Audiovisual Binding for Speech Perception in Noise and in Aging

Attigodu Chandrashekara Ganesh, Frédéric Berthommier, Jean-Luc Schwartz

Audiovisual Binding for Speech Perception in Noise and in Aging

Attigodu Chandrashekara Ganesh, Frédéric Berthommier and Jean-Luc Schwartz

GIPSA-lab (Grenoble Images Parole Signal Automatique), Speech and Cognition Department

Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France

* Institute of Engineering Univ. Grenoble Alpes

Address correspondence to Jean-Luc Schwartz, CNRS, GIPSA-lab, UMR 5216, Grenoble University, Grenoble, France, jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

Abstract

Speech Perception involves fusion of multiple sensory input and it doesn't fuse automatically, perhaps it depends on numerous external/internal factors (e.g. attention, noise or age). In this paper, we exploit a specific paradigm in which a short audiovisual context made of coherent or incoherent speech material is displayed before an incongruent audiovisual target likely to provide fusion (McGurk effect, McGurk & MacDonald, 1976). We confirm that incoherent context leads to unbinding, that is a reduction in the amount of fusion. Importantly, adding acoustic noise in the context though not in the target increases fusion. This suggests that listeners systematically evaluate the reliability of their sensory channels and weight them accordingly in the fusion process. We also show that older subjects display more unbinding, and discuss the potential consequences concerning their ability to understand speech in adverse conditions. We relate all these data to a "Binding-and-Fusion" model of audiovisual speech perception.

Keywords: Audiovisual Integration, McGurk effect, Speech perception, Audiovisual Binding

Introduction

Audiovisual Binding in Speech Perception

Speech perception is a multisensory process. The human brain is able to exploit the visual input provided by the vision of the speaker's face to enhance perception in noise (Benoit, Mohamadi, & Kandel, 1994; Erber, 1969; Sumby & Pollack, 1954) or when audition is impaired (Auer & Bernstein, 2007; Bernstein, Demorest, & Tucker, 2000; Grant, Walden, & Seitz, 1998; Tye-Murray, Sommers, & Spehar, 2007; Walden, Busacco, & Montgomery, 1993).

The discovery of the McGurk effect (McGurk & MacDonald, 1976), replicated in several studies, suggests that the human brain is able to combine the auditory and the visual input even though they are discordant. However, the audiovisual scene comprises a large amount of auditory and visual information that must be analyzed and selected before adequate fusion may occur: this is *Audiovisual Binding*.

The McGurk effect was first considered as pre-attentive and "automatic". McGurk and MacDonald (1976) note that the effects of the incongruent visual input "do not habituate over time, despite objective knowledge of the illusion involved", Summerfield & McGrath (1984) insist on the fact that the illusion is "compelling", and note that the effect remains even if subjects are instructed to pay attention only to the sound. Since automaticity is seldom defined precisely in these papers, we propose here a computational definition. Let us consider a subject provided with an audio stimulus A combined with a video stimulus V, being possibly identified within a given set of N phonemic categories C_i ($i \in \{1..N\}$). Audiovisual fusion is considered automatic if the perceptual response to the audiovisual pair (A, V) depends only on the auditory and visual inputs:

$$p_{AV}(C_i) = f(A,V) \quad (1)$$

where f is some function describing the processing and categorization mechanisms enabling to compute perceptual outputs from sensory inputs.

In the years, several experimental data have been gathered showing that fusion is actually *not* automatic in the McGurk paradigm. Indeed, the McGurk effect can be reduced if there is competing information in the visual modality providing distracting cues (see e.g. Tiippana, Andersen, & Sams, 2004). It is also decreased by loading the audiovisual speech perception task at hand with a second task performed at the same time (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Buchan & Munhall, 2012).

Nahorna, Berthommier, and Schwartz (2012) showed that if a "McGurk" target made of an audio "ba" and a video "ga" was preceded by an audiovisual incoherent context made of incompatible auditory and visual speech material (e.g., audio syllables dubbed on video sentences), then the amount of perception of the McGurk fusion "da" was largely decreased. The authors interpreted the whole set of results on the modulation of the McGurk effect in the framework of a "two-stage model of audiovisual fusion" (Figure 1) (Berthommier, 2004). According to this model, a first "Binding" stage evaluates the coherence of the auditory and visual inputs all along time, to assess whether they are associated with the same source. Then a second "Fusion" stage integrates the auditory and visual inputs to produce a fused percept, but the integration result depends not only on the auditory and visual inputs but also on the output of the Binding stage. The classical McGurk effect without context would occur because the subject would be in a "default state" characterized by binding. However, if the audiovisual context is incoherent, this provides evidence that the auditory and visual inputs do not correspond to the same source, which results in "unbinding" the sound and image in the decision process. Then, if a "McGurk" target is presented after such incoherent context, the subject experiences less fusion and provides more auditory responses. Unbinding is rapid (1 or 2 incoherent syllables suffice to produce a maximum fusion decrease), and a given amount of coherent material enables to "rebind" and recover the original McGurk effect (at

least 3 coherent syllables were required to recover binding after unbinding: see Nahorna, Berthommier, & Schwartz, 2015).

In the "two-stage model", the "Binding" box has two roles (Figure 1A). Firstly, it selects the adequate pieces of information to be fused for speech decoding (e.g. the adequate speaker's face and sound in a cocktail party scene). Secondly, it specifies the weight of each sensory channel, w_A and w_V –their "reliability" in the fusion process– decreasing the weight of the visual channel in the case of distracting visual input (Tiippana et al. (2004), "cognitive load" conditions (Alsius et al., 2005, 2007) and "unbinding" (Nahorna et al., 2015). The fusion process hence now includes the auditory and visual weights w_A and w_V:

$$p_{AV}(C_i) = f(A, V, w_A, w_V) \quad (2)$$

This renders fusion "non-automatic" in the meaning defined previously: fusion now depends not only on the auditory and visual inputs but also on contextual variables w_A and w_V likely to vary with attention, cognitive load or previous audiovisual material.

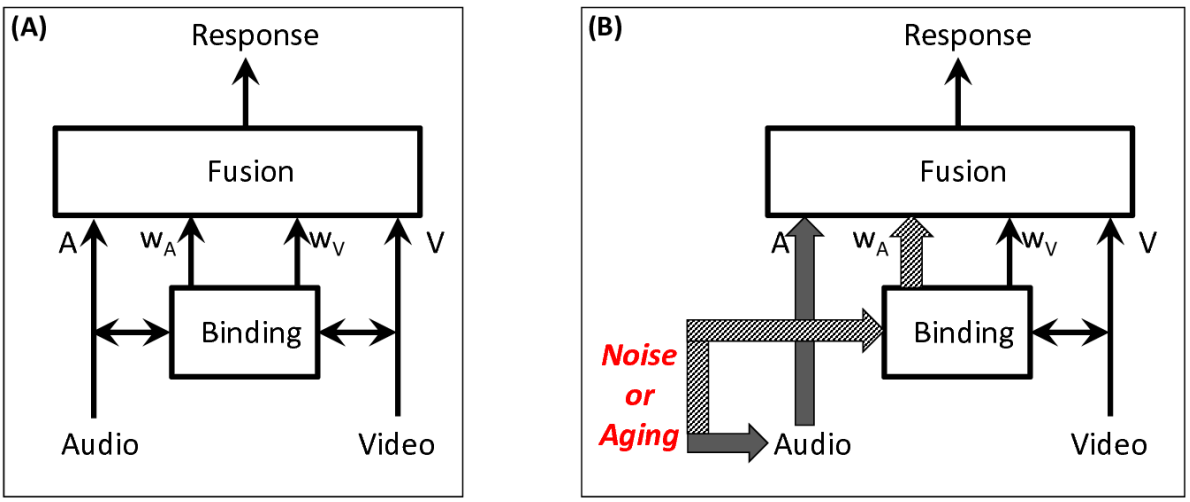


Figure 1-A) The two-stage model of audiovisual speech perception- The binding box selects the appropriate information in the audiovisual scene. The selected auditory (A) and visual (V) information is weighted by channel reliability (w_A and w_V). **B) Two contrasting hypotheses about the role of noise and aging in audiovisual fusion.** When the audition is degraded, by noise or aging, the auditory input is less reliable in the fusion process (grey arrows), but the weight of the audio channel is also decreased (hatched arrows). In the first

hypothesis (Hyp. 1), noise or aging modify $p_{AV}(C_i)$ through A in Eq. 1. In the second hypothesis, (Hyp. 2) they also intervene through the auditory weight w_A in Eq. 2.

Audiovisual Binding in Audition Degraded by Noise or Aging

The increased role of the visual channel when the audition is impaired or degraded by acoustic noise is related to the concept of inverse effectiveness according to which the lower the quality of individual sensory stimuli the higher the magnitude of multisensory enhancements (Meredith & Stein, 1983, 1986). In the McGurk effect, it is commonly assumed that the respective ambiguities of the auditory or visual inputs drive the output of the fusion process (Massaro, 1989, 1998). Hence, degrading a given input would lower its role in fusion (gray arrows in Figure 1B): this will be Hypothesis 1 (Hyp. 1) in the following. It is typically what happens in the case of acoustic noise increasing the amount of fusion (Sekiyama, 1994; Sekiyama & Tohkura, 1991) or visual noise decreasing it (Fixmer & Hawkins, 1998; Kim & Davis, 2011).

However, the effect of noise could also be assumed to intervene at the binding stage preliminary to fusion. In this interpretation, the noise would operate not only at the stimulus level but also at the channel level: if a sensory channel were degraded, its weight would decrease in the fusion process, whatever the value of instantaneous inputs. Hence, binding would incorporate an evaluation of the level of noise within each sensory channel, controlling the weight of each modality in the fusion process (w_A and w_V in Figure 1B) (see e.g. Huyse, Berthommier, & Leybaert, 2013; Schwartz, 2010): this will be Hypothesis 2 (Hyp. 2) in the following.

Interestingly, the "context+target" paradigm elaborated by Nahorna et al. (2015) provides a way to disentangle between the two hypotheses about the role of noise in audiovisual fusion that are input vs. channel reliability. Indeed, suppose we introduce some amount of acoustic noise in the context, be it coherent or incoherent, but *not* in the "McGurk"

target following context. If the decision about the target only depends on the ambiguity of each component of the target (Hyp. 1), then, since there is no noise during the target, the McGurk effect should *not* vary. However, if there is indeed an evaluation of the quality of the audio and video channels all along time (Hyp. 2), then adding acoustic noise in the context should decrease the reliability of the audio channel and hence decrease its role in fusion, with a corresponding increase in the amount of McGurk effect. *This principle will provide the basis of the first experiment in the present study.*

Another situation where audition is degraded is aging. As adults age, their sensory, perceptual and cognitive abilities tend to decline (Baltes & Lindenberger, 1997; Pichora-Fuller & Singh, 2006). Presbycusis (age-related hearing loss) is one of the common disorders seen in older adults, which can affect the ability to understand speech, especially in adverse conditions (CHABA, 1988). As a matter of fact, many older adults indicate that listening in noisy situations is a challenging and often exhausting experience. In addition, lip-reading skills also seem to decrease with age in spite of normal or corrected vision (Cienkowski & Carney, 2002; Dancer, Krain, Thompson, Davis, & et al., 1994; Feld & Sommers, 2009; Shoop & Binnie, 1979; Sommers, Tye-Murray, & Spehar, 2005).

This general unisensory deficit seems to be accompanied by *greater* multisensory integration (see Mozolic, Hugenschmidt, Peiffer, & Laurienti, 2012 for review). Indeed, a number of studies suggest an aging-related increase in the McGurk effect (Behne et al., 2007; Setti, Burke, Kenny, & Newell, 2013; Thompson, 1995). This is somewhat debated (Cienkowski & Carney, 2002; Hay-McCutcheon, Pisoni, & Kirk, 2005; Sommers et al., 2005; Tye-Murray et al., 2007; Walden et al., 1993), but Sekiyama, Soshi, and Sakamoto (2014) confirmed that visual influence was larger in older compared with younger Japanese adults, even in calibrated SNRs accounting for differences in auditory thresholds. Altogether, seniors

appear to exploit the visual speech input and fuse it with the auditory speech input at least as much as youngers, and possibly more.

Other modifications in the integration process have been found in the literature. Aging could involve a larger temporal window of multisensory integration, and less efficient selective attention processes so that seniors would be more distracted than youngers by spurious events coming from an unattended modality, irrelevant for the task at hand (e.g. Alain & Woods, 1999; Poliakoff, Ashworth, Lowe, & Spence, 2006). In conclusion of their enlightening review of the literature, (Mozolic et al., 2012) introduce the proposal that the whole range of differences between youngers and seniors in multisensory integration could be associated to a possible single explanation that they call "increased noise at baseline", that is the level of sensory noise associated with each modality. This level would be higher in seniors in all modalities, resulting in increased activity related to a given modality and hence larger multisensory interactions.

Importantly, this discussion about aging is once again related to the distinction between fusion and binding. The internal noise hypothesis means that the main difference between youngers and seniors would deal with sensory representations and the way they modify fusion: aging would essentially result in increasing internal noise associated with auditory and visual representations (Hyp. 1, gray arrows in Figure 1B). However, another possibility could be that in addition, aging would produce modifications in the respective weights of the sensory inputs, w_A and w_V (Hyp. 2, hatched arrows in Figure 1B), resulting in modifications in the output of the fusion process.

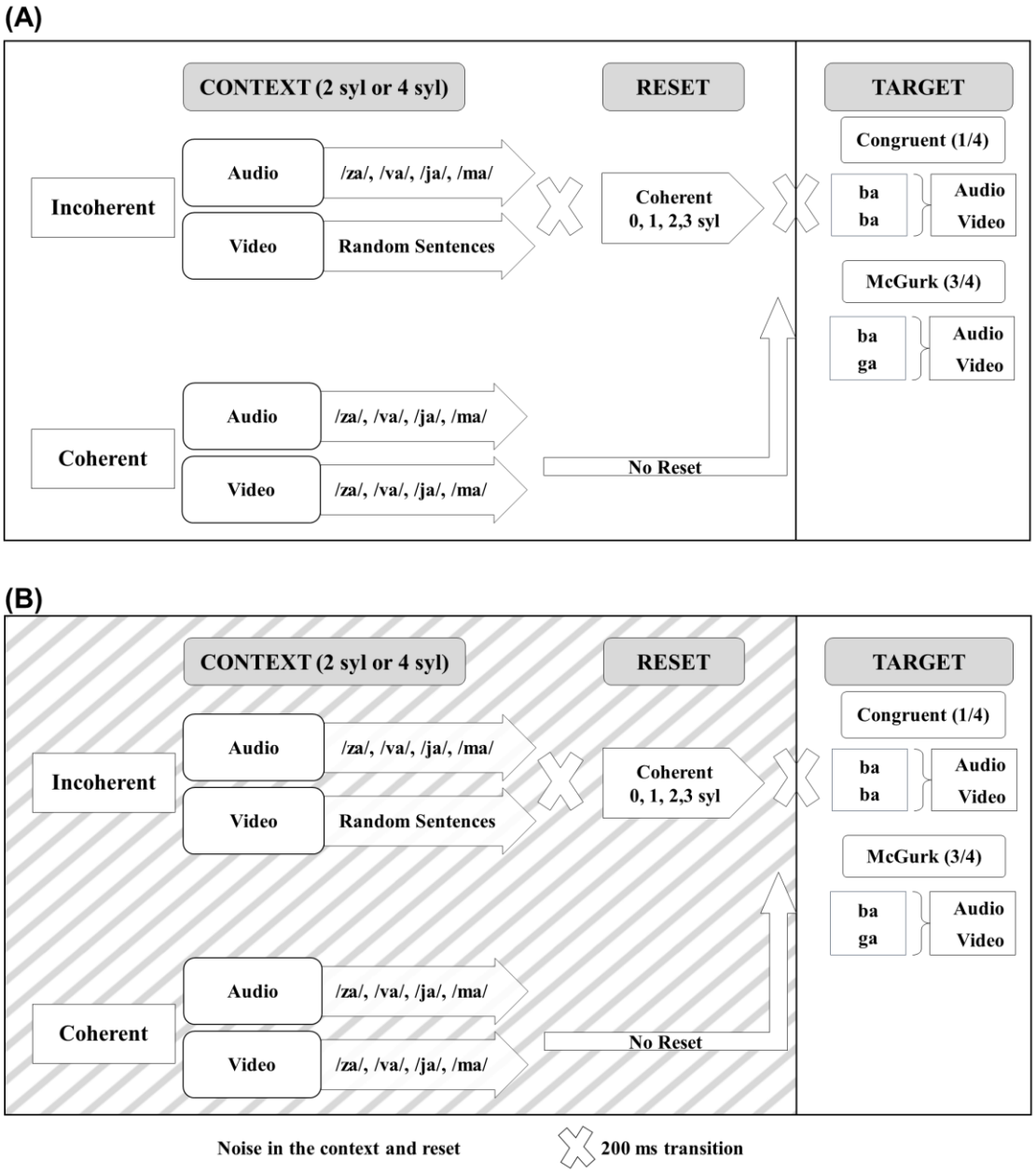
Once again, the (context+target) paradigm introduced by Nahorna et al. (2015) could shed some light on these two contrasting hypotheses. Indeed, if two sets of younger and senior participants are tested and controlled such as to provide similar levels of McGurk effect, binding processes could differ from one population to the other. If aging in audiovisual

fusion mainly results in increased sensory noise and does not intervene in binding per se, then the effect of context should be similar in the two populations. However, if context effects differ from one group to the other in spite of similar levels of fusion in the target without context, this would indicate differences in binding associated with aging. *This principle will provide the basis of the second experiment in the present study.*

Experiment 1 – Effect of noise on audiovisual binding in speech perception

Material and Methods

Twenty-nine participants (21 women and 8 men; 29 right-handed; from 18 to 50 years, mean age=30.0 years; SD=10.0 years) took part in this study. All of them were native French speakers with self-reported normal or corrected-to-normal vision and without hearing disorders. Written informed consent was obtained from all participants and all procedures were approved by the Grenoble Ethics Board (CERNI).



208

209 **Figure 2- Stimuli – A)** In Experiments 1 and 2, the stimuli are comprised of a context, coherent or not, a
210 reset and a target. **B**) In Experiment 1, the context may be degraded by acoustic noise.

211 The stimuli, procedure and response analyses closely replicate those implemented by
212 Nahorna et al. (2015). Stimuli began with a "context" and ended with a "target". The context
213 could be either incoherent (Figure 2A, top) or coherent (Figure 2A, bottom). In the case of
214 incoherent context, a "reset" was introduced between the context and the target (see

Supplementary Materials for detailed information about the precise construction of the stimuli).

The coherent context consisted of 2 or 4 coherent audiovisual syllables and acted as a control providing a reference for the McGurk effect in the incoherent context. The incoherent context was prepared by dubbing a sequence of 2 or 4 acoustic syllables on a video stream containing excerpts of sentences with the adequate duration. The reset stimulus, which was always presented after the incoherent context, consisted of 0, 1, 2 or 3 coherent audiovisual syllables. The "0" syllable reset was nothing but pure incoherent context where there was no reset material presented. In the statistical analyses that will be presented later, the stimuli were grouped into context/reset type (5 variants: coherent vs. incoherent with 0, 1, 2 or 3-syllables reset) and context duration (2 vs. 4 syllables).

The target was either a congruent audiovisual "Ba" syllable or an incongruent "McGurk" stimulus with an audio "ba" mounted on a video "ga". It was expected that the "McGurk" stimuli should be perceived as either "ba" or "da" (McGurk & MacDonald, 1976) while congruent "ba" stimuli should be unambiguously perceived as "ba". The "McGurk" targets were the main interest in the present study while the congruent "Ba" targets only served as controls. Therefore, "McGurk" targets were presented three times more than congruent "Ba" targets. Exactly the same audiovisual targets were associated with either coherent or incoherent context.

The target stimuli were never corrupted by acoustic noise. In one condition, however, acoustic Gaussian white noise at 0 dB SNR was added to the context and reset periods of the stimuli (see Figure 2B). The whole experiment consisted of two blocks, one without acoustic noise and the other one with acoustic noise. The order of the two blocks ("without noise" and "with noise") was counterbalanced between participants. The participant's task was to detect online "ba" or "da" syllables (syllable monitoring task), without knowing when they could

occur in the sequence. Detection was achieved by pressing as rapidly as possible on a keyboard. Therefore, subjects could provide responses at any time along the monitoring process (see more information on the experimental procedure in Supplementary Materials).

The analysis was based on the evaluation of the response time relative to the acoustic onset of target syllables. Responses were taken into account only if they occurred within a [200-1200 ms] time window. Responses outside this window were ignored and double different responses within the time window were also discarded (these two cases are summarized as "misses" in the following). For each condition of target and context and for each participant, the percentage of "ba" responses – not including misses – was taken as the response score and the response time (RT) was estimated by averaging the response times for all stimuli in the corresponding condition.

Analyses of variance (ANOVAs) were performed on both response scores and response times applying a Greenhouse – Geisser correction in case of violation of the sphericity assumption. *Post-hoc* analyses with Bonferroni correction were done when appropriate and reported at the [($p < 0.05$) level]. To ensure quasi-Gaussian distribution of the variables, the response scores were processed with an arcsine square root transform, and the response times were logarithmically transformed (see more information on the whole analysis protocol in Supplementary Materials). All effects reported in the following are significant (detailed statistical analyses, including significant and non-significant effects, effect sizes and all *post-hoc* tests are provided in the Supplementary Materials).

Results

The target was missed 6.4% of the cases, for the whole experiment and in average over the 29 subjects. There were significantly less misses in the "without-noise" condition (3.8%) than in the "with-noise" condition (9.1%) (see the detailed pattern of errors in Supplementary Materials). The relatively large number of errors is likely due to the

complexity of the task, reminding that the time of apparition of the target was unpredictable and that the temporal structure of the stimuli with the context + reset + target structure was rather complex. Importantly, these values were rather stable for both congruent and McGurk targets and from one context condition to the other in both noise conditions.

The amount of errors was however largely variable between subjects, reaching more than 25% in some subjects and some conditions. In the following, we discarded from the analyses all subjects with more than 25% errors in either the without-noise condition or the with-noise condition or both. This resulted in keeping only 23 subjects for further analyses.

Proportion of "ba" responses

The "Ba" targets were classified as "ba" more than 98% of the cases (not including misses) in all contexts. They will not be considered anymore in this analysis. On Figure 3A we display the response scores for "McGurk" targets in all conditions of context and noise, averaged over the 23 subjects. A preliminary test of possible "block effects" (noise first vs. noise second) showed no effect of block or any interaction effect except with context/reset type, but further *post-hoc* analysis showed no significant difference between any context/reset value from one block to another. Hence, blocks were averaged in all the following.

Three factors, context/reset type (coherent vs. incoherent with 4 reset durations, hence 5 possibilities altogether), context duration (two vs. four syllables) and noise (with noise vs. without noise) were analyzed using repeated-measures ANOVA.

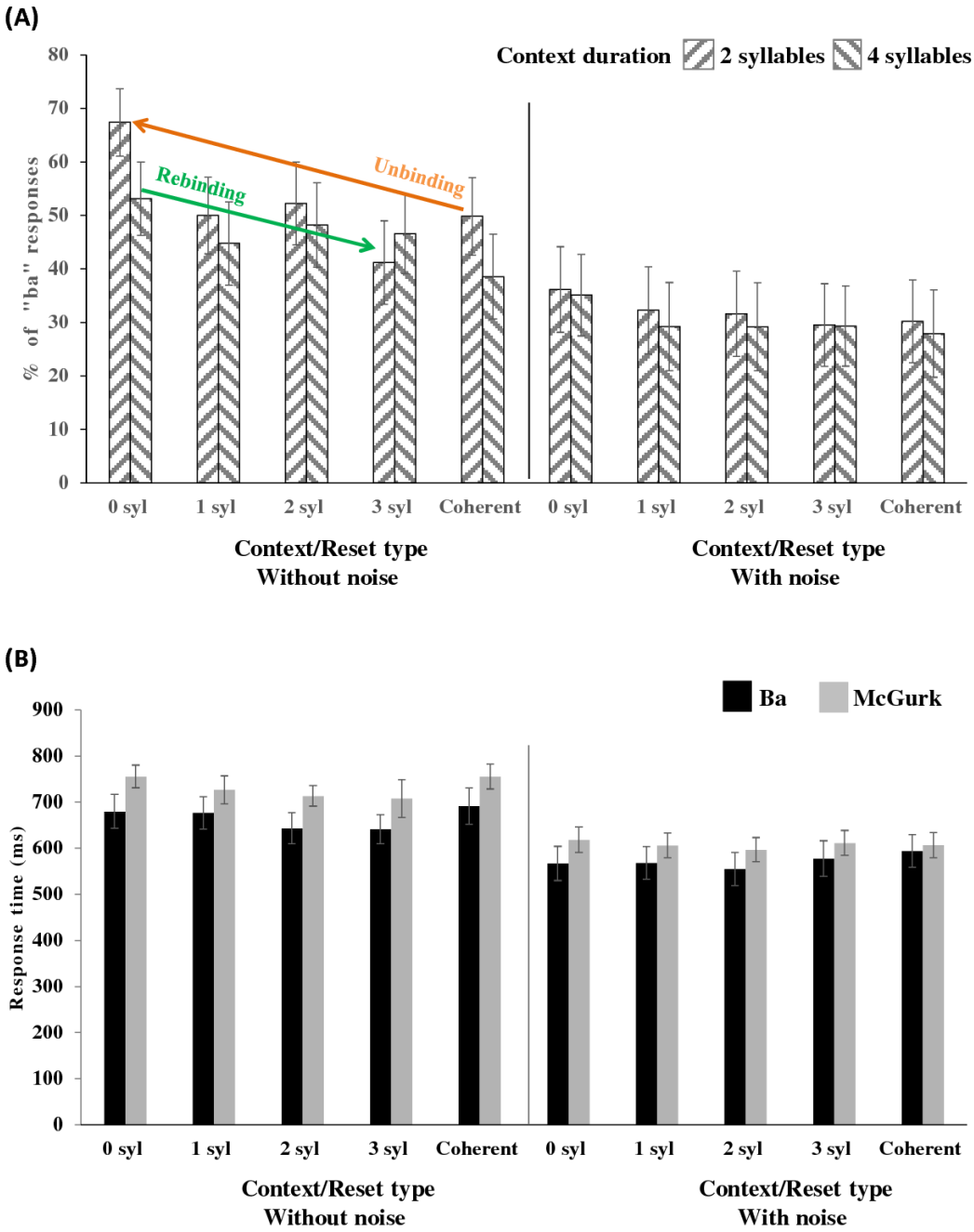


Figure 3 A) Proportion of "ba" responses for "McGurk" targets, without noise (left) or with noise (right) for incoherent context with four reset durations (0syl, 1syl, 2syl or 3syl), compared with coherent context, and for both context durations (2 or 4 syllables). Unbinding and rebinding associated to significant variations of "ba" responses without noise are displayed by colored arrows (color online) – though note that arrows do not indicate that the modification is linear. **B)** Response times for "McGurk" and "Ba" targets without noise (left) or with

noise (right) for all context/reset types, averaged over context duration. In both A and B, Error bars display standard errors computed from the residual error in the corresponding ANOVA (subject variability removed).

The effect of context/reset type [F (4, 88) = 12.29, p<0.001], context duration [F (1, 22) = 13.35, p<0.005] and noise [F (1, 22) = 21.81, p<0.001] were significant. All interactions between 2 or 3 factors were also significant (see detailed outputs of the ANOVA in Supplementary Materials). *Post-hoc* analyses show that the effects of context/reset type and context duration are significant only without noise. This lets emerge the main following outcomes.

Unbinding and rebinding associated to context/reset type and duration in the "without-noise" condition.

Here we discuss only data in the without-noise condition. Globally, the proportion of "ba" responses increases (hence the McGurk effect decreases) from the coherent to the incoherent-without-reset (0 syl reset) condition: this is unbinding. Conversely, the proportion of "ba" responses decreases in the incoherent context when reset duration increases from 0 syllable to 3 syllables: this is rebinding. *Post-hoc* analyses confirm that the incoherent context without reset provides a significantly higher percentage of "ba" responses than both the coherent context and the incoherent context with non-zero reset (1, 2 or 3 reset syllables). These effects are rather large, as displayed by variations of "ba" percentage between contexts in Figure 3A.

Furthermore, as in Nahorna et al. (2015), the proportion of "ba" responses is larger for the smaller context duration (2 syllables). Analysis of interactions and post-hocs shows that this effect is significant only in the two shortest context conditions that are coherent and incoherent without reset.

Effect of acoustic noise in the context

Globally, noise decreases "ba" scores (increases the McGurk effect) for all conditions. The effect is large. Indeed, in the coherent context or totally rebound 3syl reset conditions the percentage of "ba" responses decreases from 44% without noise to 29% with noise (averaged over the two context durations), while in the unbound 0 syl reset condition it decreases from 60% without noise to 35% with noise (averaged over the two context durations). The consequence is that all statistically significant interaction effects with noise are basically ceiling trends, in which the effects of context/reset type and context duration are strongly decreased and become non-significant in the "with-noise" condition.

Analysis of response times

On Figure 3B we display the response times for "Ba" and "McGurk" targets, averaged over the 23 subjects and over context duration. A four-way repeated-measures ANOVA on response times displays an effect of target [F (1, 22) = 23.94, p<0.001], noise [F (1, 22) = 51.55, p<0.001], context/reset type [F (4, 88) =6.40, p<0.005], and context duration [F (1, 22) =11.33, p<0.005], but no interaction between any variables (see detailed results in Supplementary Materials). The responses were quicker for all "Ba" targets compared to "McGurk" targets (51 ms average difference). The lack of interaction between the target and other variables shows that the effect of audiovisual incongruence in the "McGurk" target produces the same amount of delay compared with a congruent "Ba" target, whatever the noise, context/reset type and context duration.

As in Nahorna et al. (2015), shorter contexts (that is, 2-syllable context duration or context without reset) produce larger response times: 2syl vs. 4syl context duration increased RT by 25 ms, context without reset vs. context with 3-syllable reset increased RT by 28 ms.

Surprisingly, the response was quicker for both targets with noise compared to without noise, with a large difference of 109 ms in average. This might seem surprising, but the interpretation is straightforward. Indeed, since noise stops soon after context, it provides a

clear temporal cue for participants regarding the arrival of the target stimuli, which results in quicker responses in the "noise" condition.

Discussion

The results of this experiment, in the case of context without noise, replicate the major findings in Nahorna et al. (2015): (1) unbinding by incoherent context, that is decrease in the amount of McGurk responses, already maximal for a 2-syllable context duration, (2) rebinding by a coherent context, total for a 3-syllable reset duration, that is complete recovery of the McGurk effect, and (3) increase in response time from "Ba" to "McGurk" targets, with no interference with context duration or context/reset type, that is no significant interaction of these variables with the "target" variable. There were also larger response times for shorter contexts (2-syllable contexts without reset), together with an increase in "ba" responses from 2- to 4-syllable context without reset. More detailed analysis of this pattern of responses is provided in Nahorna et al. (2015) and will be developed in the General Discussion.

But the major result of Experiment 1 concerns the role of acoustic noise in the fusion process, with an original paradigm in which acoustic noise was present in the context but not in the target. The result is clear: noise does matter and modulates the output of the fusion process, by decreasing the number of auditory "ba" responses in all conditions of context coherence, context duration and reset duration. This is rather in line with the "noise-channel" hypothesis introduced in the Introduction section (Hyp. 2, see hatched arrows in Figure 1B). It suggests that noise in the context plays a direct role in the fusion process, and contributes to modify the weight of the unisensory components.

Experiment 2 – Effect of aging on audiovisual binding in speech perception

Methods and Materials

Twenty-five native French speaking older adults participated in the experiment (2 women and 23 men, 21 right-handed and 4 left-handed, from 60 to 75 years, mean age= 65.3

years, SD=3.9 years). None of them reported any hearing, vision (after correction) or neurological disorders. Written informed consent was obtained from each participant and all procedures were approved by the Grenoble Ethics Board (CERNI).

Audiometric thresholds were obtained at octave intervals from 250 to 8000 Hz. In all participants, pure-tone averages (calculated as the average threshold from 500 to 2000 Hz) were ≤ 25 dB HL and 35 to 40 dB HL in higher frequencies.

In addition to the screening audiometry, we administered a French version of the Speech, Spatial, and Qualities of hearing scale which is a self-reported questionnaire developed to assess how effectively auditory information is being processed in various everyday listening situations. We also administered the French version of the color-word Stroop task (Stroop, 1935), considered to measure various executive functions such as selective attention and cognitive flexibility, interference control, response inhibition and brain's processing speed. A detailed description of the use of these tests is provided in the Supplementary Materials. It happens that no significant correlation was obtained between these additional measurements and the participants' performance in the audiovisual binding task, as also described in the Supplementary Materials.

The stimuli were the same as those in Experiment 1 in the "without noise" condition (Figure 2A). The procedure, response processing, and statistical analyses were exactly the same as in Experiment 1.

The principle of this experiment consisted of comparing the effect of binding on younger vs. older adults, starting from a baseline state (McGurk effect with coherent context) similar in both groups. We observed that discarding in both groups participants with more than 90% "ba" scores in the "coherent condition" for "McGurk" targets, considered as participants with a poor level of audiovisual fusion, and hence unlikely to display large unbinding/rebinding effects, led to similar amounts of McGurk effect in the two groups.

Indeed, while this resulted in discarding 9 participants over 29 in the younger group (not taking into account the number of misses per subject at this stage) and 8 participants over 25 in the older group, the resulting mean amount of "ba" responses in the "McGurk" targets with coherent context (averaged over the two context durations) respectively reached 28% for the younger adults group with 20 remaining subjects, and 32% for the older adults group with 17 remaining participants. This difference is non-significant, as will be shown in the next section.

It is important to stress at this stage that differences between younger and older subjects in the amount of audiovisual fusion may emerge from various causes: inter-individual variability (Schwartz, 2010) and, in the case of elder subjects, hearing loss (evidenced by previously reported audiometric thresholds), decrease in lipreading abilities, or increase in audiovisual fusion, according to e.g. Sekiyama et al. (2014). Therefore the equalization in McGurk scores is used in this study to provide a similar global baseline in both populations, with no claim about the underlying processes. Binding/unbinding/rebinding processes are then considered to operate from this baseline. Hence any difference between younger and older participants associated with audiovisual context is taken as a direct measure of the difference in binding processes between tested populations.

Results

The target was missed 11.5% of the cases, averaged over all contexts and over the 17 senior subjects kept in this experiment. This amount is significantly larger than for younger participants both in silence and in noise (see Supplementary Materials). This high value shows that the task is relatively difficult for senior subjects. Here again, these values were rather stable from one context condition to the other (see more information on the pattern of errors in Supplementary Materials). The amount of errors was however largely variable between subjects, from 0 to 41%, and here again we discarded from the analyses subjects

with more than 25% errors in average. This eliminated 2 more subjects, resulting in keeping only 15 subjects for further analyses.

Analysis of the proportion of "ba" responses

The "Ba" targets were classified as "ba" more than 98% of the cases (not including misses) in all contexts. They will not be considered anymore in this analysis. Firstly, we assessed the response scores for "McGurk" targets for this group independently on the younger group, in a two-factor repeated-measures ANOVA. The effects of context duration [F (1, 15) =11.01, p<0.005], and context/reset type [F (4, 56) =39.40, p<0.001] were significant. The interaction between factors was not significant (see detailed outputs of the ANOVA in Supplementary Materials).

Post-hoc analyses show that there were significantly less "ba" responses for the coherent context than for the "0 syl" incoherent condition, with a 38% difference (unbinding). Complete rebinding required a "3syl" reset duration: indeed, the scores for the coherent context were significantly lower than the scores for the "1syl" and "2syl" reset duration but not significantly different from the "3syl" reset duration. Here again, the proportion of "ba" responses was higher (with less McGurk fusion) for the shorter context duration (comparing 2- vs. 4-syllable context duration for context without reset).

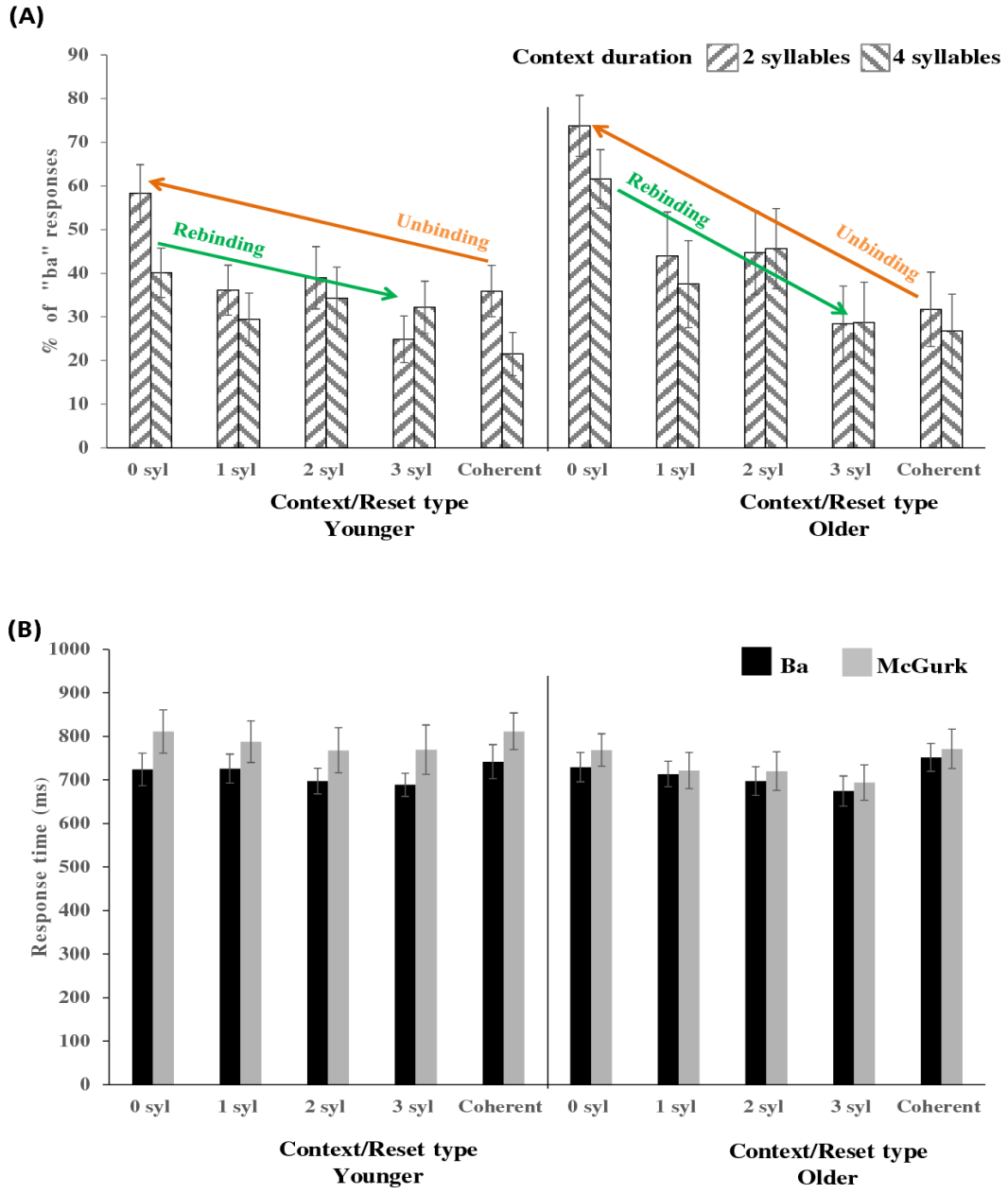


Figure 4 A) Proportion of "ba" responses for "McGurk" targets, for the 18 younger (left) vs. the 15 older participants (right) for incoherent context with four reset durations (0syl, 1syl, 2syl or 3syl), compared with coherent context, and for both context durations (2 or 4 syllables). Unbinding and rebinding are displayed by colored arrows (color online) – though note that arrows do not indicate that the modification is linear. **B)** Response times for "McGurk" and "Ba" targets for younger (left) vs. older participants (right) for all context/reset types, averaged over context duration. In both A and B, Error bars display standard errors computed from the residual error in the corresponding ANOVA (subject variability removed).

Then we compared data for the two groups, noting that the modulations of the McGurk effect with context appear larger for older participants (Figure 4 – notice that the values for youngers in Figure 4 do not correspond to those of the same group without noise in Figure 3, since only 18 young participants are considered here, instead of 23 in Experiment 1). A mixed ANOVA was conducted to compare "ba" scores between the younger and older groups according to the context/reset type and context duration. Though there was a significant main effect of context/reset type and a significant interaction effect between context/reset type and group, we could not report the results due to a violation of homogeneity of variance, since the Box's M test of equality of covariance matrices and Leven's test of equality of level variance were both significant.

Therefore, we focused on the amount of unbinding that is the modulation of binding from the coherent to the most incoherent condition. For this aim, we considered only the coherent context and the incoherent context without reset (0syl reset duration), averaging over both context durations to keep the focus on the important point, which was the comparison between coherent and incoherent contexts. We performed a two-way mixed ANOVA with age as the between-group variable (young vs. adult), and context as the within-subject variable (0syl incoherent vs. coherent condition), checking that in this case the assumption of homogeneity of variances was not violated. There was no significant difference between groups, but the context effect was significant [F (1, 31) =169.71, p<. 0001] and there was a significant interaction between context and groups [F (1, 31) =23.15, p<.001]. *The post-hoc* analysis shows that there was a significant difference between older and younger groups for the incoherent condition "0syl" reset duration (49% in youngers vs. 67% in elders) while the values in the coherent context were not significantly different (see blue arrows in Figure 4). Therefore, it appears that the dynamics of unbinding by incoherent context are larger for the older participants.

Analysis of response time

Response times were first analyzed separately for the older group. A three-factor repeated-measures ANOVA displayed an effect of target [F (1, 14) =5.38, p<0.05], context/reset type [F (4, 56) =3.80, p<0.005], and context duration [F (1, 14) =26.53, p<0.001], but no interaction between any variables (see detailed results in Supplementary Materials). As in previous experiments, (1) the responses were 22 ms quicker for "Ba" compared to "McGurk" targets, and (2) they were longer for shorter contexts: 65 ms longer for the "2syl" than for the "4syl" context duration and 78 ms longer for the context without reset than the context with 3-syllable reset. Once again, the lack of interaction between target and context/reset type shows that the effect of audiovisual incongruence in the "McGurk" target produces the same amount of delay compared with a congruent "Ba" target, whatever the context/reset type.

Finally, a mixed ANOVA was conducted comparing RTs between younger and older groups with targets and context/reset type (averaging over both context durations) as within-subjects factors. Importantly, there was no significant effect of group, alone or in interaction. Therefore, elders performed the task with a speed similar to younger adults. The main effect of target [F (1, 31) =7.83, p<0.05], with no interaction with any other factor, confirms the general pattern for RTs reported previously.

Discussion

Overall, the results produce three major outcomes. Firstly, they provide a replication of the "unbinding" and "rebinding" effects in older adults. Secondly and more importantly, the unbinding effect appears larger in older adults compared with younger ones. Indeed, while fusion scores are similar in the coherent context, the increase in "ba" responses due to unbinding is around 39% in older adults vs. 21 % in younger adults. The rebinding dynamics

seem similar (around 3 syllables) though a direct comparison of the rebinding dynamics between groups could not be afforded in a mixed ANOVA.

Finally, the pattern of response times is similar in younger and older participants, with, as in all our previous experiments, a significant delay for "McGurk" stimuli compared with congruent "Ba" stimuli, independently on context and age.

General Discussion

The two experiments in this paper replicate and confirm the general pattern of unbinding/rebinding processes reported by Nahorna et al. (2012, 2015). They extend this pattern to the case of stimuli contaminated with acoustic noise in their contextual part (Experiment 1) and to older subjects (Experiment 2). It appears that noise in the context leads to a global decrease in the percentage of "ba" responses, hence a global increase in the rate of audiovisual fusion (Experiment 1) and that seniors display larger unbinding with incoherent context (Experiment 2).

We will first discuss the global coherence of the experimental data with the Binding-and-Fusion model introduced previously. Then, we will first address a number of methodological questions that are recurrently raised concerning the paradigm at work in this study and the previous ones, and discuss possible interpretations alternative to the Binding-and-Fusion model. Finally, we will come back to the role of noise and aging in this architecture.

Interpretation in the framework of the two-stage model

Independent on the role of noise and aging, the experimental data provide three major findings, recurrently displayed in Experiments 1 and 2 and perfectly in line with the previous study by Nahorna et al. (2015):

(1) *Modulation of the McGurk effect by context/reset type*. Applying an incoherent context strongly decreases fusion while a coherent reset after a period of incoherence enables

to come back to the original level. This is explained in the two-stage model by positing unbinding and rebinding processes. Context would vary the w_A and w_V weights at the output of the binding stage, hence modulate the amount of fusion displayed by the percentage of "ba" responses for "McGurk" targets. This effect is large, producing variations up to 21% for younger adults without noise and even 39% for elders.

(2) Stable differences between RTs for "Ba" and "McGurk" targets. "Ba" targets are detected 20 to 50 ms earlier than "McGurk" targets. The difference appears stable and independent on context/reset, context duration, noise, and age. Later responses for "McGurk" targets is likely related to their incongruence, delaying participants' responses. The important point is the stability of the difference, which will be of importance in the following.

(3) Shorter contexts (with smaller context duration, i.e. 2 syllables, or with no reset) lead to slower RTs compared to longer ones. This is accompanied for the 2syl context duration by an increase in "ba" responses hence a fusion decrease, small but systematic. A possible interpretation is that short contexts produce a surprise effect for the subject, the target arriving earlier than expected. This could result in decreasing audiovisual binding, just as cognitive load happens to decrease fusion in dual tasks experiments (Alsius et al., 2005, 2007).

Questions about the involved methodology

Are the monitoring paradigm and the two-alternative forced-choice task adequate for measuring fusion?

The interest of the monitoring paradigm is that it forces subjects to take their decision rapidly and in a situation where they do not precisely know when the target will happen. Hence they are forced to constantly process the audiovisual input, expecting an adequate audiovisual target. This is likely to enhance the role of binding and scene analysis processes, which is precisely the objective. However, it could be wondered whether the decision is

really based on fusion, or rather just on uncertainty about the audiovisual category associated to the coherent vs. conflicting auditory and visual inputs. In fact, this question has been recurrently asked since the beginning of studies about the McGurk effect and in her recent review, Tiippana (2014) notes that "it is impossible to be certain that the responses the observer gives correspond to the actual percepts" (p. 1). It has also been shown that changing the categorisation task, e.g. from open- to close-choice responses, modifies the amount and pattern of fusions (Colin et al., 2005). However, the fact that responses depend on context, in direct relation with the coherence and noise level of its audiovisual content, does show that the cognitive binding process changes, whatever the precise meaning of the subject's response. Therefore it can safely been considered that the variations of the amount of "ba" responses with context do provide a direct correlate of the binding process.

What do response times represent in this experiment?

The pattern of response times over conditions is remarkably stable among subjects (as displayed by the low standard deviations in Figure 3 and 4) and even among groups, as displayed by the lack of group effect in response times in Experiment 2. This last point is striking, considering the classical trend for slower responses in most perceptual tasks in older subjects (Ratcliff et al., 2001) – though it must be remembered that the amount of response misses does significantly increase in seniors. A possible interpretation is that response times are actually driven by evidence that the stimulus is finished and a new stimulus (including context + reset + target) will be played. Evidence is provided to the subject by the 200-ms transition stimulus inserted between reset and target (see Figure 2 and Supplementary Materials). Importantly, two contextual cues accelerate the participants' responses: longer contexts increase the probability that target should arrive, and noise drop from context to target provides a strong auditory cue in the noise condition in Experiment 1. Finally,

audiovisual incongruence plays a key role, leading to a remarkably stable difference in response times between congruent "Ba" and incongruent McGurk targets.

Do auditory or visual attention change globally across the experiment?

It can be asked whether the audiovisual context might result in a global decrease of auditory or visual attention in the target categorisation task. Firstly, it could be questioned whether incoherent context might decrease visual attention, and even encourage the participants to no more look at the visual scene during the task. Conversely, it could be wondered whether acoustic noise in the "noise" block could globally decrease auditory attention. However, the stability of the difference between response times for "Ba" and "McGurk" targets shows that both sensory inputs are accurately processed, leading to an increase in response time for incongruent inputs. More importantly, the global pattern of responses in Experiment 1 displays a complex portrait mixing effects of context, reset and noise. Simple auditory or visual shifts in auditory or visual attention cannot suffice to produce such a complex pattern of modulation of participants' responses. In fact, the Binding-and-Fusion model is precisely a way to computationally embed complex auditory and visual processes possibly related to attention within a general scene analysis process, able to explain the whole set of results in a single coherent architecture.

The role of sensory degradation in audiovisual fusion

Experiment 1 shows that adding acoustic noise before a "McGurk" target though not on the target itself dramatically increases the McGurk effect. It is unlikely that auditory target intelligibility could be modified by noise in the context. Indeed, the effects of forward masking are known to decrease to zero after 200 ms at most (Moore, 2004). Here, the 200-ms transition component between context/reset and target ensures that noise in the context/reset cannot decrease the audibility of the acoustic component of the target stimulus.

Therefore, our interpretation is that the effect occurs at the channel level in the fusion process (Hyp. 2). The addition of acoustic noise would contaminate the channel by making it less reliable, which would result in an increase of the relative reliability of the visual input (hatched arrows in Figure 1B). This suggests that audiovisual fusion is monitored by the output of two evaluation devices, the first one estimating audiovisual coherence (and decreasing visual weight in the case of incoherence) and the other estimating channel reliability (and increasing/decreasing channel weights in relation to their relative reliability).

A logical prediction from this hypothesis is that degrading the visual component of an audiovisual context stimulus presented before a "McGurk" target should, on the contrary, decrease fusion. In sum, the data by Sekiyama and Tohkura (1991) showing fusion increase in acoustic noise and those of Fixmer and Hawkins (1998) and Kim and Davis (2011) displaying fusion decrease in visual noise would be largely due to channel estimation effects weighting fusion accordingly.

This adds to a number of previous studies showing that audiovisual fusion is not automatic, but rather depends on subjects (Schwartz, 2010), language (Sekiyama & Tohkura, 1991), attention (Alsius et al., 2005; Tiippana et al., 2004) and context coherence (Nahorna et al., 2012, 2015). It suggests that human listeners are able to constantly evaluate the level of noise and the conditions of communication, and to monitor the audiovisual fusion process accordingly.

Aging and the potential role of attention in audiovisual fusion

Experiment 2 displays clear differences between young and old adults in the binding/unbinding/rebinding paradigm. There could of course exist differences in unisensory performances between groups, at the level of either audio or visual processing. However, the fact that the amount of fusion in the coherent context was similar between the older and

younger groups (see Figure 4) suggests that the difference is mainly due to the way the incoherent context was processed.

The experimental data show that the incoherence of the audio and video streams within context led older subjects to selectively decrease the role of the visual input in the fusion process more than younger ones. This could appear at odds with both the observation that seniors might exhibit more dependency on visual information (Sekiyama et al., 2014), and the hypothesis of larger internal sensory noise raised by Mozolic et al. (2012).

We propose to relate the increase in unbinding in seniors to the fact that under cognitive load, integration reduces (see Alsius et al., 2005; Alsius, Navarra, & Soto-Faraco, 2007). In these studies, participants engaged in a double task appear to experience a large decrease in their ability to bind together the incongruent auditory and visual stimuli characteristic of the McGurk effect. The authors' interpretation is that audiovisual binding would not be automatic but rather require a certain amount of cognitive attention to solve the cognitive problem caused by the audiovisual conflict. When attention is already engaged in a side task, binding would become more difficult. This is compatible with a number of perceptual phenomena reported to be pre-attentive and which in fact appear to be modulated and possibly totally erased when a concurrent task is proposed simultaneously (Alsius et al., 2005).

It is widely accepted that attentional processes are exploited specifically by seniors to compensate for sensory degradation to maintain cognitive performance as stable as possible (e.g. Cabeza, Anderson, Locantore, & McIntosh, 2002; Fullgrabe & Rosen, 2016). Therefore, the amount of available attention for audiovisual binding would be lower. This would not be problematic for small incongruence as displayed in "McGurk" stimuli. But in the case of large conflicts associated to incoherent contexts, it is likely that a larger amount of attention is required for keeping audition and vision bound together and hence produce binding despite

636 evidence that the auditory and visual streams are incongruent. If the ability to maintain this
637 amount of attention is decreased in seniors, this would result in less fusion and more
638 unbinding, which is actually what happens in the experiment. Therefore, it appears that aging
639 in this experiment produces an effect at the level of audiovisual fusion (hatched arrows in
640 Figure 1B, Hyp. 2) rather than just an increase in internal noise at the sensory level (gray
641 arrows in the same figure).

642 These data hence show for the first time a situation in which seniors would be poorer
643 than young adults in audiovisual integration. This could be of importance to explain part of
644 their difficulty in understanding their speaking partners in a complex audiovisual scene made
645 of interacting speakers (cocktail party effect). Indeed, the results of Experiment 2 suggest that
646 in this case, seniors might experience difficulty in keeping binding efficient in face of the
647 complex pattern of coherent and incoherent stimuli. This would result in strong unbinding
648 effects, in which the older participants would be led to disconnect the visual from the
649 auditory input, hence dramatically decreasing their ability to understand.

650 **Conclusions**

651 In this paper, we presented two studies on the audiovisual binding, confirming that
652 audiovisual fusion is *not* automatic but controlled by an audiovisual binding process prior to
653 fusion. This process would evaluate both the coherence of the auditory and visual inputs, and
654 the reliability of the auditory and visual channels, and weight the unisensory evidence
655 accordingly. The Binding-and-Fusion model would be part of a general audiovisual scene
656 analysis process enabling the speech perception system to extract and combine the adequate
657 pieces of information before decoding.

658 This system appears to be more fragile in senior subjects, as displayed by larger
659 unbinding effects in the case of incoherent contexts before a “McGurk” target. This could

partly explain the seniors' difficulties in understanding a conversation in a cocktail-party like situation.

Importantly, this suggests that audiovisual binding might be a plastic process, likely to display degradations with age. A matter of interest is its development in childhood. Furthermore, it could be questioned whether this system could, on the contrary, be subject to learning processes in which listeners would be guided to reinforce the efficiency of audiovisual binding, possibly increasing their ability to take profit of speechreading in adverse conditions. The Binding-and-Fusion process is hence an important topic for future research for both theoretical and practical reasons.

Acknowledgements

This research was funded by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s", PI J.L. Schwartz). This project has been supported by Academic Research Community "Quality of life and ageing" (ARC 2) of the Rhône-Alpes Region, which provided a doctoral funding for Ganesh Attigodu Chandrashekara.

References

Alain, C., & Woods, D. L. (1999). Age-related changes in processing auditory stimuli during visual attention: evidence for deficits in inhibitory control and sensory memory. *Psychology and Aging, 14*(3), 507-519. doi:10.1037/0882-7974.14.3.507

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*(9), 839-843. doi:10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research, 183*(3), 399-404. doi:10.1007/s00221-007-1110-1

Auer, E. T., Jr., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research, 50*(5), 1157-1165. doi:10.1044/1092-4388(2007/080)

Baltes, P. B., & Lindenberger, U. (1997). Emergence of a powerful connection between sensory and cognitive functions across the adult life span: a new window to the study of cognitive aging? *Psychology and Aging, 12*(1), 12-21. doi:10.1037/0882-7974.12.1.12

Behne, D., Wang, Y., Alm, M., Arntsen, I., Eg, R., & Valso, A. (2007). *Changes in auditory-visual speech perception during adulthood.* Paper presented at the Proceedings of AVSP 2007.

Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research, 37*(5), 1195-1203. doi:10.1044/jshr.3705.1195

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception and Psychophysics, 62*(2), 233-252. doi:10.3758/BF03205546

Berthommier, F. (2004). A phonetically neutral model of the low-level audio-visual interaction. *Speech Commun, 44*(1–4), 31-41. doi:10.1016/j.specom.2004.10.003

Buchan, J. N., & Munhall, K. G. (2012). The effect of a concurrent working memory task and temporal offsets on the integration of auditory and visual speech information. *Seeing Perceiving, 25*(1), 87-106. doi:10.1163/187847611X620937

Cabeza, R., Anderson, N. D., Locantore, J. K., & McIntosh, A. R. (2002). Aging gracefully: compensatory brain activity in high-performing older adults. *Neuroimage, 17*(3), 1394-1402. doi:10.1006/nimg.2002.1280

CHABA. (1988). Speech understanding and aging. Working Group on Speech Understanding and Aging. Committee on Hearing, Bioacoustics, and Biomechanics, Commission on Behavioral and Social Sciences and Education, National Research Council. *Journal of the Acoustical Society of America, 83*(3), 859-895.

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear and Hearing, 23*(5), 439-449. doi:10.1097/01.aud.0000034781.95122.15

Dancer, J., Krain, M., Thompson, C., Davis, P., & et al. (1994). A cross-sectional investigation of speechreading in adults: Effects of age, gender, practice, and education. *The Volta Review, 96*(1), 31-40.

Erber, N. P. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *Journal of Speech, Language, and Hearing Research, 12*(2), 423-425. doi:10.1044/jshr.1202.423

Feld, J. E., & Sommers, M. S. (2009). Lipreading, Processing Speed, and Working Memory in Younger and Older Adults. *Journal of speech, language, and hearing research : JSLHR, 52*(6), 1555-1565. doi:10.1044/1092-4388(2009/08-0137)

Fixmer, E., & Hawkins, S. (1998). *The influence of quality of information on the McGurk effect.* Paper presented at the Proceedings of AVSP 1998, Terrigal, Australia.

728 Fullgrabe, C., & Rosen, S. (2016). Investigating the Role of Working Memory in Speech-in-
729 noise Identification for Listeners with Normal Hearing. *Advances in Experimental*
730 *Medicine and Biology, 894*, 29-36. doi:10.1007/978-3-319-25474-6_4
731 Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by
732 hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-
733 visual integration. *Journal of the Acoustical Society of America, 103*(5 Pt 1), 2677-
734 2690. doi:10.1121/1.422788
735 Hay-McCutcheon, M. J., Pisoni, D. B., & Kirk, K. I. (2005). Audiovisual speech perception
736 in elderly cochlear implant recipients. *Laryngoscope, 115*(10), 1887-1894.
737 doi:10.1097/01.mlg.0000173197.94769.ba
738 Huyse, A., Berthommier, F., & Leybaert, J. (2013). Degradation of labial information
739 modifies audiovisual speech perception in cochlear-implanted children. *Ear and*
740 *Hearing, 34*(1), 110-121. doi:10.1097/AUD.0b013e3182670993
741 Kim, J., & Davis, C. (2011). *Audiovisual speech processing in visual speech noise.* Paper
742 presented at the Proceedings of AVSP 2011, Volterra, Italy.
743 Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of
744 speech perception. *Cognitive Psychology, 21*(3), 398-421. doi:10.1016/0010-
745 0285(89)90014-5
746 Massaro, D. W. (1998). *Perceiving Talking Faces*. Cambridge: MIT Press.
747 McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588),
748 746-748. doi:10.1038/264746a0
749 Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the
750 superior colliculus. *Science, 221*(4608), 389-391. doi:10.1126/science.6867718

751 Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on
752 cells in superior colliculus results in multisensory integration. *Journal of*
753 *Neurophysiology, 56*(3), 640-662.
754 Moore, B. (2004). *An Introduction to the Psychology of Hearing* (5th Edition ed.). Oxford:
755 Elsevier.
756 Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., & Laurienti, P. J. (2012). Multisensory
757 Integration and Aging. In M. M. Murray & M. T. Wallace (Eds.), *The Neural Bases of*
758 *Multisensory Processes*. Boca Raton (FL): CRC Press/Taylor & Francis Llc.
759 Nahorna, O., Berthommier, F., & Schwartz, J. L. (2012). Binding and unbinding the auditory
760 and visual streams in the McGurk effect. *Journal of the Acoustical Society of*
761 *America, 132*(2), 1061-1077. doi:10.1121/1.4728187
762 Nahorna, O., Berthommier, F., & Schwartz, J. L. (2015). Audio-visual speech scene analysis:
763 characterization of the dynamics of unbinding and rebinding the McGurk effect.
764 *Journal of the Acoustical Society of America, 137*(1), 362-377.
765 doi:10.1121/1.4904536
766 Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive
767 processing: implications for hearing aid fitting and audiologic rehabilitation. *Trends*
768 *Amplif, 10*(1), 29-59. doi:10.1177/108471380601000103
769 Poliakoff, E., Ashworth, S., Lowe, C., & Spence, C. (2006). Vision and touch in ageing:
770 crossmodal selective attention and visuotactile spatial interactions. *Neuropsychologia,*
771 *44*(4), 507-517. doi:10.1016/j.neuropsychologia.2005.07.004
772 Schwartz, J. L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in
773 speech perception is subject-dependent. *Journal of the Acoustical Society of America,*
774 *127*(3), 1584-1594. doi:10.1121/1.3293001

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E), 15*(3), 143-158. doi:10.1250/ast.15.143

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology, 5*, 323. doi:10.3389/fpsyg.2014.00323

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90*(4 Pt 1), 1797-1805. doi:10.1121/1.401660

Setti, A., Burke, K. E., Kenny, R., & Newell, F. N. (2013). Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes. *Frontiers in Psychology, 4*, 575. doi:10.3389/fpsyg.2013.00575

Shoop, C., & Binnie, C. A. (1979). The Effects of Age Upon the Visual Perception of Speech. *Scand Audiol, 8*(1), 3-8. doi:10.3109/01050397909076295

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263-275. doi:10.1097/00003446-200506000-00003

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643-662. doi:10.1037/h0054651

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America, 26*(2), 212-215. doi:10.1121/1.1907309

Thompson, L. A. (1995). Encoding and memory for visible speech and gestures: a comparison between young and older adults. *Psychology and Aging, 10*(2), 215-228. doi:10.1037/0882-7974.10.2.215

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457-472. doi:10.1080/09541440340000268

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28*(5), 656-668. doi:10.1097/AUD.0b013e31812f7185

Walden, B. E., Busacco, D. A., & Montgomery, A. A. (1993). Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons. *Journal of Speech and Hearing Research, 36*(2), 431-436. doi:10.1044/jshr.3602.431

Supplementary Material

Detailed description of stimuli and experimental paradigm in Experiments 1 and 2

Audiovisual material

We utilized the material that was prepared for the initial audiovisual binding experiments by (Nahorna, Berthommier, & Schwartz, 2012, 2015). The stimuli for all experiments were prepared from two sets of audiovisual material, a "syllable" material and a "sentence" material, produced by a French male speaker with lips painted in blue to allow precise video analysis of lip movements (Lallouache, 1990). The recordings were carried out in a soundproof room. Stereo soundtracks were digitized in Adobe Audition at 44.1 kHz with 16-bit resolution. Videos were edited in Adobe Premiere Pro into a 720/576 pixels movie with a digitization rate of 25 frames/s (1frame = 40 ms).

The stimuli in the "syllable" material consisted of successive French syllables randomly selected within the set "pa", "ta", "va", "fa", "za", "sa", "ka", "ra", "la", "ja", "cha", "ma", "na" – before producing a final syllable in the set "ba", "da" or "ga". The speaker produced the syllables with a short temporal gap between two consecutive syllables enabling easy cuts for stimuli preparation, with a mean syllable duration (including temporal gaps) of 700 ms (typically varying between 650 and 750 ms). The stimuli in the "sentence" material consisted of sequences of sentences freely uttered by the speaker during the recording session. These two materials were used to prepare either coherent contexts made of coherent audiovisual excerpts from the "syllable" material or incoherent contexts dubbing sounds from the "syllable" material with video coming from the "sentence" material. The final syllables "ba" or "ga" in the "syllable" material were extracted and utilized to construct the target stimuli.

Context

The coherent context was made of 2 or 4 audiovisual syllables extracted from the "syllable" material. The incoherent context was prepared by dubbing a sequence of 2 or 4 acoustic syllables extracted from the "syllable" material (same syllables that were used in preparing the coherent context) on a video stream extracted from the "sentence" material with the adequate duration. The durations of 2 or 4 syllables have been shown by (Nahorna et al., 2015) to be sufficient to produce maximal effects of the incoherent context compared with the coherent one that is a maximal decrease of the McGurk effect. Indeed, longer incoherent contexts produce the same decrease compared with coherent context. Sound and video files were automatically extracted from the audiovisual material with the desired length using Matlab (Mathworks, Natick, MA, USA).

Reset

The reset stimulus, which was always presented after the incoherent context, consisted of 0, 1, 2 or 3 coherent audiovisual syllables extracted from the "syllable" material. The "0" syllable reset was nothing but pure incoherent context where there was no reset material presented. Visual continuity between context and reset was achieved by a linear transition between the last three images of the context and the first two images of the reset. In the statistical analyses that will be presented later, the stimuli were grouped into context/reset type (5 variants: coherent vs. incoherent with 0, 1, 2 or 3-syllables reset) and context duration (2 vs. 4 syllables).

Target

The target was either a congruent audiovisual "Ba" syllable or an incongruent McGurk stimulus. The McGurk stimuli were prepared from an audio occurrence of the "ba" syllable dubbed on the sequence of images of an occurrence of the "ga" syllable. The audio "ba" and video "ga" were synchronized by using the precise temporal localization of the acoustic bursts of the original "ba" and "ga" stimuli, obtained with the Praat software

(Boersma & Weenink, 2014). The same set of audiovisual targets was associated with either coherent or incoherent context.

To construct various combinations of context and target from the audiovisual materials, we need to join different sequences of images from the "syllables" and "sentences" material. This could create abrupt breaks and thus continuity could be lost. To ensure continuity between context+rest and target, a 200 ms transition stimulus (5 images) was inserted 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. This transition stimulus provided a small cue for the arrival of the target stimulus. From this cue, the acoustic burst of the target stimulus arrived precisely 240 ms later (see Figure S1).

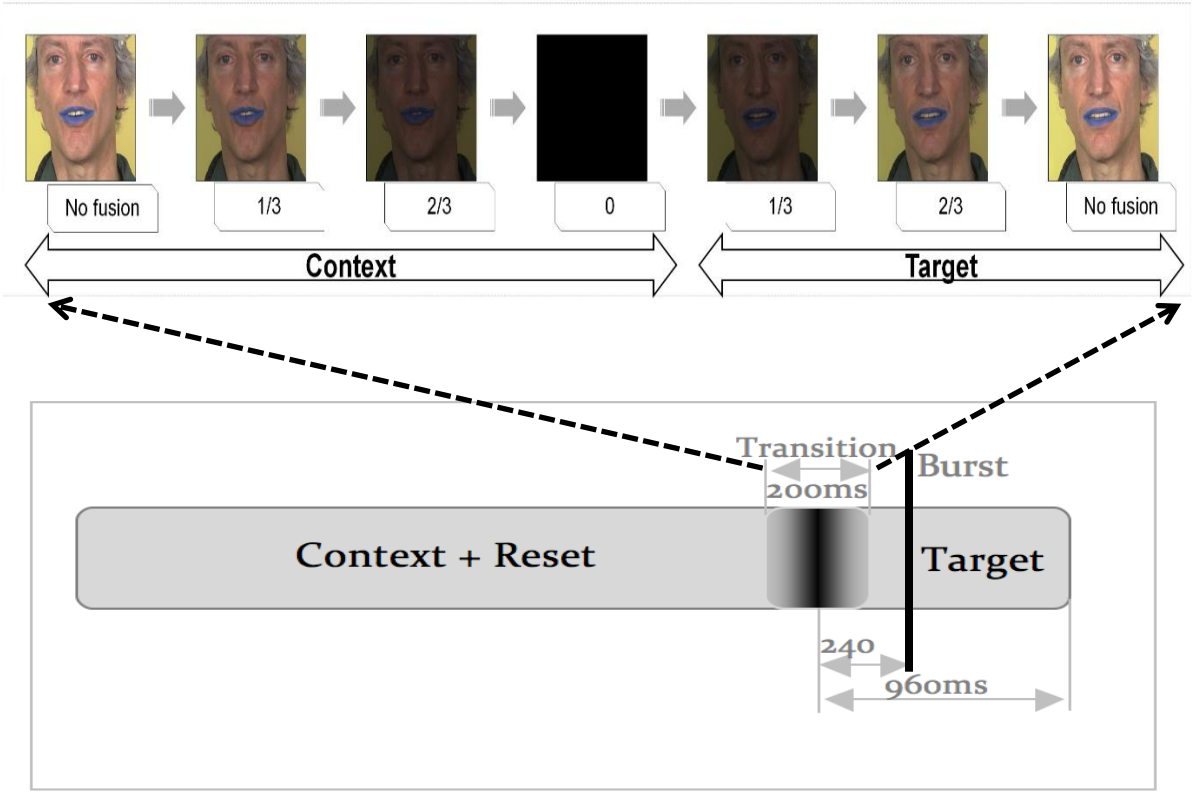


Figure S1 – Fusion between context + reset and target.

Addition of noise on the context and reset parts in Experiment 1

The target stimuli were always presented without acoustic noise in all conditions. However, in one condition in Experiment 1, acoustic noise was added to the context and reset

periods of the stimuli (see Figure 2). We used Gaussian white noise at 0 dB SNR generated using Matlab (Mathworks, Natick, MA, USA). SNR values were computed on the portions of the speech input removing all silent portions between syllables.

Final audiovisual film preparation

Stimuli were mixed randomly to produce films containing all possible stimuli in a given experiment. An 840-ms inter-stimulus silent interval was inserted between the end of one (context+reset+target) stimulus and the beginning of the next one. 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 and necessity to constantly search for new targets, to decrease as much as possible post-decision biases on target detection. All the auditory stimuli were normalized to keep the same mean energy for all "contexts" and "targets" stimuli throughout the experiment.

Experiment 1 consisted in two blocks, one without acoustic noise and the other one with acoustic noise. As explained previously, McGurk targets were presented three times more than congruent "Ba" targets, which served as controls. For each (context+reset) condition (2 context durations; coherent context + incoherent context with 4 possible reset durations; 2 noise conditions; hence altogether 20 conditions) there were 4 occurrences of a "Ba" target and 12 occurrences of a McGurk target. Hence there were 320 sequences in total, spread over 2 blocks of 10 min each, one for each noise condition (see Table S1). All stimuli were randomized and we prepared five different films with five different orders in each block. The five different films were randomly distributed among the subjects. Experiment 2 comprised only the block without noise.

	2-syl context duration		**4-syl context duration**	
Targets	Coherent context	Incoherent context with reset of	Coherent context	Incoherent context with reset of

		0 syl	1 syl	2 syl	3 syl		0 syl	1 syl	2 syl	3 syl
"Ba"	4	4	4	4	4	4	4	4	4	4
"McGurk"	12	12	12	12	12	12	12	12	12	12

Table S1 - Number of stimuli presented for each condition in each block (without noise or with noise).

Procedure

All experiments were carried out in a soundproof booth. Stimulus presentation was coordinated with the Presentation® software (Neurobehavioral Systems Inc., Albany, CA). The participant's task was to monitor for the arrival of target stimuli "ba" or "da" within the displayed films, by pressing as soon as possible the appropriate key (two-alternative-forced-choice identification task). This is different from classical speech recognition tests where participants know when the target stimuli will be presented. Participants were instructed to look constantly at the screen and, each time a "ba" or a "da" was perceived, to press the corresponding button immediately. The response button was evenly interchanged between subjects.

The distance of the participant to the screen at about 50 cm from the screen and the intensity of the audio stimulus were kept fixed. The films were presented on a computer monitor with high-fidelity headphones set at a comfortable fixed level. Trial sessions were provided before each block to enable participants to familiarize with stimuli and task. In Experiment 1 comprising two blocks, the order of the blocks was counterbalanced across participants.

Detection of responses

The expectation in this monitoring task was that for each congruent "Ba" target the participants should detect a "ba", while for each incongruent "McGurk" target they should detect either a "ba" or a "da". Since the context material contained no "ba", "da" or "ga" in

the audio stream, we expected that no target should be detected during the context periods. However, such an online monitoring task may lead to either wrong detections – that is the detection of “ba” or “da” during the context – or failure of target detection. Therefore, the first step in the analysis process was to define a protocol for detecting responses to target stimuli.

Each subject’s response was associated with a temporal value provided by the Presentation® software (Neurobehavioral Systems Inc., Albany, CA). The response time was evaluated by the difference in milliseconds between this value and the acoustic onset of target syllables – defined as the plosive burst onset. Any response provided with a response time larger than 1200 ms, or smaller than 200 ms, was considered as a false detection and discarded from the analysis. The value of 1200 ms has been proposed from the analysis of response time histograms (Nahorna et al., 2012), showing that it enabled to accept most responses while discarding spurious responses that could actually be due to the beginning of the next context period (remember that the inter-stimulus interval was short, namely 840 ms). We systematically report the number of missed targets and show that indeed most targets are detected by the participants in all experiments. In the cases of double responses within the acceptable [200-1200] window, we accepted the first response together with its corresponding response time if both responses were the same, and rejected the response in case of two different responses. All the possible outcomes are described in Figure S2.

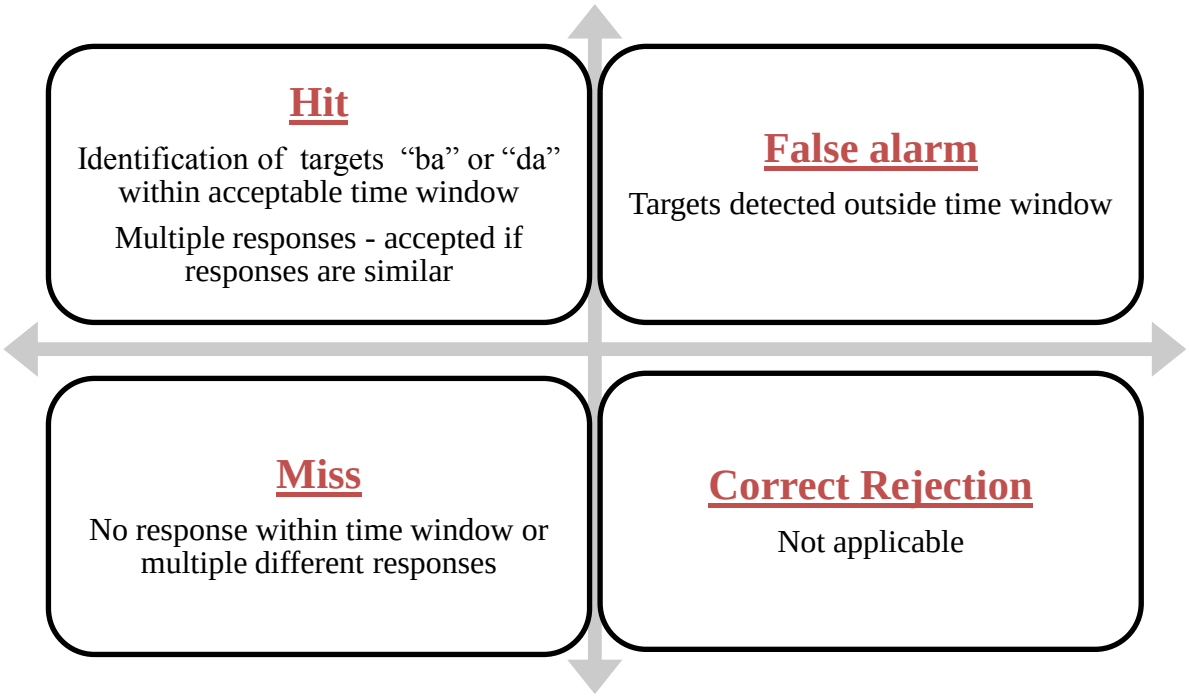


Figure S2 - Classification of responses.

Detailed statistical results

Analysis of responses

For each experiment, the total number of "ba" and "da" hit responses was calculated for each condition of context and target and for each participant. Then the percentage of "ba" responses – that is the ratio "total number of ba responses" divided by "total number of ba or da responses" – was taken as the score of responses by this participant in this context for further statistical analyses presented in the next section. The number of "no responses" and "multiple different responses" within the acceptable time window was also systematically computed.

Analysis of response time

For each experiment, for each condition of target and context and for each participant, the mean response time was estimated by averaging the response times for all stimuli in the corresponding condition.

Statistical Analysis

954 The suitable statistical analysis was performed on both response scores ("ba/ (ba+da)"
955 scores) and response times (mean response times) using the SPSS Statistics 17 © IBM
956 software. The response scores to the "Ba" targets were systematically close to 100%, and not
957 considered in the analysis since these targets only served as a control stimulus. To ensure
958 quasi-Gaussian distribution of the variables, the response scores were processed with arcsine
959 square root transformation [asin (sqrt)] transform, and the mean response times were
960 logarithmically transformed. Then analyses of variance (ANOVAs) were performed on both
961 transformed response scores and transformed response times, applying a Greenhouse –
962 Geisser correction in case of violation of the sphericity assumption. *Post-hoc* analyses with
963 Bonferroni correction were done when appropriate and were reported at the [($p < 0.05$) level].

964 ***Experiment 1***

965 *Analysis of missing responses*

966 The pattern of missing responses (Figure S2) is displayed in Table S2 averaged over
967 the 29 subjects. Overall, the mean percentage of missed targets amounted to 6.4%. The
968 "without-noise condition" led to lesser misses (3.8%) compared with the "with-noise
969 condition" (9.1%). The difference between noise conditions was significant (Wilcoxon signed
970 rank test for paired data over the 29 subjects, $V = 80.5$, $p = 0.016$). These values largely
971 varied between subjects. Without noise, the miss rate averaged over all conditions of target,
972 context/reset type and context duration varied among subjects from 0% to 30%. With noise, it
973 varied from 0% to 51%. Only subjects with values of these average miss rates lower than
974 25% in both the without-noise and the with-noise condition were kept in further analyses
975 (providing 23 remaining subjects).

Condition	Context duration	Target	Coherent context		Incoherent context with reset of							
					0 syl		1 syl		2 syl		3 syl	
			NR	MR	NR	MR	NR	MR	NR	MR	NR	MR

Without noise	2 syl	"Ba"	0.87	2.58	2.59	1.72	0.86	5.17	0.00	2.59	0.86	2.59
		"McGurk"	2.01	2.87	2.01	3.74	2.30	1.72	1.72	2.30	1.44	4.02
	4 syl	"Ba"	0.00	3.45	0.86	0.00	0.86	0.86	0.86	2.59	2.59	2.59
		"McGurk"	1.15	3.74	0.57	1.44	0.57	2.59	1.44	2.01	0.86	3.45
With noise	2 syl	"Ba"	0.00	6.03	0.00	7.76	2.59	8.62	0.86	6.90	1.72	8.62
		"McGurk"	0.86	7.18	0.86	8.62	0.86	5.46	0.57	7.47	1.15	10.06
	4 syl	"Ba"	0.86	13.79	0.00	12.93	0.00	6.90	0.00	12.07	0.00	6.90
		"McGurk"	0.57	6.03	0.86	8.05	0.29	8.05	0.57	9.20	0.57	7.76

Table S2 - Mean number of missed targets averaged over the 29 subjects for each condition of noise, target, context/reset type and context duration (NR= No response in %, MR=Multiple different responses in %).

Analysis of the proportion of "ba" responses

A repeated-measures ANOVA with three factors, context/reset type (coherent vs. incoherent with 4 reset durations, hence 5 possibilities altogether), context duration (two vs. four syllables) and noise (with noise vs. without noise) was realized on the proportion of "ba" responses for McGurk targets. Detailed results of the analysis are reported in Table S3. *Post-hoc* results are displayed in Table S4.

Source	**d.f=F**	**Sig.**	**η2**	**η2 within**
Noise (with noise vs. without noise)	(1, 22) =21.81	.000	0.0649	0.3036
Context duration (2 syl vs. 4 syl)	(1. 22)=13.35	.001	0.0028	0.0132
Context/Reset nature (Coherent, 0, 1, 2 & 3 syl reset duration)	(4, 88) =12.29	.000	0.0105	0.0491
Noise * Context duration	(1, 22) = 3.08	.093	0.0003	0.0018
Noise * Context/Reset nature	(4, 88) = 2.61	.060	0.0015	0.0072
Context duration * Context/Reset nature	(4, 88) = 2.85	.049	0.0017	0.0079

Noise*Context duration * Context/Reset nature	(4, 88) = 3.07	.027	0.0016	0.0078
Subjects	(1. 22)=40.59	.000	0.7860	

985 **Table S3-** Detailed results of the three-way repeated-measures ANOVA for response scores for the McGurk
986 target.
987 Effect sizes are estimated by eta-squared values. Since the contribution of inter-subject variance, displayed in
988 the last line of the Table, is extremely high, we also provide eta-squared values among within-subject factors,
989 removing the contribution of the "subjects" factor in the computation of the total variance.
990

Tested Effect	**Tested Variable**	**Post-Hoc Results**
Noise		Without noise > With noise**
Context duration		2 syl > 4 syl*
Context/Reset nature		0 syl >1 syl, 2 syl, 3 syl & coherent context* 1, 2, 3 syl & coherent context (n.s.)
Context Duration * Context/Reset nature	0 syl	2 syl > 4 syl*
	Coherent context	2 syl > 4 syl*
	2 syl	0 syl > 1, 2, 3 syl & coherent context * coherent context > 3syl*
	4 syl	0 syl > 1 syl & coherent context *
Context/Reset nature *Noise*Context Duration		
Between Noise	2 syl	0 syl & coherent context (Without noise > With noise) ** 1, 2 & 3 syl (Without noise > With noise) *
	4 syl	0, & 2 syl (Without noise > With noise) ** 1, 3 syl & coherent context (Without noise > With noise) *
Between context duration	Without noise	0 & coherent context (2 syl > 4 syl) ** 1, 2 & 3 syl (n.s.)
	With noise	0, 1, 2, 3 syl & coherent context (n.s.)
Between Context/Reset nature	Without noise	2 syl (0 syl>1, & 2 syl) * 2 syl (0 syl>3 syl & coherent context) ** 2 syl (1 syl > 3syl)*

		2 syl (coherent context > 3syl)* 4 syl (0 syl > 1 syl & coherent context) *
	With noise	2 & 4syl (0 syl >1, 2, 3 syl & coherent context) n.s.

Table S4. *Post-hoc* analysis for response scores for the McGurk target (**=p<0.001, *=p<0.05, n.s= not significant).

Analysis of response times

Response times were analyzed in a four-way repeated-measures ANOVA with factors target ("Ba" vs. McGurk), context/reset type (coherent vs. incoherent with 4 reset durations, hence 5 possibilities altogether), context duration (two vs. four syllables) and noise (with noise vs. without noise). Detailed results are reported in Table S5 and *post-hoc* results in Table S6.

Source	**d.f=F**	**Sig.**	**η2**	**η2 within**
Noise (with noise vs. without noise)	(1, 22) = 51.55	.000	0.0899	0.2308
Targets ("Ba" vs. "McGurk")	(1, 22) = 23.94	.000	0.0205	0.0526
Context Duration (2 syl vs. 4 syl)	(1, 22) = 11.33	.003	0.0062	0.0159
Context/Reset Nature (coherent, 0, 1, 2 & 3 syl reset durations)	(4, 88) = 6.40	.001	0.0056	0.0144
Noise*Target	(1, 22) = 1.38	.251	0.0004	0.0012
Noise*Context Duration	(1, 22) = 12.18	.002	0.0029	0.0076
Target*Context Duration	(1, 22) = .834	.371	0.0002	0.0005
Noise*Target*Context Duration	(1, 22) = 1.55	.226	0.0005	0.0013
Noise* Context/Reset Nature	(4, 88) = 1.72	.179	0.0020	0.0052
Target* Context/Reset Nature	(4, 88) = .688	.586	0.0004	0.0011
Noise*Target* Context/Reset Nature	(4, 88) = .903	.444	0.0007	0.0019
Context Duration* Context/Reset Nature	(4, 88) = .884	.425	0.0009	0.0024
Noise*Context Duration* Context/Reset Nature	(4, 88) = 1.89	.138	0.0014	0.0038
Target*Context Duration* Context/Reset Nature	(4, 88) = 1.53	.224	0.0014	0.0037

Noise*Target*Context Duration* Context/Reset Nature	(4, 88) = 1.35	.259	0.0011	0.0030
Subjects	(1, 22) =19551.70	.000	0.6103	

Table S5 - Detailed results of the four-way repeated-measures ANOVA for response times.

Tested Effect	Tested Variable	Post-Hoc Results
Noise		Without noise > With noise**
Target		Ba < McGurk**
Context duration		4 syl < 2 syl*
Context/Reset nature		0 syl > 2syl* 2 & 3 syl < coherent context*

Table S6 -*Post-hoc* analysis for response time for the ba & McGurk target (**=p<0.001, *=p<0.05, n.s= not significant).

Experiment 2

Analysis of missing responses

17 senior subjects were kept in the experiment, under the criterion to display less than 90% "ba" responses in the "coherent condition" for "McGurk" targets, averaged over the two context durations. The pattern of missing responses (see Figure S2) averaged over these 17 subjects is displayed in Table S7. Overall, the mean percentage of missed targets amounted to 11.5%.

The difference between the amount of missing responses between the 17 selected old participants and the 20 selected young participants – under the same criterion to display less than 90% "ba" responses in the "coherent condition" for "McGurk" targets, averaged over the two context durations – was significant for both conditions of noise (Wilcoxon signed rank test: young without noise vs. old, W = 284, p = 0.0005; young with noise vs. old, W = 268, p = 0.003). Miss amount varied largely between subjects, with values averaged over all conditions of target, context/reset type and context duration varying from 0% to 41%. Two

senior subjects displaying an average miss rate higher than 25% were discarded, hence only 15 senior subjects were kept in further analyses.

Context duration	Target	Coherent context		Incoherent context with reset of							
				0 syl		1 syl		2 syl		3 syl	
		NR	MR	NR	MR	NR	MR	NR	MR	NR	MR
2 syl	"Ba"	3.33	11.67	3.33	13.33	0.00	28.33	0.00	6.67	1.67	6.67
	"McGurk"	5.00	11.67	0.56	7.78	2.78	5.00	1.67	16.11	1.11	8.33
4 syl	"Ba"	0.00	15.00	3.33	1.67	0.00	6.67	0.00	13.33	0.00	6.67
	"McGurk"	0.00	7.78	0.00	7.22	0.00	8.89	0.00	15.56	1.11	8.33

Table S7 - Mean number of missed targets averaged over the 15 subjects for each condition of target, context/reset type and context duration (NR= No response in %, MR=Multiple responses in %)

Analysis of the proportion of "ba" responses for the 15 seniors

A repeated-measures ANOVA with two factors, context/reset type (coherent vs. incoherent with 4 reset durations, hence 5 possibilities altogether), and context duration (two vs. four syllables) was realized on the proportion of "ba" responses for McGurk targets. Detailed results of the analysis are reported in Table S8. *Post-hoc* results are displayed in Table S9.

Source	d.f=F	Sig.	η 2	η 2 within
Context Duration (2 syl vs. 4 syl)	(1, 14) = 11.01	.005	0.0065	0.0207
Context/Reset Nature (coherent, 0, 1, 2 & 3 syl reset durations)	(4, 56) = 39.40	.000	0.1728	0.5485
Context Duration*Context/Reset Nature	(4, 56) = 2.34	.076	0.0056	0.0179
Subjects	(1, 14) = 37.83	.000	0.6849	

Table S8 - Detailed results of the two-way repeated-measures ANOVA for response scores.

Tested Effect	Tested Variable	Post-Hoc Results

Context duration		2 syl > 4 syl*
Context/Reset nature		0 syl > 3 syl & coherent context** 0 syl > 1 & 2 syl* 2 syl > 3 syl & coherent context**

Table S9 - *Post-hoc* analysis for response scores for the McGurk target (**=p<0.001, *=p<0.05, n.s= not significant).

Analysis of response times for the 15 seniors

Response times were analyzed in a three-way repeated-measures ANOVA with factors target ("Ba" vs. McGurk), context/reset type (coherent vs. incoherent with 4 reset durations, hence 5 possibilities altogether), and context duration (two vs. four syllables). Detailed results are reported in Table S10 and *Post-hoc* results in Table S11.

Source	d.f=F	Sig.	η2	η2 within
Targets ("Ba" vs. "McGurk")	(1, 14) = 5.38	.036	0.0147	0.0282
Context Duration (2 syl vs. 4 syl)	(1, 14) =26.53	.000	0.0278	0.0531
Context/Reset Nature (coherent, 0, 1, 2 & 3 syl reset durations)	(4, 56) = 3.80	.016	0.0308	0.0587
Target*Context Duration	(1, 14) = 1.30	.272	0.0015	0.0029
Target*Context/Reset Nature	(4, 56) = 0.35	.773	0.0022	0.0043
Context Duration*Context/Reset Nature	(4, 56) = 2.14	.130	0.0085	0.0162
Target*Context Duration*Context/Reset Nature	(4, 56) = 0.66	.560	0.0028	0.0054
Subjects	(1, 14) = 17284.11	.000	0.4754	

Table S10 - Detailed results of the three-way repeated-measures ANOVA for response scores for the McGurk target.

Tested Effect	Tested Variable	Post-Hoc Results
Target		Ba < McGurk*

Context duration		4 syl < 2 syl**
Context/Reset nature		0, 1, 2, 3 syl & coherent context (n.s.)

Table S11 - *Post-hoc* analysis for response time for the ba & McGurk target (**=p<0.001, *=p<0.05, n.s= not significant).

Comparing youngers and seniors

A two-way mixed ANOVA with age as the between-group variable (18 younger vs. 15 older participants), and context/reset type (coherent vs. 0syl incoherent), as the within-subject variable was realized on the proportion of "ba" responses for McGurk targets. Detailed results of the analysis are reported in Table S12. *Post-hoc* results are displayed in Table S13.

Source	**d.f=F**	**Sig.**	**η2**
Context (coherent, & 0 syl incoherent)	(1, 31) = 169.71	.000	0.2615
Context*Group	(1, 31) = 23.15	.000	0.0357
Group (young vs. older adult)	(1, 31) = 0.92	.244	0.0189

Table S12- Detailed results of the two-way mixed ANOVA for response scores for the McGurk target.

	Tested Variable	**Post-Hoc Results**
Context		0 syl > coherent context**
Context*Group	0 syl	older > younger*
	older	0 syl >coherent condition**
	younger	0 syl >coherent condition**

Table S13-*Post-hoc* analysis for response scores for the McGurk target (**=p<0.001, *=p<0.05, n. s= not significant).

Response times were analyzed in a three-way mixed ANOVA with age as the between-group variable (young vs. adult), and target ("Ba" vs. "McGurk") and context as the within-subject variable (0 syl incoherent vs. coherent condition) was realized on the

proportion of "ba" responses for McGurk targets. Detailed results of the analysis are reported in Table S14. *Post-hoc* results are displayed in Table S15.

Source	d.f=F	Sig.	η2
Target ("Ba" vs. McGurk)	(1, 31) = 7.83	.009	0.0317
Target*Group	(1, 31) = 0.95	.337	0.0038
Context/Reset Nature (coherent, 0, 1, 2 & 3 syl reset durations)	(1, 31) = 0.21	,646	0.0007
Context/Reset Nature*Group	(1, 31) =0.31	.579	0.0010
Context/Reset Nature*Target	(1, 31) = 0.31	.581	0.0009
Context/Reset Nature*Target*Group	(1, 31) = 0.49	.487	0.0014
Group (young vs. older adult)	(1, 31) = 0.40	.532	0.0083

Table S14 - Detailed results of the three-way mixed ANOVA for response time for the Ba & McGurk target.

Tested Effect	Tested Variable	Post-Hoc Results
Target		Ba < McGurk*

Table S15 - *Post-hoc* analysis for the three-way mixed ANOVA for response time for the Ba & McGurk target.

Subjective assessment of hearing for senior participants (Experiment 2)

In addition to the screening audiometry, we also administered a French version of the Speech, Spatial, and Qualities of hearing scale (SSQ; Gatehouse & Noble, 2004) which is a self-reported questionnaire developed to assess how effectively auditory information is being processed in various everyday listening situations. Recently, this questionnaire has been validated in the French language and found good reproducibility of scores. Inter-subject variability was obtained between French and other languages including the English version that was primarily developed (Moulin, Pauzie, & Richard, 2015) and it was concluded that the SSQ has potential to be used as an International standard for hearing disability evaluation.

The SSQ includes questions related to speech in quiet and noise, ASA, cognitive abilities and similar abilities which are very relevant to our experimental paradigm (e.g. question on multiple speech streams: "You are listening to someone talking to you, while at the same time trying to follow the news on TV. Can you follow what both people are saying?"). For both "Speech Hearing" items, and "Qualities Hearing" items, participants were instructed to estimate their abilities by selecting an 11-point response scale ranging from "0" (complete disability) to "10" (no disability). Overall, we obtained average scores respectively equal to 7.8 out of 10 for the "Speech Hearing" sub-scale and 8.6 out of 10 for the "Qualities Hearing" sub-scale, to compare to mean scores from 8.4 to 8.6 in the older English-speaking population (Füllgrabe, Moore, & Stone, 2015) and from 9 to 9.5 for the younger French-speaking population (Moulin et al., 2015).

Cognitive assessment of executive functions

In order to measure participant's attentional control, cognitive flexibility, and processing speed, we administered the French version of the color-word Stroop task (Stroop, 1935), a very popular measure in the neuropsychological and cognitive domain, considered to measure various executive functions such as selective attention and cognitive flexibility (Charchat-Fichman & Oliveira, 2009; Homack & Riccio, 2004), interference control (van Mourik, Oosterlaan, & Sergeant, 2005), response inhibition (Pocklington & Maybery, 2006) and brain's processing speed (Lamers, Roelofs, & Rabeling-Keus, 2010). We administered the color-word Stroop test, with two conditions, Word naming with incongruent (word "red" written in blue ink) and neutral stimuli (word "red" written in gray color), and Color naming with incongruent (word "blue" written in red ink) and neutral stimuli (list of "X"s in red ink). Stroop Interference (incongruent responses–neutral stimuli) was calculated for both word naming and color naming tasks. It amounted to 152 ms for incongruent color naming and 34 ms for word naming, respectively, with 5.2% errors for incongruent color naming and 4.0 %

errors for incongruent word naming. Overall, the Stroop Interference, mean reaction time as well as error rate for both word naming and color naming were within the normal range when compared to other similar studies on normal older healthy adults (Hutchison, Balota, & Duchek, 2010; Spieler, Balota, & Faust, 1996). For example, Spieler et al. (1996) obtained Stroop Interference ranges around 175-177 ms for color naming, and 19-43 ms for word naming, and error rates for color naming ranging from 1.3 to 3.8% for the neutral condition and from 3.9 to 7.2% for the incongruent condition. Our data suggest that all the participants may have normal processing speed and executive functional skills.

Correlations with cognitive variables

Pearson product-moment correlation coefficients were computed to assess the relationship between the SSQ and Stroop values for senior participants and a number of characteristics of their behavior in Experiment 2 (e.g. mean amount of McGurk responses, the amount of unbinding, differences in response times between “Ba” and McGurk targets). No significant correlation was found in any of these tests.

References

Boersma P. & Weenink D. (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.63, retrieved 24 January 2014 from http://www.praat.org/

Charchat-Fichman, H., & Oliveira, R. M. (2009). Performance of 119 Brazilian children on Stroop paradigm-Victoria version. *Arquivos de Neuro-Psiquiatria, 67*(2b), 445-449. doi:10.1590/S0004-282X2009000300014

Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience, 6*. doi:10.3389/fnagi.2014.00347

Gatehouse, S., & Noble, W. (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). *International Journal of Audiology, 43*(2), 85-99. doi:10.1080/14992020400050014

Homack, S., & Riccio, C. A. (2004). A meta-analysis of the sensitivity and specificity of the Stroop Color and Word Test with children. *Archives of Clinical Neuropsychology, 19*(6), 725-743. doi:10.1016/j.acn.2003.09.003

Hutchison, K. A., Balota, D. A., & Duchek, J. M. (2010). The Utility of Stroop Task Switching as a Marker for Early Stage Alzheimer's Disease. *Psychology and Aging, 25*(3), 545-559. doi:10.1037/a0018498

Lallouache, M. T. (1990). *Un poste 'visage-parole.' Acquisition et traitement de contours labiaux (A 'face-speech' workstation. Acquisition and processing of labial contours).* Paper presented at the Proceedings XVIII Journées d'Etudes sur la Parole Montréal.

Lamers, M. M., Roelofs, A., & Rabeling-Keus, I. (2010). Selective attention and response set in the Stroop task. *Memory and Cognition, 38*(7), 893-904. doi:10.3758/MC.38.7.893

Moulin, A., Pauzie, A., & Richard, C. (2015). Validation of a French translation of the speech, spatial, and qualities of hearing scale (SSQ) and comparison with other language versions. *International Journal of Audiology, 54*(12), 889-898. doi:10.3109/14992027.2015.1054040

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America, 132*(2), 1061-1077. doi:10.1121/1.4728187

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2015). Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. *Journal of the Acoustical Society of America, 137*(1), 362-377. doi:10.1121/1.4904536

Pocklington, B., & Maybery, M. (2006). Proportional Slowing or Disinhibition in ADHD? A Brinley Plot Meta- analysis of Stroop Color and Word Test Performance. *International Journal of Disability, Development and Education, 53*(1), 67-91. doi:10.1080/10349120500510057

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 461-479. doi:10.1037/0096-1523.22.2.461

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643-662. doi:10.1037/h0054651

van Mourik, R., Oosterlaan, J., & Sergeant, J. A. (2005). The Stroop revisited: a meta-analysis of interference control in AD/HD. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 46*(2), 150-165. doi:10.1111/j.1469-7610.2004.00345.x

1154

1155