following equation:

$$P_{AV}(t) = F(I_A, I_V, S, A) \quad \text{(Eq. 3)}$$

where A represents a global attentional factor, modulated in the leaf-face experiment by the visual distractor reducing the weight of the I_V visual input in the fusion process.

Later, experiments by Soto-Faraco' group showed that an attentional load applied to the fusion process (consisting in superposing to the McGurk audiovisual speech perception

task an additional task involving the processing of other auditory, visual or tactile stimuli: Alsius et al., 2005, 2007) decreased the McGurk effect. The authors concluded that the fusion process was not automatic, but rather under the control of a global attentional process modulated by the attentional load. In the framework of (Eq. 3), it could be suggested that the attentional load factor is integrated inside the A term, resulting in a decrease of the weight of the I_V visual input in the fusion process.

The passage from (Eq. 1) to (Eq. 3) can be computationally implemented in various ways. We ourselves proposed an implementation based on the late-fusion multiplicative FLMP model where fusion only depends on the unisensory inputs, in accordance with Eq. 1. From that basis, we introduced a weighted fuzzy-logical model of perception, WFLMP, in which fusion would also involve specific weights controlling the role of each modality in the fusion process. This led to various implementation of the WFLMP, in which weights depend on the subject's individual characteristics (Schwartz, 2010; Huyse et al., 2013), attentional processes (Schwartz et al., 2010), or degradation of the auditory or visual input (Heckmann et al., 2002; Huyse et al., 2013).

C. Audio-Visual Speech Scene Analysis and the binding and fusion hypothesis

A remarkable point in the studies by Tiippana et al. (2004) and Alsius et al. (2005) is that the subjects were simultaneously processing multiple auditory or visual inputs (see also Andersen et al., 2009; Alsius and Soto-Faraco, 2011). Then a question arises: how do subjects succeed in segregating mixed sources in each unisensory flow before attempting to fuse the adequate pieces of information? This is the issue of perceptual scene analysis. The concept of auditory scene analysis (ASA) popularized by Bregman (1990) has largely renewed our understanding of auditory processing, gradually imposing a model in which

a perceptual organization stage should intervene in the auditory categorization process by specifying the different sources of information mixed in the scene before they could be efficiently identified. Auditory scene analysis involves segmenting the scene into sensory elements that should be grouped in respect to their common source, either by bottom-up innate primitives or by learnt top-down schemas. The way various primitives, likely detected in different auditory maps in the human brain, are grouped together to form a whole percept is generally called the *binding problem*.

A multisensory scene such as a mixture of audiovisual speech sources contains both acoustic and optic cues, likely resulting in auditory and visual primitives. The question addressed by our group since a number of years concerns whether audiovisual scenes, including multiple audiovisual speech streams, could involve an Audio-Visual Speech Scene Analysis process in which auditory and visual primitives would be adequately bound together before audiovisual fusion could occur. Studies in this area are rare, and the classical conception is rather that monosensory grouping precedes multisensory interactions, with a number of data in support of this view (Sanabria et al., 2005; Keetels et al., 2007). However, some data suggest that audiovisual interactions could intervene at various stages of the speech decoding process.

This includes the audiovisual speech detection advantage in which the presence of the speaker's face has been shown to improve the detection of speech embedded in acoustic noise (Grant and Seitz, 2000) and produce specific gains in intelligibility (Schwartz et al., 2004). The audiovisual speech detection advantage happens to operate independently of the possibility to understand speech, even in a foreign language (Kim and Davis, 2003) or with time-reverse speech. The temporal correlation between the auditory and visual components plays a crucial role in this process (Kim and Davis 2004). On the other way round, an auditory stimulus comodulated with the visual stimulus of a talking face

improves the visibility of the talking face masked by interocular suppression (Alsius and Munhall, 2013). In all these studies, it is suggested that audiovisual comodulation provides a binding process able to fuse together acoustic and optic cues, improving the detection of an audiovisual source or the extraction of audiovisual cues masked by auditory or visual noise.

Furthermore, electrophysiological experiments display early audiovisual interactions in the auditory cortex (Colin et al., 2002; Besle et al., 2004), showing that visual speech can speed up the cortical processing of the auditory input as soon as 100ms after the stimulus onset (van Wassenhove et al., 2005). Altogether, these data suggest that the visual speech flow could modulate ongoing auditory feature processing at various levels (Bernstein et al., 2004; Bernstein et al., 2008; Arnal et al., 2009; Eskelund et al., 2011).

This led Berthommier (2004) propose a two-stage model in which audiovisual coherence between the auditory and the visual input would be computed prior to fusion, to determine whether the two inputs are coherent and hence should be bound together and produce perceptual fusion. This binding and fusion process would consist in conditioning fusion on binding, just as Bregman reasoned that auditory perception should be conditioned by auditory binding thanks to an auditory scene analysis process. It may be described by an additional expansion of (Eq. 3):

$$P_{AV}(t) = F(I_A, I_V, S, A, C_{AV}) \quad \text{(Eq. 4)}$$

wherein C_{AV} represents an audiovisual coherence index enabling the subject estimate whether the auditory and visual inputs should be fused or not.

This assumption found an experimental support in a series of experiments that we conducted recently (Nahorna et al., 2012). In these experiments, we manipulated the

audiovisual coherence index C_{AV} by providing an audiovisual context prior to the McGurk target. The context was either coherent (auditory and visual inputs from the same source, namely a speaker producing a series of audiovisual syllables) or incoherent (auditory and visual input from two different sources, for example the sound of the speaker producing a sequence of acoustic syllables, dubbed on the image of the speaker producing a sequence of sentences unrelated with the sequence of acoustic syllables). There were two targets, one congruent (audiovisual "ba") and one incongruent (the McGurk target made of an auditory "ba" with a visual "ga"). The subject's task consisted in attempting to detect online "ba" or "da" syllables inside a film made of a series of such (context + target) sequences, without knowing when they would occur in the film. The online monitoring procedure aimed at emphasizing the role of audiovisual scene analysis processes, the assumption being that with incoherent context, the subject would unbind to a certain extent the auditory and visual streams and hence display less McGurk effect, with more "ba" and less "da" responses to McGurk targets. It appeared that the McGurk effect was indeed largely reduced in the incoherent context condition in respect with the coherent context condition.

We interpreted these results in the binding and fusion framework, by assuming that:

(1) Without context, the subjects would be in a default state where pieces of information are bound together, as it seems to be the case for auditory scene analysis (see e.g. Bregman & Pinker 1978), and also for visual scene analysis (Hupé and Pressnitzer, 2011). Therefore the auditory and visual inputs are supposedly coherent and hence bound together;

(2) Subjects would estimate the audiovisual coherence index C_{AV} by the context. In the incoherent context condition, this index suggests that sound and image should not be bound together, which would decrease the role of the visual input in the

fusion process and hence decrease the amount of McGurk responses. This was called by Nahorna et al. (2012) unbinding;

(3) In the coherent context condition on the contrary, the index would confirm that sound and image should be bound together, hence the subject would stay in the default state and display a stable McGurk effect.

D. Dynamics of the binding process in audiovisual speech scene analysis

We assume that the computation of the audiovisual coherence C_{AV} index is part of a general scene analysis process, generalizing Bregman' ASA to audiovisual scenes. We therefore consider that a major issue of current research on audiovisual fusion in speech perception is the characterization of this binding and fusion process, and more generally the understanding of what constitutes the audiovisual speech scene analysis system.

In this paper we capitalize on the "context + target" experimental paradigm developed by Nahorna et al. (2012) to focus on the dynamics of the binding-unbinding process, around two major questions.

1. Time constant of the unbinding process

The first one deals with the precise time constant of the unbinding process. The experiments in our previous work used rather long contextual stimuli, from around 3 s to around 10 s. It appeared that the amount of unbinding – displayed by the amount of decrease in the McGurk effect – was constant over this duration range. While McGurk stimuli in a coherent context were identified as "ba" 60% to 70% of the time and as "da" the remaining 40% to 30%, the application of an incoherent context decreased the amount of "da" responses to about the half of their value without context, independent of

context duration. This result was obtained for both a strongly incoherent context consisting in acoustic syllables dubbed on a completely different video material extracted from sequences of unscripted sentences, and for a phonetically incoherent context obtained by dubbing audio syllables on video syllables having a different phonetic value, while maintaining audiovisual synchrony.

It remains to be established what happens for smaller context durations. This is the objective of the first experiment in which we will assess the role of short incoherent contexts, from 0 to 3 seconds, to see what is the minimal duration of incoherence necessary for providing significant unbinding (as displayed by a significant decrease in the amount of the McGurk effect) and when does maximal unbinding occur.

2. Conditions for rebinding after unbinding

Supposing that the decrease in the McGurk effect produced by an incoherent audiovisual contextual stimulus is indeed due to an unbinding mechanism, a question is to know what kind of information is able to reset the system and put it back in its supposedly bound default state.

The objective of the second experiment in the present paper is to attempt to answer this question. For this aim, we will test whether applying a reset period of either coherent material or silence after the incoherent unbinding context would enable to recover the McGurk effect. The driving hypothesis of this experiment is the following: (1) the incoherent context alone should decrease the McGurk effect and hence increase the amount of "ba" responses; (2) the additional reset context, if it is efficient for rebinding, should result in recovering the McGurk effect (possibly with a cumulative effect, that is the amount of McGurk responses should increase for increasing durations of the reset stimulus, back to its initial value without context when reset is long enough).

II. Experiment 1: Time constant of the unbinding process

The first experiment aimed at estimating whether short incoherent audiovisual contexts could indeed modulate the McGurk effect and at assessing the role of context duration in the range corresponding to 0 to 3 seconds of incoherence. The paradigm was quite similar to the one used in Nahorna et al. (2012), consisting in online monitoring of congruent and incongruent McGurk targets embedded in a coherent or incoherent context. The general hypothesis was that incoherent contexts should decrease the amount of fusion responses "da" to McGurk targets, the experimental question being to know how this decrease would depend on context duration. Response times, which are seldom studied in audiovisual perception experiments, were also analyzed to assess how they would depend on the target and context.

A. Materials and Methods

1. Participants

20 subjects, French native speakers without any reported history of hearing disorders and with normal or corrected-to-normal vision participated in the experiment (4 women and 16 men, from 23 to 54 years old with mean 26.6, 19 right-handed and 1 left-handed). They all gave informed consent to participate in the experiment and were not aware of the purpose of the experiments.

2. Stimuli

Subjects were presented with audiovisual films consisting of an initial part called context followed by a second part called target (Figure 1). All stimuli were prepared from

audiovisual material produced by a French male speaker, JLS, with lips painted in blue to allow precise video analysis of lip movements (Lallouache, 1990). The videos consisted of the entire speaker's face, keeping natural colors apart from the blue make-up. Recordings were digitized at an acoustic sampling frequency of 44.1 kHz and a video sampling frequency of 50 Hz (25 images per second with two frames per image). All the stimuli that will be described here under are exactly the same as those in Nahorna et al. (2012), apart from smaller context durations in the present experiment compared with Experiments 1 and 2 in Nahorna et al. (2012).

The target was either a congruent audiovisual "ba" syllable, or an incongruent McGurk stimulus with an audio "ba" dubbed on a video "ga". To prepare incongruent "McGurk" stimuli, the auditory channel of videos finishing with a "ga" was edited by replacing the "ga" sound with a "ba" excerpt extracted from appropriate acoustic files. The "ba" sound was positioned exactly at the same temporal position as the "ga" sound. Synchronization was ensured by superposing temporal positions of the plosive burst at the onset of the target stimulus. Congruent audiovisual "ba" syllables should be perceived as "ba", while incongruent McGurk stimuli should often be perceived as "da" (McGurk and MacDonald, 1976). The focus was actually on McGurk targets and the congruent "ba" targets were only presented as controls.

There were three types of contexts in this experiment. The first type was coherent. It consisted in a series of 1 to 5 audiovisual syllables extracted from random sequences containing "pa", "ta", "va", "fa", "za", "sa", "ka", "ra", "la", "ja", "cha", "ma" or "na". The speaker was instructed to produce a short silence between consecutive syllables, which was necessary for further audio editing. The syllable rhythm was about 1.5 Hz, hence the context duration varied between 0.6 and 3 s depending on the number of uttered syllables.

303 The second type was called strongly incoherent. This context consisted of either 1,2,3,4
304 or 5 acoustic syllables dubbed on an equally long stretch of a video of a speaker saying
305 sentences.

306 The third type was called phonetically incoherent. It was obtained by swapping the audio
307 content from one syllable to the other – keeping exactly the same video material as in the
308 coherent context condition – while maintaining a precise synchrony in time between the
309 auditory and visible syllables, hence the term phonetically incoherent. To maximize
310 audio-visual incoherence, syllables were firstly organized in five groups known to be
311 visually rather distinguishable (visemes): “pa, ma”, “fa, va”, “ta, na, sa, za”, “cha, ja”
312 and “ka, la, ra, ga”. Then the audio content of each syllable was swapped with the
313 content of a syllable from a different group. For each syllable, care was taken to maintain
314 perfect synchrony between the sound and the image by dubbing the sound with the burst
315 onset at exactly the same position as the original sound. Again, context duration was
316 varied, such that the context consisted of either 1,2,3,4 or 5 audiovisual syllables.

317 As recalled in Section I.C.1, both sets of incoherent contexts have already been shown in
318 Nahorna et al. (2012) to produce a significant decrease in the McGurk effect for context
319 durations larger than 5 syllables (typically 3 seconds). Therefore the question in
320 Experiment 1 is to know what happens for smaller durations.

321 A fixed set of target stimuli (comprising “ba” and “McGurk” stimuli) was used all along
322 the experiment. McGurk stimuli were presented three times more than congruent stimuli,
323 which served as controls. There were 4 different “ba” targets and 12 different McGurk
324 targets, positioned at the end of each of the three sets of context sequences and for each of
325 the 5 context durations (all 12 McGurk tokens and 4 ba targets were used equally often in
326 each condition). To ensure continuity between the end of the context stimulus and the

onset of the target stimulus, a 200-ms transition stimulus (5 images without sound) was inserted between context and target (with a progressive linear shift from face to black from images 1 to 3, and a progressive linear shift from black to face from images 3 to 5). Fading is a "necessary evil" to be able to carefully control both contexts and targets, hence finding a way to stick together these two pieces of audiovisual material. It could potentially predict the occurrence of targets, but does so then for all conditions. This would in fact provide some reset ingredient potentially decreasing the role of incoherent contexts, hence we cannot dismiss the assumption that incoherence effects could be *underestimated* because of a possible resetting effect due to fading. Subjects however never complained that there was a perturbing discontinuity from context to target, discontinuity actually being very difficult to notice thanks to the dubbing procedure described above[(2)].

An additional set of stimuli consisting in targets without context (4 "ba" and 12 McGurk targets) was also presented. These stimuli, introduced to provide a kind of reference for evaluation of the role of context, were not contained in the experimental plan (with three contexts and five context durations) hence they had a special status in the statistical analyses (see later).

This provides altogether 256 stimuli: 3 contexts * 5 durations * (12 McGurk targets + 4 "ba" targets) + (12 McGurk targets +4 "ba" targets) without context. The 256 stimuli were concatenated into a single film, with a 840-ms inter-stimulus silent interval. The video component of this silent interval was made of the repetition of the last image of the previous stimulus. Such a short inter-stimulus interval was selected to put the subjects in a real monitoring task where there was large uncertainty about the temporal arrival of possible targets, to decrease as much as possible post-decision biases on target detection. A film was hence made of a random succession of coherent and incoherent contexts at all durations, and of targets without context (this was *not* a context-blocked experiment). All

acoustic files were globally normalized in intensity to ensure that they were presented at the same level. We prepared 5 different films with 5 different orders of the 256 stimuli (each film lasted about 15 minutes). Each subject was presented with one film, the 5 films being equally distributed between the 20 subjects (4 subjects per film).

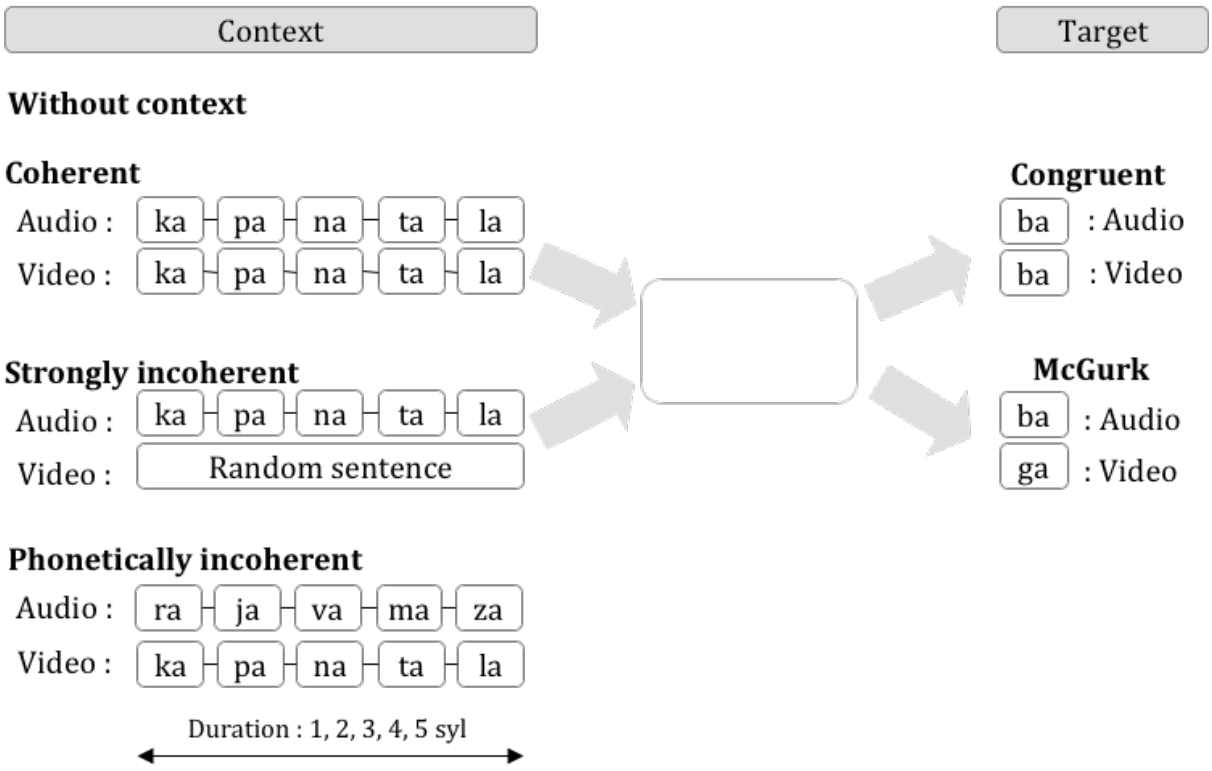


Figure 1 – Organization of stimuli in Experiment 1

3. Procedure

The subject's task was to detect online "ba" or "da" syllables (syllable monitoring task), without knowing when they could occur in the sequence. The experiment consisted of syllable monitoring with two possible responses – "ba" or "da" (responses entered on a keyboard, with one button for "ba" and one for "da", the order of buttons being equally

distributed across subjects). Therefore, subjects could provide responses at any time along the monitoring process.

The experiment was monitored by the Presentation® software (Version 0.70, www.neurobs.com). It was carried out in a soundproof booth with the sound presented through an earphone at a fixed level for all subjects, the level being adjusted to be comfortable for the task (around 60 dB Sound Pressure Level). The video stream was displayed on a screen at a rate of 25 images per second, the subject being positioned at about 50 cm from the screen. Instructions were to constantly look at the screen, and each time a "ba" or a "da" was perceived, to immediately press the corresponding button (displayed by the experimenter at the beginning of the experiment).

4. Processing of responses

The number of "ba" and "da" responses to the targets was computed for each subject and each condition. Since the task was syllable monitoring and the subjects did not know when the targets would occur, they could detect "ba" or "da" at any time and also fail to detect the target (failures either due to lack of response or multiple different responses to the target stimulus).

Analysis of response times enabled us to specify a protocol in which only responses within 1200 ms after the target syllable acoustic onset were considered (target onset was manually detected with the support of the MATLAB 7.6.0 software). This choice was constrained by the short inter-stimulus interval (840 ms): 1200 ms after the burst onset of the target stimulus was typically the onset time of the next stimulus. Furthermore, responses intervening less than 200 ms after the burst were also discarded (see e.g. Ratcliff & Rouder, 1998; van Maanen et al., 2012). In the case of two different responses inside this [200-1200] window, the responses were discarded. Altogether (that is adding

the number of misses or different responses to the target), this resulted in a total of 8.9% of cases with no response to a target stimulus. This amount is not surprising considering that the subjects only had two possible answers at their disposal while McGurk stimuli could result in percepts other than "ba" and "da" in French (Cathiard et al., 2001), and that they had less than 1.2 s to answer online. The number of no-response was actually larger for McGurk than for "ba" targets. Importantly, the amount of cases with no response was rather stable for McGurk targets across the three context conditions, varying between 9.3 and 11%, hence this protocol did not bias the following analyses.

Response time was defined as the time separating the plosive burst at the onset of the target stimulus and the response (within the 1200 ms cutoff) measured with the Presentation® software. For each (subject, target, context, duration) condition, the mean response time was estimated by averaging the response times for all stimuli in the corresponding condition.

5. Statistical analyses

Considering responses, analyses were performed on proportions of "ba" responses over the total number of "ba" plus "da" responses (ignoring cases where no response was provided by the subjects), after processing them with an asin(sqrt) transform to ensure quasi-Gaussian distribution of the variables involved. A systematic check was made that other analyses performed either on the proportions of "ba" responses over the total number of stimuli ("ba" plus "da" plus no response) or on the proportions of "da" responses over the total number of stimuli provided the same significant and non-significant effects. Since "ba" targets were only there as controls, the analysis of responses was focused on McGurk targets.

To quantitatively assess the comparative role of the three contexts and their five durations, a repeated-measures ANOVA was done on transformed proportions of "ba" responses for McGurk targets, with context (3 values) and context duration (5 values) as independent variables and subject as a random-effect factor. Greenhouse–Geisser correction was applied in case of violation of the sphericity assumption. When appropriate, we used post-hoc analyses of differences between two conditions with Bonferroni corrections, and reported differences as significant in case of a Bonferroni-corrected value $p<0.05$. Importantly, the data for targets without context were not considered in the ANOVA since they are not part of the experimental plan with 3 contexts and 5 context durations. However, since they were recorded to provide a reference, specific t-tests comparing the context conditions to this no-context condition have been conducted following the results of the ANOVA.

Considering mean response times per subject and condition, a repeated-measures ANOVA was performed on the logarithm of these values for ensuring normality of the distributions, with the same independent variables as previously. A repeated-measures ANOVA was done on logarithms of mean response times with target (2 values), context (3 values) and context duration (5 values) as independent variables and subject as a random-effect factor. Once again, the no-context condition was not introduced in these ANOVAs and rather played the role of a baseline for evaluating the role of context.

B. Results

1. Effect of context on the amount of "ba" responses

The results of the subjects' responses (proportion of "ba" responses relative to the total number of "ba" + "da" responses) for both targets in the three contexts and without context are set out in Figure 2. "ba" targets are classified as "ba" in all contexts with a score close to 100% (varying between 98.3% and 99% in the three contexts). McGurk targets produce a smaller proportion of "ba" responses, but this proportion is much larger in the strongly incoherent and slightly larger in the phonetically incoherent contexts than in the coherent context. The repeated-measures two-factor ANOVA on scores for McGurk targets shows that the effect of context is indeed significant [$F(2,38)=58.425$, $p<0.001$]). Post-hoc analysis confirms that the differences between the three contexts are significant. The increase in the proportion of "ba" responses to McGurk targets from the coherent (45%) to the strongly incoherent context (73%) is very large and corresponds actually to a reduction of the McGurk effect by half (from 55% of "da" responses with coherent context to 27% with strongly incoherent context). The difference is much smaller – though significant – with the phonetically incoherent context (10% increase in "ba" responses from 45% to 55%). Paired t-tests comparing either the target with coherent context or the target with phonetically incoherent context to the reference provided by the target without context provide no significant difference (without context compared to coherent context: 55% vs. 45%, [$t(19)=1.54$, $p>0.139$]; without context compared to phonetically incoherent context: 55% vs. 55%, [$t(19)=0.001$, $p=1$]).

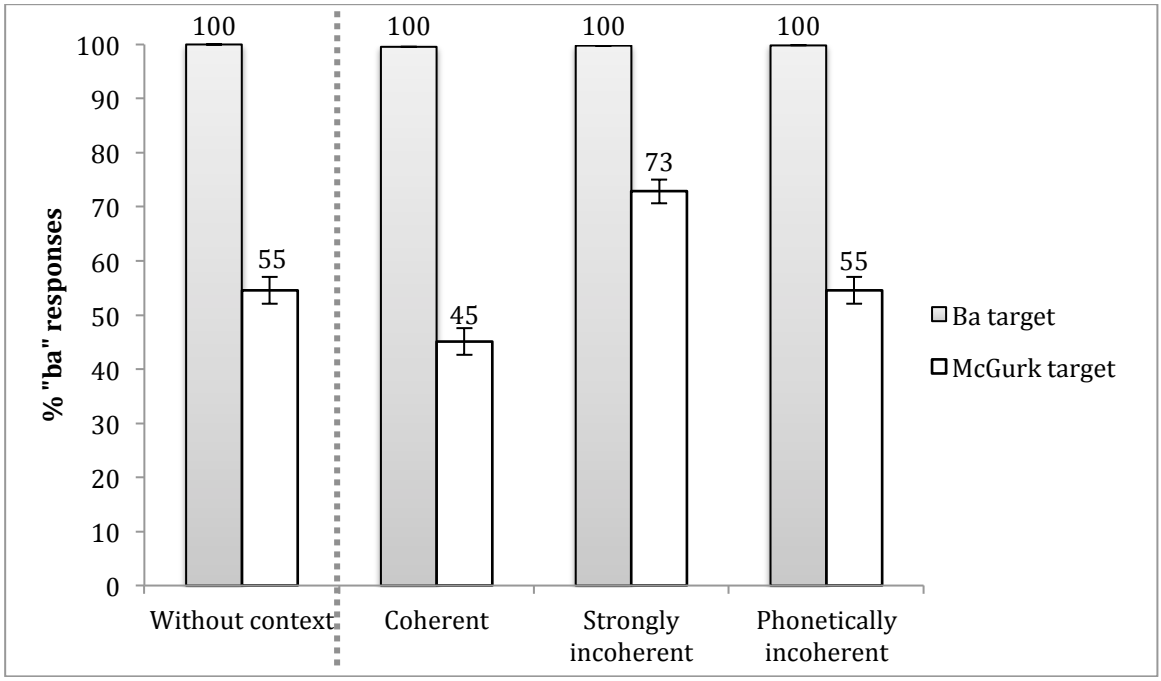


Figure 2 – Percentage of "ba" responses (relative to the total number of "ba" + "da" responses) for the two targets in the three contexts and without context.

2. Effect of context duration

Concerning durations, the ANOVA displays a main effect of the duration factor [$F(4,76)=5.44$, $p<0.001$] and a significant interaction with context [$F(8,152)=3.558$, $p<0.001$] (Fig. 3). Post-hoc analyses show that the duration factor is significant only for the strongly incoherent context. For this condition, the only significant differences are between durations 1 or 2 syllables on one hand and 4 syllables on the other hand. Globally, the trend for the strongly incoherent context is that the strong reduction of the McGurk effect is not only quick, complete as soon as one syllable of incoherent context, but even larger for the smallest context durations. We will propose possible interpretations of this unexpected fact later in the discussion. Concerning the phonetically incoherent context, since duration does not seem to matter, this suggests that the small

reduction of the McGurk effect with this context compared with the coherent context is rapid and complete for a one-syllable duration, as for the other incoherent context.

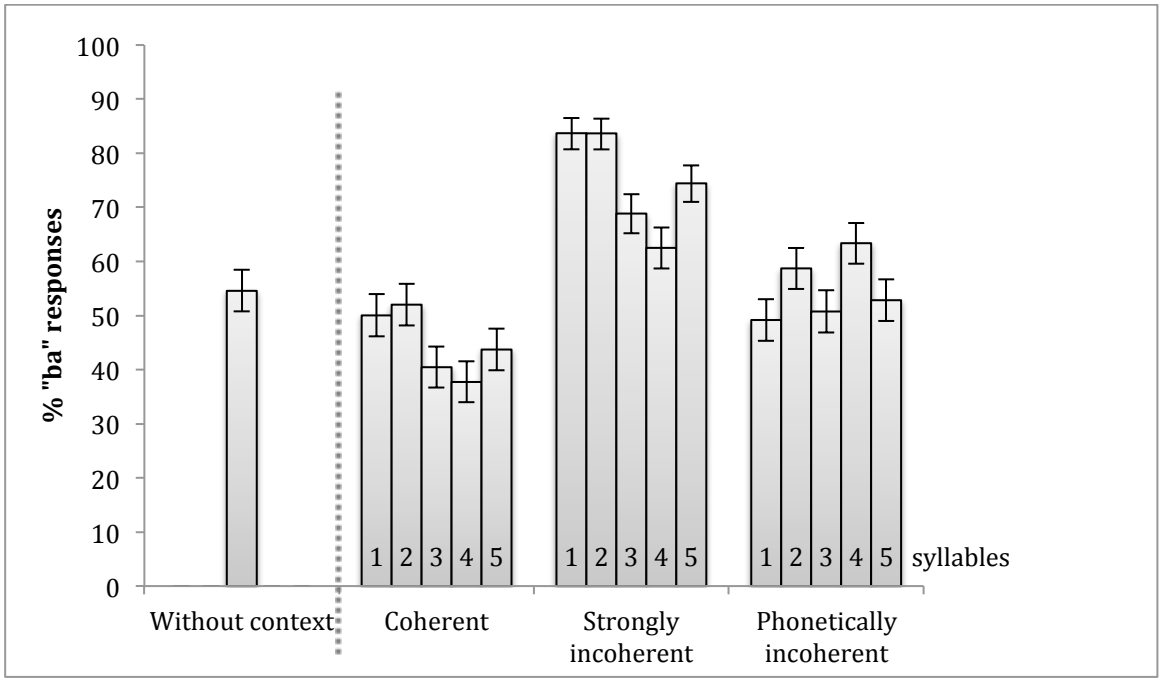


Figure 3 – Percentage of "ba" responses for McGurk targets for the three contexts and their five durations, compared to targets without context.

3. Contextual effects provided by previous stimuli

A possible problem in the previous analyses concerns the possibility that the response to a given stimulus may be influenced by the previous stimulus. This would produce possible spillover effects, e.g. the no-context condition would in fact be influenced by the previous coherent or incoherent contexts; or the coherent context condition would be contaminated by a previous stimulus with an incoherent context, etc. This question was already discussed in our previous study (Nahorna et al., 2012), and we will provide the same kind of analyses to evaluate this question. Firstly we performed a new repeated-

490 measures ANOVA on global scores (all context durations together) for McGurk targets,
491 with three factors: subject (random), context and preceding context (fixed). Notice that
492 although the set of target stimuli is of course the same from one context to the other, it
493 is not controlled for being the same from one previous context to the other, which
494 makes this analysis arguable. It appears that both the effects of context [$F(2,38)=51.192$,
495 $p<0.001$] and preceding context [$F(2,38)=4.252$, $p=0.022$] are significant, but not their
496 interaction [$F(4,76)=0.335$, $p=0.854$]. The significant effect of context corresponds to the
497 results presented previously (see Section II.B.1 and Fig. 2). The significant effect of
498 preceding context suggests that it plays a role in the binding and decision process, with a
499 mean 5.5% increase in "ba" responses (averaged over all McGurk targets for the three
500 contexts) from a preceding context which is coherent to a preceding context which is
501 strongly incoherent. The lack of significant interaction means that the effect of preceding
502 context is the same for all current contexts.

503 However, we reasoned in Nahorna et al. (2012) that another important bias could come
504 not from the previous *stimulus* but from the previous *response*. Indeed, if the preceding
505 context is strongly incoherent, the preceding response to McGurk targets is more often a
506 "ba". Might this play a role in the decision for the next McGurk target? Actually, this
507 should be the case, considering two classical response biases that are recalibration and
508 contrast (Bertelson et al., 2003; Vroomen and Baart, 2011). Recalibration effects appear
509 when subjects modify their categories – and hence their decisions – in relation with the
510 decision they took for previous stimuli. The possibility here would be that when a subject
511 categorizes a given McGurk stimulus as "ba" (respectively "da"), there is an increased
512 chance that the next McGurk stimulus will stay perceived as "ba" (respectively "da").
513 Contrast effects appear when the response to a stimulus in a given category C1

(contrasted to another category C2) is more likely to be "C1" if the preceding stimulus was in category C2 than if it was in category C1.

These two kinds of effects were indeed clearly displayed in the data analyzed by Nahorna et al. (2012). The same phenomenon appears in the present study, as can be seen on Fig. 4 where we report the scores for McGurk targets depending on context, preceding context and preceding response (incoherent context in this figure is the strongly incoherent context: we do not present results for phonetically incoherent context to make the figure clearer). On this figure, we observe the difference between the coherent and incoherent contexts with more "ba" responses in the second case (the "ba" score increases when comparing the first set of 3 bars with the second one, or the third one with the fourth one). However, there is in each case a large modulation depending on the preceding stimulus and response. Indeed, for each set of 3 bars (that is for each configuration of precedent context and present context) there is a recalibration effect with a much larger "ba" score when the precedent target was a McGurk target with "ba" response, compared with the "ba" score when the precedent target was a McGurk target with "da" response. There is also probably a contrast effect with a decrease in "ba" responses when the previous target was a "ba" compared to when it was a McGurk target with "ba" response – though it is not easy to disentangle contrast from recalibration.

Of course, since the preceding context modifies the amount of "ba" responses to the McGurk targets, the induced response biases may explain the effect of preceding context displayed in the ANOVA. Actually, the size of recalibration effects (50% or more in Fig. 4) is much larger than the size of the global effect due to the preceding context. Once the previous decision is taken into account, if we compare the first set of three bars with the third one or the second one with the fourth one in Fig. 4, we notice that in most cases the amount of "ba" responses is in fact *higher* when the preceding context is coherent

compared with when it is incoherent. Therefore altogether, we may consider that the present results are not contaminated – or at most very weakly – by the context of a previous stimulus, though they are subject to classical contrast and recalibration phenomena providing some decision biases. It might appear surprising that context effects are more or less restricted to one target and seem more or less "reset" when the next stimulus is presented: we will come back on this point in the General Discussion (Section IV.3).

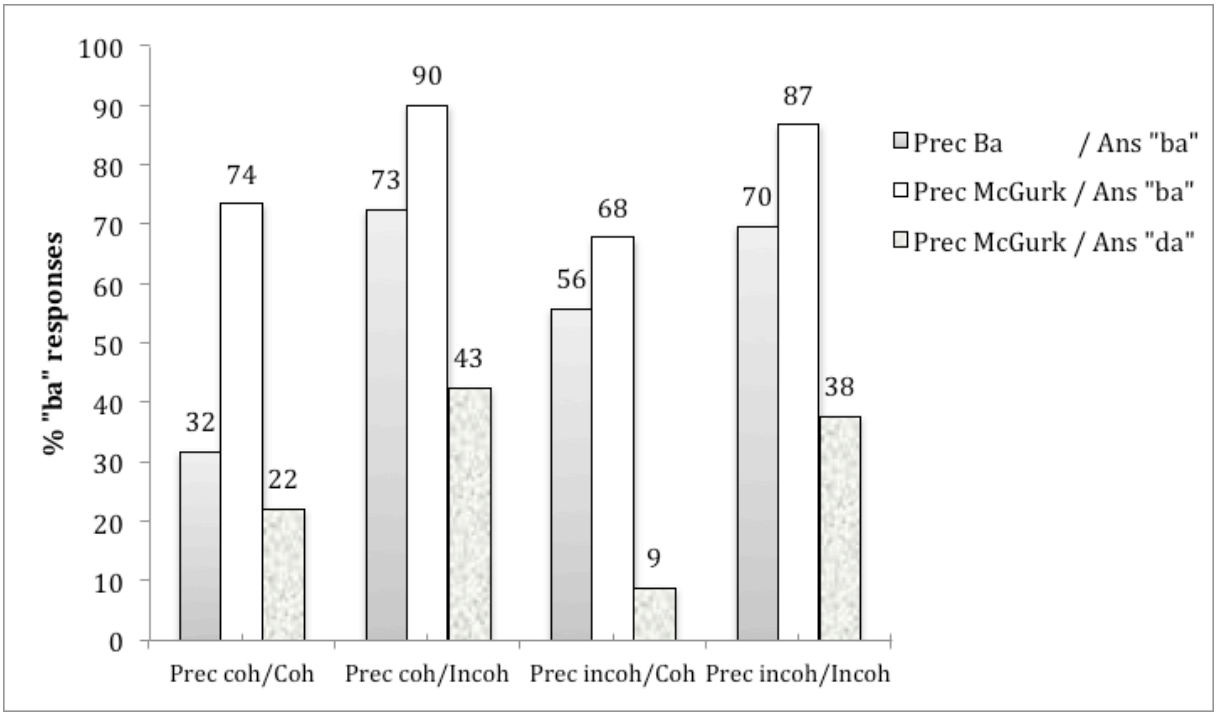


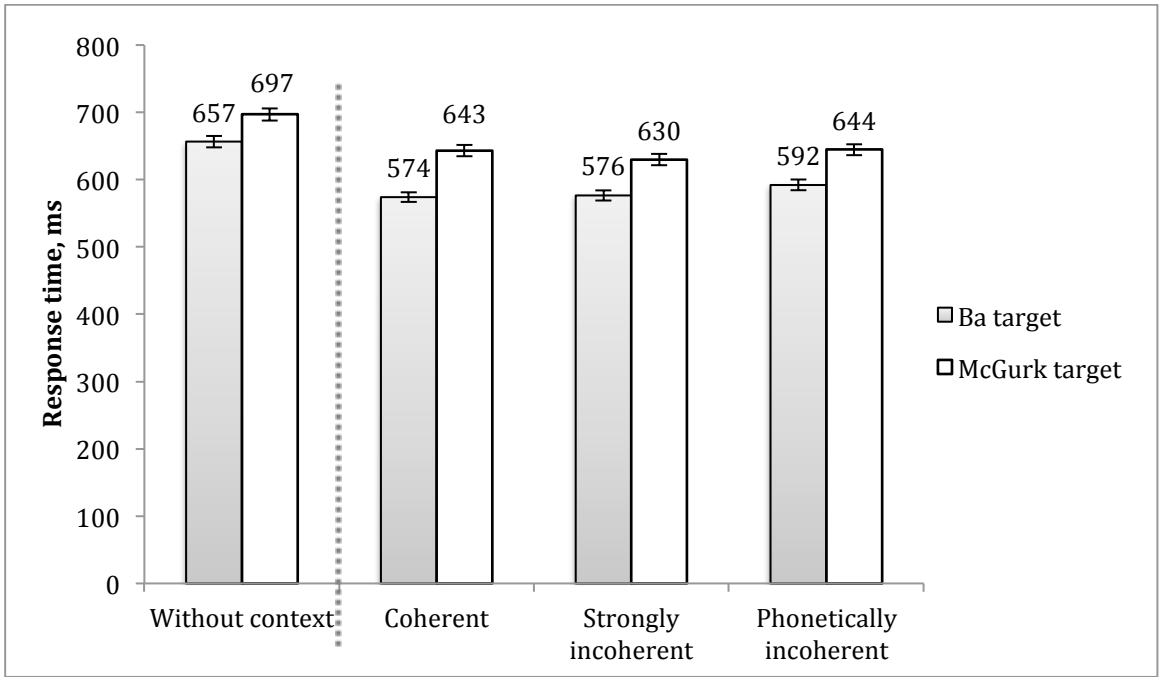
Figure 4 – Effect of the preceding decision in Experiment 1. Responses to McGurk stimuli depending on context ("Coh" for coherent, "Incoh" for incoherent), preceding context ("Prec coh" for coherent preceding context, "Prec incoh" for incoherent preceding context), preceding target stimulus ("Prec ba" vs "Prec McGurk") and previous answer ("Ans ba" for previous "ba" target, "Ans ba" and "Ans da" for previous

553 "McGurk" target). Incoherent context in this figure is the strongly incoherent context: we
554 do not present results for phonetically incoherent context to make the figure clearer.

555

556 ***4. Analysis of response times***

557 Mean response times for both targets in the three contexts and without context are set out
558 in Figure 5. Response times appear to be globally larger without context, and not
559 different from one context to the other. Response times are also systematically larger for
560 McGurk targets. These trends are confirmed by the three-way ANOVA. There is a
561 significant effect of target [$F(1,19)=28.52$, $p<0.001$], with a 58.3 ms difference between
562 mean response times for "ba" and McGurk targets. There is no effect of context, either
563 alone or in interaction with any other factor.



564
565

566 **Figure 5** – Mean response times for the two targets
567 in the three contexts and without context.

568

There is also a significant effect of context duration [$F(4,76)=3.41$, $p<0.03$] and of the interaction between target and context duration [$F(4,76)=4.16$, $p<0.004$]. The effect of duration is displayed in Figure 6. It appears a global trend in which response time decreases with context duration, from no context to 5 syllables. Post-hoc analyses display significant differences between response times (averaged over "ba" and McGurk targets) at 1 vs. 2 and 3 syllables. The effect of duration could be due to the fact that context enables the subjects to prepare the arrival of the target stimulus and hence respond more quickly when it finally arrives. This could explain the trend for having larger response times without context (Figure 5).

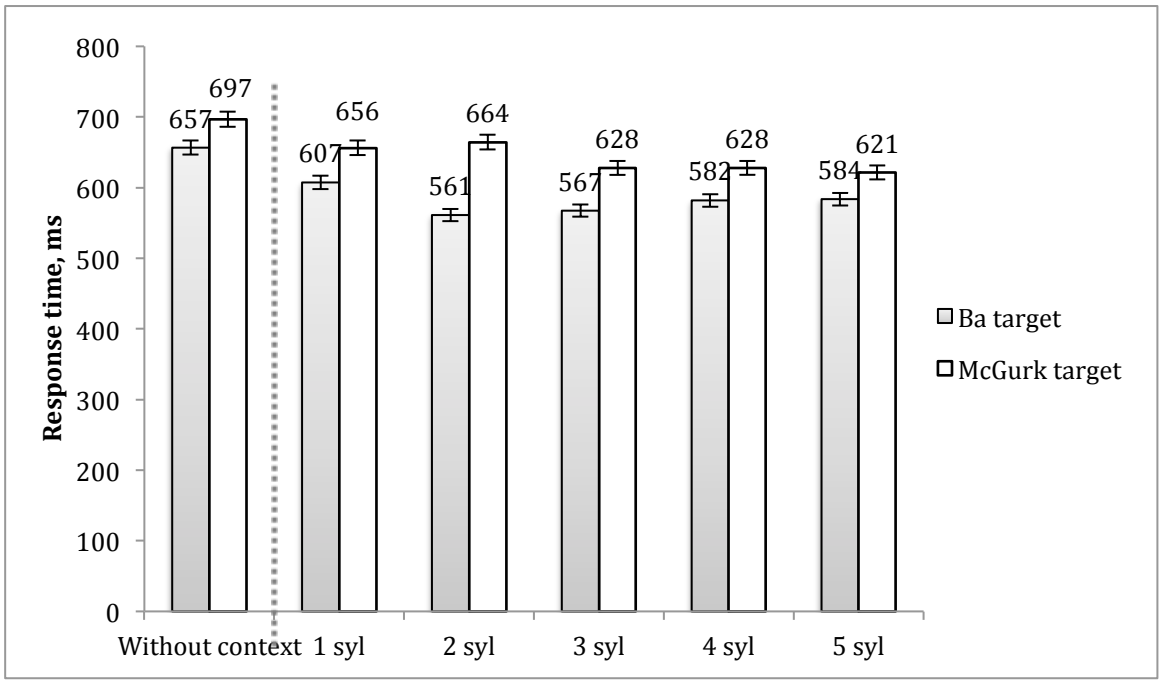


Figure 6 – Mean response times for the two targets in the five context durations and without context.

C. Discussion

Four major facts emerge from this experiment. Firstly, the present data confirm those obtained in the princeps study by Nahorna et al. (2012): various kinds of incoherent audiovisual contexts decrease the strength of the McGurk effect and increase the amount of auditory responses to McGurk targets. For strongly incoherent contexts the size of the reduction in the McGurk effect is similar in the present data and in the previous ones by Nahorna et al. (2012): typically a reduction by half. For phonetically incoherent contexts the size is much smaller, though significant: while there was also a reduction of the McGurk effect by half compared with the coherent context in the princeps paper (see Experiment 2 in Nahorna et al., 2012) it is much smaller here (55% ba" responses with phonetically incoherent context vs. 45% for coherent context, see Figure 2). This is likely due to the fact that both incoherent contexts were presented in the same experiment here while they were studied in two separate experiments in the previous study. This seems to have induced a kind of calibration process for subjects of the present study, in which the size of incoherence is compared from one stimulus to another. However the present data confirm that pure phonetic audiovisual incoherence keeping a perfect audiovisual synchrony allows some amount of unbinding between sound and image when compared with coherent context. But they show that this is only a small part of the total amount of incoherence available in the strongly incoherent context: hence the corresponding amount of decrease in the McGurk effect is much smaller for pure phonetic incoherence.

Secondly, we now have a clear confirmation that the unbinding effect is rapid. One syllable seems to suffice to produce an effect as large as the effect of five syllables – and in

Nahorna et al. (2012) there was no difference between 5 and 20 syllables. Hence it seems that unbinding is almost complete with a small duration of incoherence (around 600 ms), typically one syllable. This appears to be the case for both contexts. For the strongly incoherent context, there is even a trend that small durations (1 or 2 syllables) produce a larger decrease in the McGurk effect than larger ones (4 syllables). This is rather counterintuitive. It could be due to non-monotonous contrast effects in the computation of audiovisual coherence (with a kind of incoherence adaptation effect that would increase the size of perceived incoherence at the first time when some incoherence is perceived). It could also be related with the increase in response times for short contexts compared to longer ones (see Figure 6). Indeed, this could be taken as an indicator that the subject is surprised by the arrival of the target for short contexts, and that surprise could lead to decreased fusion, considering the audiovisual integration has been shown to falter under high attention demands (Alsius et al., 2005).

The third point concerns the nature of the default state. Our hypothesis was that without context subjects would be in a default state of binding. The mere fact that the McGurk effect exists shows that there is indeed a certain amount of binding without context. It remains to be known if binding is maximal in the default state The fact that there is no significant difference between the no-context and coherent context conditions and no effect of context duration for the coherent context condition suggests that this might be the case. This is further supported by the results from our previous study, where we found no effect of context duration from 5 to 20 syllables. However, since the phonetically incoherent context also displays no significant difference with the no-context condition, we cannot dismiss the possibility that there would be in fact no unbinding effect of the phonetically incoherent context compared with the no context condition (the default state) and some increase of the amount of binding when a coherent context is applied to

633 the default state. The (non significant) 10% decrease of the "ba" percentage from the no-
634 context condition to the coherent context condition (see Fig. 2), together with the (non
635 significant) decrease trend of the "ba" score in the coherent context when context
636 duration increases from 1 to 5 syllables (see Fig. 3) might call for further experiments to
637 test this assumption. Let us conclude, to summarize the discussion of this third point,
638 that the default state (without context), which we will still consider as "bound" since it
639 displays a certain amount of audiovisual integration, is perhaps not maximally bound;
640 and that the possible increase in binding that could be produced by a coherent context, if
641 it exists, does not seem very large.

642 The last important finding in Experiment 1 is that response times are consistently larger
643 for McGurk targets than for congruent "ba" targets independently on the effects of
644 context (Figure 5). This is rather striking considering the size of context effects on the
645 scores of "ba" responses. Indeed, it is classically considered that response times in such
646 experiments rely heavily on the ambiguity of the stimulus to process (Massaro and
647 Cohen, 2003). In the present case, the ambiguity in McGurk targets is largely reduced by
648 the very incoherent context: while these targets are identified close to 50% as "ba"
649 (actually 45% "ba" vs. 55% "da") in the coherent context, they are perceived as 73% as
650 "ba" in the very incoherent context (see Figure 2). However this does not result in any
651 significant change in response times: context seems to modify the response but not the
652 response time. This suggests that the increase in response times for McGurk stimuli is
653 due, at least partly, to the detection of a local audiovisual incoherence, which seems to
654 slower the response independently on the response itself. We will come back to this point
655 in the general discussion.

656

III. Experiment 2: Testing the existence of a rebinding process

The results of Experiment 1 clearly show that an incoherent context results in a decrease of the McGurk effect, which is due in our interpretation to an unbinding mechanism. The objective of Experiment 2 is to know what kind of information is able to reset the system and put it back in its bound default state (recalling the previous discussion about the fact that the default state is not necessarily "maximally bound"), that is enhance the McGurk effect again so that it recovers the level it has with no contextual stimulus before the McGurk target.

A. Materials and Methods

1. Participants

20 French subjects without hearing or vision problems participated in the experiment (9 women and 11 men, from 18 to 60 years old, mean 25.7, 19 right-handed and 1 left-handed). They all gave informed consent to participate in the experiment, and were not aware of the purpose of the experiments.

2. Stimuli

The stimuli, described in Figure 7, consisted in a succession of three components (with a 5-images fading between consecutive stimuli as in Experiment 1):

- A *context* which could be either coherent or "strongly incoherent" in the sense of Experiment 1. Therefore we discarded phonetically incoherent context in this

experiment, to focus on the two most extreme variants that are coherent and strongly incoherent. In the following of Experiment 2, incoherent will refer to the strongly incoherent type of context. Considering the results of Experiment 1 showing no influence of context duration for coherent context, and a small significant difference between small (1 or 2 syllables) and large (4 syllables) durations for strongly incoherent contexts, we used only 2-syllable and 4-syllable durations;

- a *reset* stimulus consisting in either 0, 1, 2 or 3 coherent audiovisual syllables (*coherent reset*) or audio silence with fixed image of duration 0, 480, 1000, 1480 ms corresponding roughly to the same duration as the 0-, 1-, 2- or 3-syllable coherent reset condition (*fixed reset*). The reset was inserted only after incoherent contexts: coherent contexts were followed directly by the target, and used only as controls in this experiment. Notice that the "0-syllable reset" conditions actually mean no reset at all, and that these conditions are of course the same for both the coherent reset and the fixed reset, though it was necessary to introduce both conditions to ensure a full factorial design;
- and finally a *target* which could be, as in Experiment 1, either a congruent audiovisual "ba" or a McGurk stimulus consisting in an audio "ba" dubbed on a video "ga". As in Experiment 1, McGurk targets were presented three times more than congruent "ba" targets, which served as controls.

Stimuli were presented to participants in two blocks, one block with coherent reset and the other one with fixed reset. Each block comprised stimuli with either the coherent context (with 2 possible durations) with no reset, or the incoherent context (2 possible durations) followed by the reset (4 possible durations). Hence there were altogether 10 conditions per block, with 4 different occurrences of a "ba" target and 12 different

occurrences of a McGurk target per condition, with a total of 160 stimuli in a block, presented in a random order and organized in a film as in Experiment 1, with the same 840-ms inter-stimulus interval. The order of blocks was randomized between the 20 subjects with 10 subjects per order.

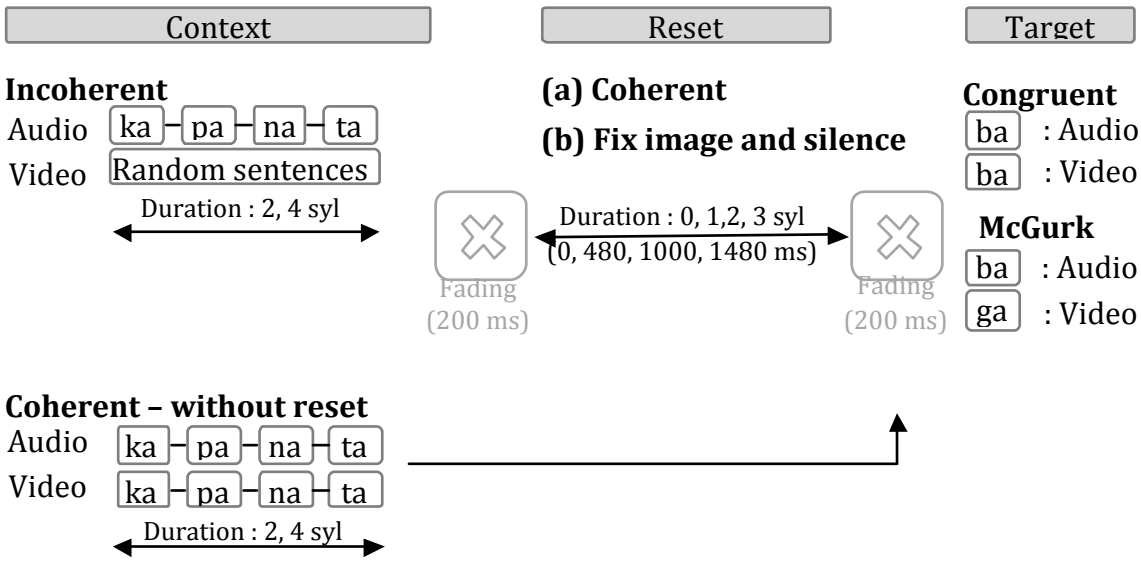


Figure 7 – Organization of stimuli in Experiment 2.

3. Procedure, processing of responses and statistical analyses

Procedure and response processing were exactly the same as in Experiment 1. The number of missing responses in this experiment (still with the [200-1200 ms] cut off procedure) was less than in Experiment 1 (7.6%), Once again however, the amount of cases with no response for McGurk targets was rather stable across the two reset conditions, varying between 7 and 9.4%.

Statistical analyses were performed on the same variables as in Experiment 1: for each subject and condition, proportions of "ba" responses over the total number of "ba" plus "da" responses processed with an asin(sqrt) transform, and logarithm of mean response times. Only the stimuli with incoherent context plus reset were submitted to repeated-measures ANOVAs, the stimuli with coherent context without reset being only considered as a baseline over which unbinding and rebinding were evaluated.

B. Results

1. Analysis of "ba" responses

As in Experiment 1, the "ba" target leads to 100% "ba" responses in both experiments and in all conditions. Therefore, as planned, we will concentrate on McGurk targets. A repeated-measures three-factors ANOVA on scores for McGurk targets with factors context duration (2 vs. 4 syllables), reset type (fixed vs. coherent) and reset duration (0, 1, 2 or 3 syllables) shows the following results.

The effect of context duration is significant [$F(1,19)=18.89$, $p<0.001$]. The shorter context with 2 syllables produces in average a percentage of "ba" responses 5.4% larger (that is a smaller McGurk effect) than the longer context with 4 syllables. There is no interaction between context duration and any other variable, hence this effect is stable for all reset conditions, whatever the reset type and duration.

The effects of reset type and reset duration are displayed on Figure 8. The effects of reset type [$F(1,19)=5.097$, $p=0.036$], reset duration [$F(3,57)=12.64$, $p<0.001$], and the interaction between reset type and reset duration [$F(3,57)=11.699$, $p<0.001$] are all significant. Actually, three major facts emerge from Figure 8.

- *Unbinding with incoherent context*. Let us first look at what happens for the incoherent context without reset, corresponding to the 0-syl condition (left bars, for both types of resets). The score of "ba" responses is around 75-80%, much larger than the score for the coherent context condition (rightmost bars), which is less than 50%. This replicates the decrease of McGurk effect from coherent (more than 50% McGurk effect) to incoherent context (less than 25% McGurk effect) displayed in Experiment 1.

- *Poor rebinding with fixed reset*. Looking at the bars corresponding to the fixed reset condition on Figure 8, it appears that this reset (made of acoustic silence + fixed image) provides almost no rebinding, since the "ba" score only slightly decreases from 0 to 1-syl (that is 480ms duration), then remains stable and stays much larger than the score for coherent context even for the longest reset duration (3-syl corresponding to 1480 ms). Post-hoc analyses confirm the initial small decrease in "ba" responses, since there is a significant difference between scores at 0 and 2 syllables. However, a t-test confirms that the score at 3 syllables (74%) is significantly different from the score with coherent context (46%): $t(19)=5.22$, $p<0.001$.

- *Good rebinding with coherent reset*. On the contrary, looking at the bars corresponding to the coherent reset condition, we observe that the "ba" score regularly decreases with reset duration and reaches the same value as for coherent context, coming back to its default state for the largest coherence period of 3 syllables. Post-hoc analyses confirm that the score at 0 is significantly higher than with 1, 2 or 3 syllables, and the score at 1 or 2 syllables is significantly higher than with 3 syllables. A t-test confirms that the score at 3 syllables (43%) is not different from the score with coherent context (45%): $t(19)=0.624$, $p=0.54$.

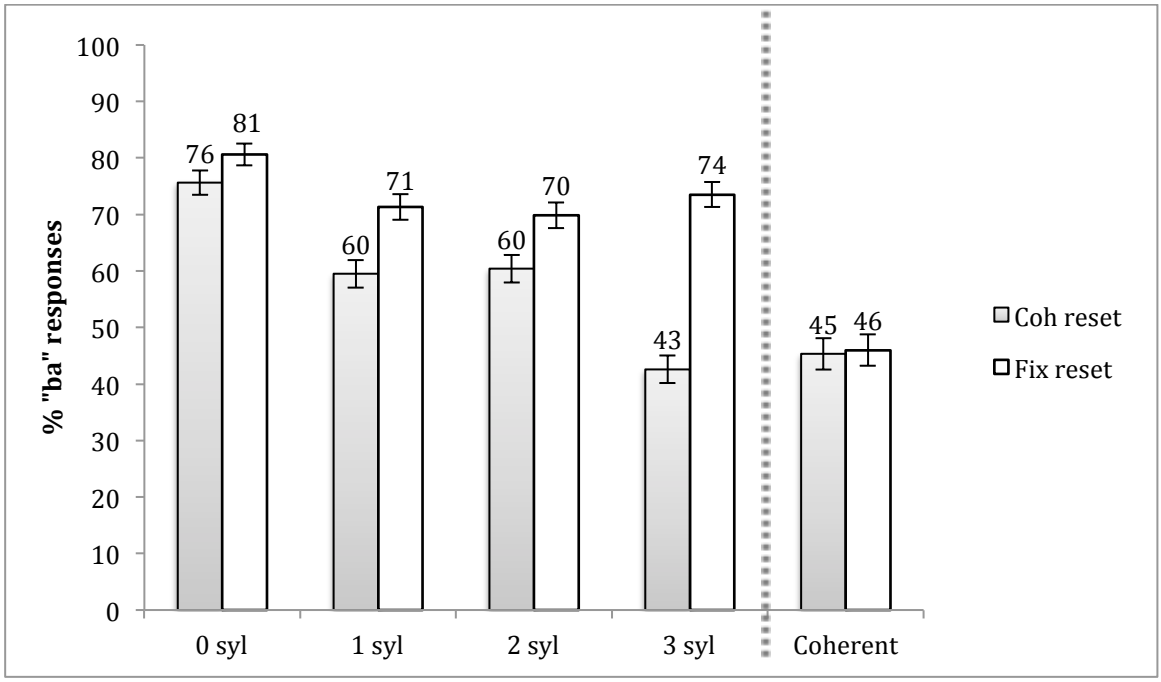
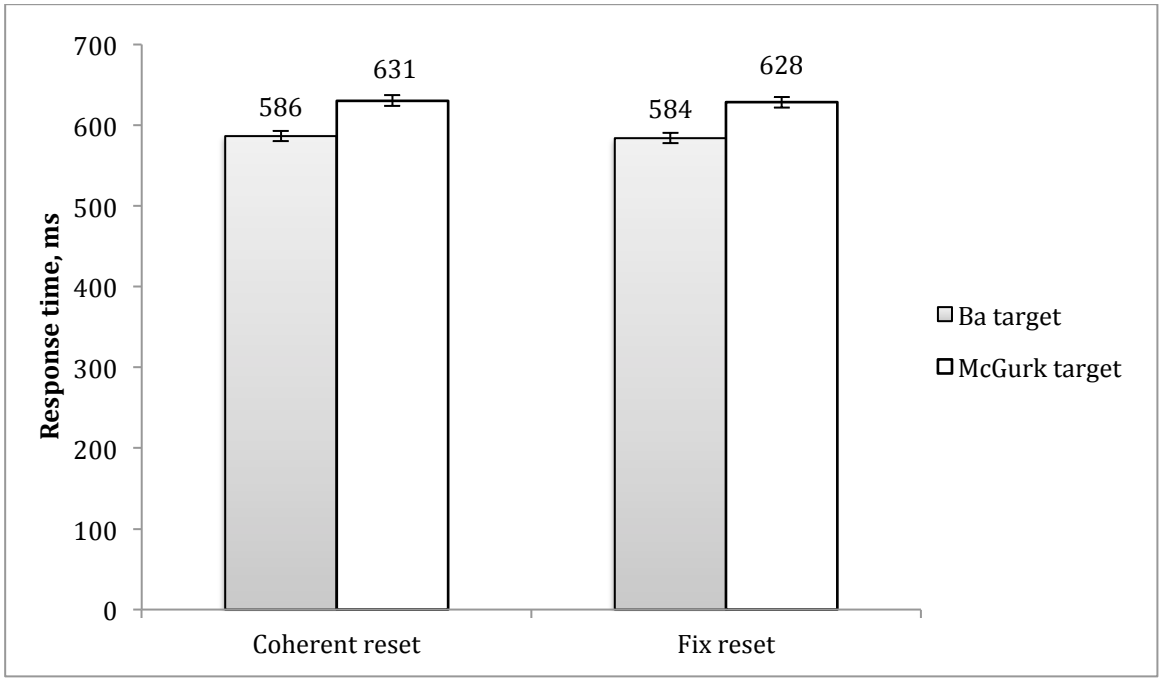


Figure 8 – Percentage of "ba" responses (relative to the total number of "ba" + "da" responses) for the McGurk targets with coherent context and with incoherent context for the two reset types and the four reset durations.

2. Analysis of response times

Mean response times for both targets in the two reset conditions are displayed in Figure 9. Response times are once again larger for McGurk targets. A two-way repeated-measures ANOVA on target and reset type shows an effect of target ([$F(1,19)= 29.57$, $p<0.001$]; difference between mean response times for "ba" and McGurk targets: 49.5 ms) but no effect of reset, alone or in interaction with target.



781

782 **Figure 9** – Mean response times for the two targets

783 in the two reset conditions.

784

785 ## C. Discussion

786 This experiment firstly confirms the amount of unbinding provided by the incoherent
787 context (corresponding to the strongly incoherent context in Experiment 1), which
788 produces a relative reduction of the McGurk effect by more than half. There is also a
789 confirmation that short incoherent contexts (2 syllables) produce a larger decrease in the
790 McGurk effect than longer ones (4 syllables), with a significant increase in the score of
791 "ba" responses around 5.4% in the first case. The fact that this increase is not dependent
792 on the reset duration (from 0 to 3 syllables) renders less plausible our interpretation in
793 Experiment 1 about the possible role of surprise since this should lead to differences
794 between short resets where the target comes rather quickly for the short context and long
795 resets where surprise is more unlikely.

The major new result of Experiment 2 is that after an incoherent context decreasing the McGurk effect, a coherent reset stimulus may increase it again until the original McGurk level is recovered. However, while the decrease is rapid in Experiment 1 with a maximum decrease already obtained for a one-syllable long context, the recovery appears slower in Experiment 2, not complete before 3 coherent syllables are presented. On the contrary, the other type of reset material composed of acoustic silence and fixed image does not allow to recover the original McGurk effect: the level of McGurk responses after a 2-syllable or 4-syllable period of incoherence remains remarkably stable at a low value after a period of fixed reset up to 1.5 s (see Figure 8).

Finally, this experiment provides a confirmation concerning the pattern of response times. Indeed, it appears (Figure 9) that response times are consistently longer for McGurk targets than for congruent "ba" targets independently on the effects of reset. This happens in spite of the strong effects of reset type and reset duration on the scores of "ba" responses: reset modifies the response but not the response time. This confirms that response times are not completely predictable from the ambiguity of the stimulus to process.

IV. General Discussion

The two experiments presented in this paper confirm that context modulates the McGurk effect in a principled way, and provide a number of quantitative data about the dynamics of this process. In the following, we will first discuss how these results fit inside the binding and fusion architecture that we propose in the framework of audiovisual speech

818 scene analysis. Then we will attempt to formalize this architecture in more detail, and
819 propose some elements of a cognitive model, to let emerge some open questions.

820 A. Characterization of the binding system in audiovisual speech
821 perception

822 *1. How context intervenes in audiovisual fusion*

823 The two experiments in this paper confirm the results of the two experiments presented in
824 our first study (Nahorna et al., 2012): the McGurk effect is not automatic, it depends on
825 the context provided by a sequence of audiovisual speech stimuli presented prior to the
826 McGurk target. Incoherent contexts of various types and durations decrease the amount
827 of fusion responses "da" in favor of auditory responses "ba", compared to coherent
828 contexts. This shows that there must exit in the audiovisual speech perception system a
829 device assessing audiovisual coherence and probably computing an audiovisual
830 coherence index of some kind: let us call this device a coherence box.

831 This coherence box is likely to be instrumental in the audiovisual speech detection
832 advantage (see Section I). Indeed, this advantage increases with the correlation between
833 visual cues (e.g. lip area or mouth opening) and audio cues (e.g. spectral features or
834 amplitude) (e.g. Grant and Seitz, 2000. Kim and Davis, 2004). It could also provide the
835 basis for audiovisual predictions, that is enable some predictions about the auditory
836 stream from the visual input, which has been proposed to be the basis for early
837 audiovisual interactions in evoked response potentials (e.g. van Wassenhove et al., 2005:
838 Arnal et al., 2009). We assume more generally that the computation of audiovisual
839 coherence index is a basic component in the audiovisual speech scene analysis system.
840 This index would enable the brain to evaluate the coherence between auditory and visual

features in a complex multi-speaker scene, in order to properly associate the adequate components inside a coherent audiovisual speech source. This is requested in a number of experimental paradigms testing audiovisual speech perception in a scene associating various faces and and/or various sounds (e.g. Andersen et al., 2009; Alsius and Soto-Faraco, 2011).

It remains to understand *how* does this coherence box intervene in the decision process leading to a given amount of fusion percepts in the present experiment. This is an open question. Since this box supposedly enables the brain to know which auditory and visual components must be associated to provide a fused percept (this is the *binding* problem), our assumption is that a low coherence index provides low evidence for fusion and hence decreases the visual weight in fusion, hence the increase in the amount of "ba" responses for incoherent contexts in Experiment 1.

It could also be envisioned that context in these experiments intervenes as a post-perceptual decision bias, according to which participants would be biased in their decision to not report a fusion response when they receive evidence about an audiovisual mismatch (provided by the context)[(3)]. However, the individual data show that the decrease in fusions is not of an all or none type. For example, we observed that in Experiment 1, most subjects display an increase in the amount of "ba" responses in the strongly incoherent context whatever their score in the coherent context condition. Therefore the decision bias would obey complex quantitative rules, not so different from a decrease in visual weight in a decision fusion process. Anyway, the global conclusion at this stage is that (1) a coherence index seems to be evaluated by the subject, and (2) its value seems to modulate the subject's decision in some way. This is captured by the formula proposed in Eq. (4) in the Introduction, and it is globally compatible with the binding and fusion architecture: binding is realized by the coherence box through the

computation of the audiovisual coherence index, and fusion, modulated by this index, provides the subject's final decision.

2. The dynamics of unbinding and rebinding

Experiments 1 and 2 confirm the study by Nahorna et al. (2012) showing that McGurk fusion depends on the previous audiovisual context. Our interpretation is that the incoherence of the audio and video streams leads the subjects to selectively decrease the role of the visual input in the fusion process. The general hypothesis is that modulation is driven by the output of a binding stage integrating information about the coherence of the auditory and visual input.

We begin to characterize the binding stage in the present paper. Firstly, Experiment 1 shows that the dynamics of unbinding is rapid. One syllable or the equivalent duration (around 0.5 s) suffices to produce a maximum decrease in the McGurk effect (around 50% decrease). There even appears a trend, confirmed in Experiment 2, according to which short durations of incoherence produce more unbinding than longer ones. The interpretation of this fact is not completely clear. It could be due to a kind of adaptation effect according to which the computation of coherence would include temporal derivatives, enhancing the incoherence index at the beginning of an incoherent sequence.

Experiment 1 also confirms that pure phonetic incoherence suffices to produce an effect on binding, since there is a difference between a coherent and a phonetically incoherent context – with a significantly smaller McGurk effect in the second case. This means that audiovisual correlations in time between audio and visual cues are probably not the single elements that intervene in the assessment of audiovisual coherence, and that the phonetic content of the incoming information also plays a part in this process.

Experiment 2 shows that unbinding processes can be followed by rebinding processes, in which coherent reset sets back the weight of the visual input and hence enables to recover the McGurk effect. However, rebinding appears slower than unbinding, since it requires at least 3 coherent syllables (for a duration around 1.5 s) to be complete. The interpretation seems to be that loosing faith in the common origin of the sound and face seems rapid, but recovering faith implies to gather a minimum amount of new coherent cues, which takes a longer time for accumulation of adequate information.

3. Binding states and reset processes

It is classically considered that auditory scene analysis involves a default grouped state followed by a possible build-up of auditory segregation (Bregman, 1990). The systematic bias towards the grouped interpretation is displayed both in the auditory and in the visual modality (Hupé and Pressnitzer, 2012). In the case of multisensory scenes, a general compatibility bias is displayed in various experiments dealing with the fusion of conflicting cues (e.g. Yu et al., 2009; Noppeney et al., 2010). This bias suggests that subjects suppose at the beginning of the task that the various cues are not conflicting before evidence of conflict progressively leads the subjects to select one cue rather than the other.

The present data are consistent with the hypothesis of a default state of the audiovisual binding mechanism in which audio and video components are fused together. Various evidence point towards this hypothesis. Firstly the existence of the McGurk effect itself seems to require this assumption. Indeed, McGurk stimuli are just a specific case of phonetic incoherence, not different from those used in Experiment 1. The fact that they can be fused together implies that subjects process these stimuli under the underlying

assumption of a default state. Notice that this underlying assumption is strong enough to resist to a number of incongruence in the components of the sensory streams: discrepancies in the spatial localisation of the auditory and visual sources (Bertelson et al., 1994), temporal asynchronies (van Wassenhove et al., 2007), and even incoherence of source identity, with a female face dubbed on a male voice (Green et al., 1991).

However, as we discussed at the end of Experiment 1 (Section II.C), our data do not allow to know for sure whether binding is maximal with no context (and hence cannot be increased by applying a coherent context, whatever its duration), or if it is actually sub-optimal, in which case coherent context could increase the confidence that the auditory and visual streams refer to a single source and hence the visual input would play a larger role in the decision process. A challenge for future studies will be to better understand how the evaluation of audiovisual coherence, and hence the amount of binding and the weight of the visual input, are constantly updated along the flow of audiovisual information.

A striking result of Experiment 2 is that a fixed reset has almost no rebinding effect, with the consequence that even for the longest duration (around 1.5s) the subjects stay frozen in an unbound state where the McGurk effect is largely decreased. It remains to study how the subjects come back to their default bound state. The fact that the influence of one stimulus on the next one seems rather weak (see Section II.B.3) makes us wonder whether giving a response also resets the system. However, as discussed in that section, there are too many confounding factors (associated to recalibration and contrast mechanisms producing decision biases), which impede to answer to this question at this stage.

A reset material should engage the subject into the understanding that the situation has

dramatically changed. This could involve changing from one speaker to another, assessing whether a piece of incoherent context from one speaker would modify the McGurk effect for another speaker. Another question deals with the speech-specific nature of the audiovisual binding system, asking whether for example an incoherent audiovisual context made of non-speech material would be as efficient as the kind of incoherent context used in the present study to reduce the McGurk effect.

4. Response is global, response time seems local

Reaction times to McGurk stimuli are seldom reported. When data are provided, they display longer reaction times for incongruent (McGurk) stimuli compared to congruent ones (e.g. Massaro and Cohen, 1983; Keane et al., 2010). Globally, there is a trend for having longer reaction times for incongruent than for congruent audiovisual stimuli (see a review in Tiippana et al., 2011). However, there are two possible interpretations of this fact. Firstly, ambiguity in categorical judgment classically increases response latency in a binary choice, and this is also in line with models of perceptual decision (e.g. Ratcliff and Rouder, 1998; Smith and Ratcliff, 2004). Since incongruence generally results in more ambiguous decisions, this should lead to longer response times. Secondly, it could also be proposed that subjects are slower to respond to the extent that the auditory and visual information give conflicting information about the speech event. These two assumptions were discussed by Massaro and Cohen (1983), with the conclusion that perceptual ambiguity was a better predictor of response times.

The results of the two experiments in this paper show that response times differ between congruent “ba” and incongruent McGurk targets but do not depend on context. In Experiment 1, response times are 58.3 ms larger for McGurk targets with no significant effect of context type and duration, though responses vary between 50% “ba” for

961 coherent context up to more than 80% "ba" for strongly incoherent context at the
962 smallest durations (1 or 2 syllables; see Figure 3). In Experiment 2, response times are
963 49.5 ms larger for McGurk targets with no significant effect of reset type and duration,
964 though responses vary once again between 50% and 80 "ba" depending on the reset
965 condition (see Figure 8).

966 Therefore, the present data suggest that ambiguity is not the sole determinant of response
967 times for McGurk stimuli embedded in the various contextual environments that we used
968 here. Indeed, while responses are modulated by context and hence appear as the product
969 of a global computation where both context (including reset) and target play a role,
970 response times appear as mainly governed by the local characteristics of the target, with
971 quicker responses for congruent compared to incongruent targets.
972

973 ## B. Elements of a cognitive model

974 The various elements summarized in the previous section may be encapsulated within a
975 tentative cognitive architecture displayed in Figure 10. This architecture has no ambition
976 to be definitive or complete, it simply aims at making clear some basic components that
977 emerge from both the first study by Nahorna et al. (2012) and the present one. This
978 architecture comprises the following element, that we progressively define starting from
979 the standard model of Section I.

980 • Audiovisual fusion for decision. The links between auditory and visual inputs and
981 the decision box provide the basic architecture in all audiovisual fusion models
982 since thirty years. Restricting the architecture to this box provides the basis for Eq.
983 (1).

- Attentional processes and individual specificities. Fusion appears to depend on individual and cultural/linguistic factors and attentional processes. Adding the corresponding arrow towards the fusion box provides the basis for Eq. (3).

- Coherence C(t). Our experimental results on the role of context suggest that the brain constantly evaluates the coherence of the auditory and visual inputs to determine whether they belong to a coherent source. This participates in our view to a general audiovisual scene analysis process in which subjects determine in a complex scene which parts of the auditory information must be associated with which parts of the visual information. We recalled in Section I.B a number of natural candidates for the computation of coherence C(t) that could be based on computations of correlation or mutual information between such cues as global envelope or envelope in specific spectral bands for the audio input, and lip or face parameter cues for the visual input. The fact that phonetic incoherence suffices to modulate the McGurk effect suggests that phonetic cues also participate to the computation of local coherence C(t). The bidirectional arrows in Figure 10 between the auditory and visual boxes on one hand and coherence C(t) on the other hand indicate that C(t) may also provide some feedback enabling better extraction of monosensory cues, as displayed by data on the audio-visual speech detection advantage (Grant and Seitz, 2000; Schwartz et al., 2004).

- Binding state. Our results also suggest that coherence enables to constantly monitor the binding state in the subject's brain, and that the binding state would play a role in the fusion-decision process: the less bound the binding state, the smaller the weight of vision in the fusion process. There seems to exist a default state which is bound to a certain extent, but it remains to know if coherent context

may drive towards a state which would be "more bound" than the default state. If we continue in the view that the binding state may vary on a quantitative scale between less and more bound, quantitative data in the present study suggest that the time constant towards less bound is more rapid than towards more bound. It would be around one syllable (less than 0.5s) in the first case, and around three syllables (more than 1s) in the second case. Interestingly, a previous work by our team on audiovisual speech source separation based on statistical modeling of audiovisual coherence showed that 400 ms suffice to adequately associate one audio stream and one video stream in a mixture of two faces and voices (Sodoyer et al., 2004). This confirms that there is enough information in less than 0.5 s to determine if a sound and a face may be bound together or not. Last but not least, results of Experiment 2 show that once the system is put in an unbound state by incoherent audiovisual material, it may stay frozen in this state for a while (at least 1.5 s) unless new evidence for coherence is provided. Altogether, the coherence and binding state boxes and the way they enter the fusion box provide the basis for Eq. (4).

- Response time (RT). While it is classically considered that response times mainly depend on the decision process, with larger response times for more ambiguous stimuli, the present study suggests that local coherence also plays a role in response times. Local incongruence in McGurk targets would be detected by the subjects and slower their response. This is in line with various studies in which it appears that subjects are both able to perceive and estimate the discrepancy between the sight and the sound of a speaking face and fuse the two inputs into a single percept (Manuel et al., 1989; Summerfield and McGrath, 1984; Soto-Faraco and Alsius, 2007, 2009). This suggests that the subjects have conscious

access to the output of the coherence box, C(t). Hence response times in our schema depend on both the decision process and the output of the local coherence computation process.

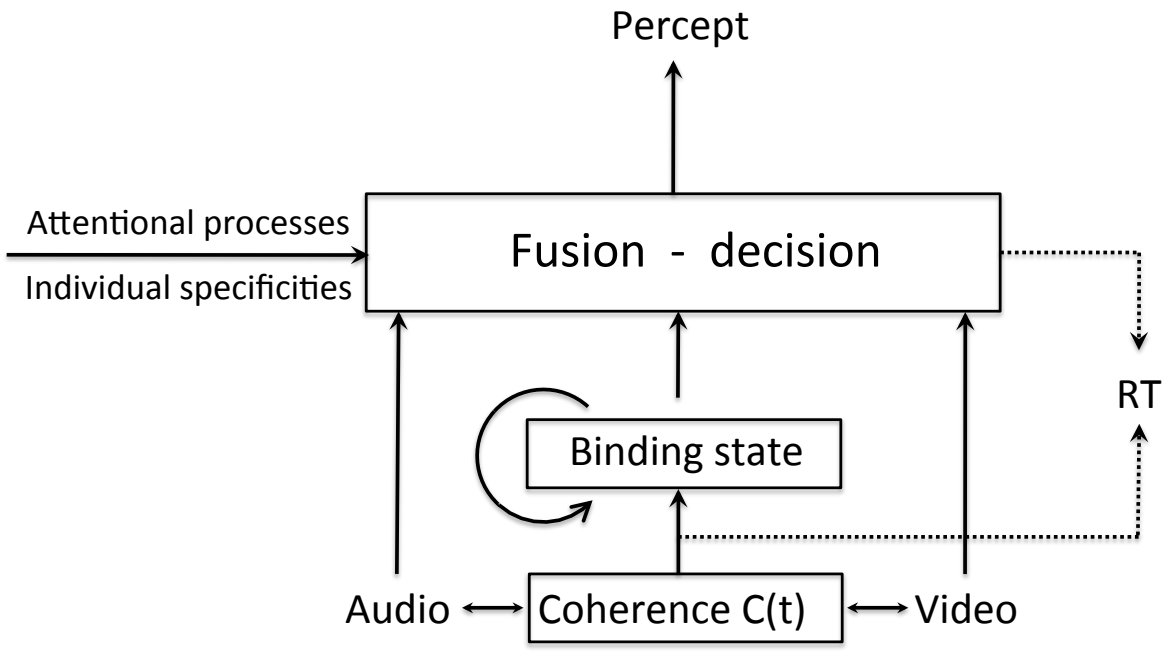


Figure 10 – A possible cognitive architecture for audiovisual binding and fusion in speech perception.

V. Conclusion

This set of experiments confirms that context may modify the McGurk effect, through a series of mechanisms, which combine unbinding (through incoherent context decreasing the role of the visual input) and rebinding (through coherent reset setting back the weight of the visual input). A first experiment displayed rapid unbinding

effects, with a reduction of the McGurk effect by half for very short incoherent contexts, made of one acoustic syllable dubbed on incoherent visual material extracted from the production of free sentences. A smaller incoherence amount, in which the phonetic content of the audio and video streams are different while keeping a perfect synchrony between the dynamics of sound and lips, resulted in a smaller but significant reduction of the McGurk effect compared with coherent context.

A second experiment tested the role of possible reset stimuli after a period of incoherence producing strong unbinding. It showed that a fixed reset (acoustic silence plus fixed image of the speaker's face) has almost no rebinding effect, with the consequence that even for the longer duration (around 1.5s) the subjects stay frozen in an unbound state where the McGurk effect is largely decreased. On the contrary, a coherent reset of 3 syllables is enough to completely recover from unbinding and restore the default binding stage.

Altogether these data can be captured inside a two-stage cognitive architecture in which a first binding stage assessing the coherence between sound and face would control the output of the fusion process and accordingly change the nature of the percept. Unbinding would result in a smaller role of vision in the decision process. Major challenges will involve a better understanding of possible binding states in the human's brain, in terms of online dynamics, neural correlates and changes in relation with age and hearing status.

Acknowledgments

This work was supported by the French National Research Agency (ANR) through funding for the MULTISTAP project (MULTISTability and binding in Audition and sPeech: ANR-08-BLAN-0167 MULTISTAP). The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152).

Endnotes

(1) Corresponding author (jean-luc.schwartz@gipsa-lab.grenoble-inp.fr)

(2) Examples of stimuli for Experiments 1 and 2 are available at http://www.gipsa-lab.grenoble-inp.fr/~jean-luc.schwartz/fichiers_public_JLS/AV_Binding_demo/AV_Binding_Demo.html

(3) We thank one of the reviewers for having suggested this possible interpretation of our data.

References

Alsius, A., & Munhall, K. (**2013**). "Detection of audiovisual speech correspondences without visual awareness," Psychological Science**24**, 423-31.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S.S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Current Biology **15**, 839–843.

Alsius, A., Navarra, J. & Soto-Faraco, S. (**2007**). "Attention to touch weakens audiovisual speech integration," Experimental Brain Research **183**, 399-404.

Alsius, A., & Soto-Faraco S. (**2011**). "Searching for audiovisual correspondence in multiple speaker scenarios," Experimental Brain Research **213**, 175-183.

Andersen, T.S., Tiippana, K., Lampinen, J. and Sams, M. (**2001**). "Modelling of Audiovisual Speech Perception in Noise," Proceedings of the Fourth International ESCA ETRW Conference on Auditory-Visual Speech Processing, Ålborg, Denmark, pp. 172-176.

Andersen, T.S., Tiippana, K., Laarni, J., Kojo I., & Sams, M. (**2009**). "The role of visual spatial attention in audiovisual speech perception," Speech Communication **51**,184-193.

Arnal, L.H., Morillon, B., Kell, C.A. & Giraud, A.-L. (**2009**). "Dual neural routing of visual facilitation in speech processing," Journal of Neuroscience **29**, 13445–13453

1108 Benoit, C., Mohamadi, T. & Kandel, S., (**1994**). "Effects of phonetic context on audio-
1109 visual intelligibility of French," Journal of Speech and Hearing Research, **37**, 1195-
1110 1203.

1111 Bernstein, L. E., Auer, E. T., & Moore, J. K. (**2004**). "Audiovisual speech binding:
1112 convergence or association?," in G.A. Calvert, C. Spence C, & B.E. Stein (eds.) *The*
1113 *handbook of multisensory processes* (pp 203–224). Cambridge: The MIT Press.

1114 Bernstein, L.E., Lu, Z.L., & Jiang, J. (**2008**). "Quantified acoustic-optical speech signal
1115 incongruity identifies cortical sites of audiovisual speech processing," Brain Research
1116 **1242**, 172–184.

1117 Bertelson, P., Vroomen, J., De Gelder, B. (**2003)**. "Visual recalibration of auditory
1118 speech identification: a McGurk aftereffect," Psychological Science **14**, 592–597.

1119 Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (**1994**). "Exploring the
1120 relation between McGurk interference and ventriloquism," in Proc. ICSLP 94 (Vol.
1121 2, pp. 559–562). Yokohama: Acoustical Society of Japan.

1122 Berthommier, F. (**2004**). "A phonetically neutral model of the low-level audiovisual
1123 interaction," Speech Communication ***44***, 31-41.

1124 Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (**2004**). "Bimodal speech: early
1125 suppressive visual effects in human auditory cortex," European Journal of
1126 Neuroscience **20**, 2225-2234.

1127 Bregman, A. S. (**1990**). *Auditory scene analysis* (773 p.), MIT Press: Cambridge, MA.

1128 Bregman, A.S. & Pinker, S. (**1978**). "Auditory streaming and the building of timbre,"
1129 Canadian Journal of Psychology **32**, 19-31.

1130 Cathiard, M.A., Schwartz, J.L., & Abry, C. (**2001**). "Asking a naive question about the
1131 McGurk Effect: why does audio [b] give more [d] percepts with visual [g] than with
1132 visual [d]?," *Proceedings AVSP-2001*, 138-142.

1133 Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (**2002**).
1134 "Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic
1135 representation within short-term memory," Clinical Neurophysiology **113**, 495–506.

1136 Erber, N.P. (**1969**). "Interaction of audition and vision in the recognition of oral speech
1137 stimuli," Journal of Speech and Hearing Research **12**, 423-425.

1138 Eskelund, K., Tuomainen, J., & Andersen, T. S. (**2011**). Multistage audiovisual
1139 integration of speech: dissociating identification and detection," Experimental Brain
1140 Research **208**, 447-57.

1141 Fuster-Duran, A. (**1995**). "McGurk effect in Spanish and German listeners. Influences of
1142 visual cues in the perception of Spanish and German conflicting audio-visual
1143 stimuli," in Proceedings of the Eurospeech 95, pp. 295–298.

1144 Grant, K. W., & Seitz, P. (**2000**). "The use of visible speech cues for improving auditory
1145 detection of spoken sentences," Journal of the Acoustical Society of America **108**,
1146 1197–1208.

1147 Green, K., Kuhl, P., Meltzoff, A., & Stevens, E. (**1991**). "Integrating speech information
1148 across talkers, gender, and sensory modality: female faces and male voices in the
1149 McGurk effect," Perception and Psychophysics **50**, 524-536.

1150 Heckmann, M., Kroschel, K., Savariaux, C., Berthommier, F. (**2002**). DCT-Based video
1151 features for audio-visual speech recognition. In: Proc. ICSLP02, Denver, pp. 1925–
1152 1928.

1153 Hupé, J.M., & Pressnitzer, D. (**2012**). "The initial phase of auditory and visual scene
1154 analysis," Philosophical Transactions of the Royal Society B **367**, 942-953.

1155 Huyse, A., Berthommier, F. et Leybaert, J. (**2013**). "Degradation of labial information
1156 modifies audiovisual speech perception in cochlear-implanted children," Ear and
1157 Hearing **34**, 110-121.

1158 Keetels, M., Stekelenburg, J., & Vroomen, J. (**2007**). "Auditory grouping occurs prior to
1159 intersensory pairing: Evidence from temporal ventriloquism," Experimental Brain
1160 Research**180**, 449-456.

1161 Keane, B. P., Rosenthal, O., Chun, N. H. & Shams, L. (**2010**). "Audiovisual integration
1162 in high functioning adults with autism," Research in Autism Spectrum Disorders **4**,
1163 276–289.

1164 Kim, J., Davis, C. (**2003**). "Hearing foreign voices: does knowing what is said affect
1165 masked visual speech detection," Perception **32**, 111–120.

1166 Kim, J., & Davis, C. (**2004**). "Investigating the audio-visual detection advantage,"
1167 Speech Communication **44**, 19-30.

1168 Lallouache, M.T. (1990). « Un poste 'visage-parole'. Acquisition et traitement de
1169 contours labiaux. (A "face-speech" workstation. Acquisition and processing of labial
1170 contours)," Proceedings XVIII Journées d'Etudes sur la Parole (pp. 282-286),
1171 Montréal.

van Maanen, L., Grasman, R.P.P.P., Forstmann, B.U., & Wagenmakers, E-J. (**2012**), "Piéron's Law and optimal behavior in perceptual decision-making," Frontiers in Decision Neuroscience **5**, 143.

Manuel, S., Repp, B. H., Liberman, A. M., & Studdert-Kennedy, M. (**1989**). "Exploring the "McGurk effect"," *Paper presented at the 24th meeting of the Psychonomic Society*, San Diego.

Massaro, D. W. (**1989**). "Multiple Book Review of *Speech Perception byEar and Eye: A Paradigm for Psychological Inquiry*," Behavioral and Brain Sciences**12**, 741–794.

Massaro, D. W. (**1987**). "*Speech perception by ear and eye*" (320 p.), Hillsdale: LEA.

Massaro, D. W., & Cohen, M. M. (**1983**). "Evaluation and Integration of Visual and Auditorial Information in Speech Perception," Journal of Experimental Psychology: Human Perception and Performance **9**, 753-771.

Massaro, D.W., Tsuzaki, M., Cohen, M.M., Gesi, A., Heredia, R. (**1993**). "Bimodal speech perception: an examination across languages," Journal of Phonetics **21**, 445–478.

McGurk, H., & MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **265**, 746–748.

Nahorna, O., Berthommier, F., & Schwartz, J.L. (**2012**). "Binding and unbinding the auditory and visual streams in the McGurk effect," J. Acoust. Soc. Am. **132**, 1061-1077.

1192 Noppeney, U., Ostwald, D., & Werner, S. (**2010**). "Perceptual decisions formed by
1193 accumulation of audiovisual evidence in prefrontal cortex," Journal of Neuroscience
1194 **30**, 7434-46.

1195 Ratcliff, R., & Rouder, J.N. (**1998**). "Modeling response times for two-choice decisions,"
1196 Psychological Science **9**, 347–356.

1197 Sanabria, D., Soto-Faraco, S., Chan, J.S., & Spence, C. (**2005**). "Intramodal perceptual
1198 grouping modulates multisensory integration: Evidence from the crossmodal
1199 congruency task," Neuroscience Letters**377**, 59-64.

1200 Schwartz, J. L. (**2006**). "Bayesian model selection: The 0/0 problem in the fuzzy-logical
1201 model of perception," Journal of the Acoustical Society of America **120**, 1795–1798.

1202 Schwartz, J. L. (**2010**). "A reanalysis of McGurk data suggests that audiovisual fusion in
1203 speech perception is subject-dependent," Journal of the Acoustical Society of
1204 America **127**, 1584-1594.

1205 Schwartz, J.L., Tiippana, K., & Andersen, T. (**2010**). "Disentangling unisensory from
1206 fusion effects in the attentional modulation of McGurk effects: a Bayesian modeling
1207 study suggests that fusion is attention-dependent," in Proceedings AVSP2010 (pp.
1208 23-27). Tokyo, Japan.

1209 Schwartz, J.L., Berthommier, F., & Savariaux, C. (**2004**). "Seeing to hear better:
1210 Evidence for early audio-visual interactions in speech identification," Cognition **93**,
1211 B69–B78.

1212 Schwartz, J.L., Robert-Ribes, J., & Escudier, P. (**1998**). "Ten years after Summerfield ...
1213 a taxonomy of models for audiovisual fusion in speech perception," in R. Campbell,

1214 B. Dodd & D. Burnham (eds.) Hearing by Eye, II. Perspectives and directions in
1215 research on audiovisual aspects of language processing (pp. 85-108). Hove (UK):
1216 Psychology Press.

1217 Sekiyama, K. & Burnham, D. (**2008**). Impact of language on development of auditory-
1218 visual speech perception. Developmental Science **11**, 306-320.

1219 Sekiyama, K. & Tohkura, Y. (**1993**). "Inter-language differences in the influence of visual
1220 cues in speech perception," Journal of Phonetics **21**, 427-444.

1221 Sekiyama, K. & Tohkura, Y. (**1991**). "McGurk effect in non-English listeners: Few visual
1222 effects for Japanese subjects hearing Japanese syllables of high auditory
1223 intelligibility," The Journal of the Acoustical Society of America **90**, 1797-1805.

1224 Smith, P.L., & Ratcliff, R. (**2004**). "Psychology and neurobiology of simple decisions,"
1225 Trends in Neurosciences **27**, 161-168.

1226 Sodoyer, D., Girin, L., Jutten, C., &Schwartz, J.L. (**2004**). "Further experiments on
1227 audio-visual speech source separation," Speech Communication**44**, 113-125.

1228 Soto-Faraco, S., & Alsius, A. (**2007**). "Conscious access to the unisensory components of
1229 a cross-modal illusion," Neuroreport **18**, 347–50.

1230 Soto-Faraco, S., & Alsius, A. (**2009**). "Deconstructing the McGurk-MacDonald
1231 illusion," Journal of Experimental Psychology: Human perception and performance
1232 **35**, 580-7.

1233 Soto-Faraco, S., Navarra, J., & Alsius, A. (**2004**). "Assessing automaticity in audiovisual
1234 speech integration: evidence from the speeded classification task," Cognition **92**,
1235 B13–B23.

1236 Sumby, W., & Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise,"
1237 Journal of the Acoustical Society of America**26**, 212–215.

1238 Summerfield, Q. (**1987**). "Some preliminaries to a comprehensive account of audio-visual
1239 speech perception," inB. Dodd & R. Campbell (eds.) *Hearing by Eye: The Psychology of*
1240 *Lipreading* (pp. 3–51) New York (NY): LawrenceErlbaum Associates.

1241 Summerfield, Q., & McGrath, M. (**1984**). "Detection and resolution of audio-visual
1242 incompatibility in the perception of vowel," Quarterly Journal of Experimental
1243 Psychology **36A**, 51-74.

1244 Tiippana, K., Andersen, T.S., & Sams, M. (**2004**). "Visual attention modulates
1245 audiovisual speech perception," European Journal of Cognitive Psychology **16**, 457–
1246 472.

1247 Tiippana, K. Puharinen, H., Möttönen, R., & Sams, M. (**2011**). "Sound Location Can
1248 Influence Audiovisual Speech Perception When Spatial Attention Is Manipulated,"
1249 Seeing and Perceiving **24**, 67–90.

1250 Vroomen, J. & Baart, M. (**2011**). "Phonetic recalibration in audiovisual speech," in M.
1251 M. Murray & M. T. Wallace (eds.) *Frontiers in the neural basis of multisensory processes*
1252 (pp. 363-379) Routledge: Taylor & Francis.

1253 van Wassenhove, V., Grant, K.W., & Poeppel, D. (**2005**). "Visual speech speeds up the
1254 neural processing of auditory speech," Proceedings of the National Academy of
1255 Sciences **102**, 1181–1186.

1256 Van Wassenhove, V., Grant, K.W., & Poeppel, D. (**2007**). "Temporal window of
1257 integration in bimodal speech," Neuropsychologia **45**, 598-607.

1258 Yu, A. J., Dayan, P., & Cohen, J. D. (**2009**). Dynamics of attentional selection under
1259 conflict: Toward a rational Bayesian account. Journal of Experimental Psychology:
1260 Human Perception and Performance **35**, 700-717.

1261

1262

Figure captions

Figure 1 – Organization of stimuli in Experiment 1.

Figure 2 – Percentage of "ba" responses (relative to the total number of "ba" + "da" responses) for the two targets in the three contexts and without context.

Figure 3 – Percentage of "ba" responses for McGurk targets for the three contexts and their five durations, compared to targets without context.

Figure 4– Effect of the preceding decision in Experiment 1. Responses to McGurk stimuli depending on context ("Coh" for coherent, "Incoh" for incoherent), preceding context ("Prec coh" for coherent preceding context, "Prec incoh" for incoherent preceding context), preceding target stimulus ("Prec ba" vs "Prec McGurk") and previous answer ("Ans ba" for previous "ba" target, "Ans ba" and "Ans da" for previous "McGurk" target). We do not present results for phonetically incoherent context to make the figure clearer.

Figure 5– Mean response times for the two targets in the three contexts and without context.

Figure 6 – Mean response times for the two targets in the five context durations and without context.

Figure 7 – Organization of stimuli in Experiment 2.

Figure 8 – Percentage of "ba" responses (relative to the total number of "ba" + "da" responses) for the McGurk targets with coherent context and with incoherent context for the two reset types and the four reset durations.

Figure 9 – Mean response times for the two targets in the two reset conditions.

Figure 10 – A possible cognitive architecture for audiovisual binding and fusion in speech perception.