Binding and unbinding the McGurk effect in audiovisual speech fusion: Follow-up experiments on a new paradigm

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz

Binding and unbinding the McGurk effect in audiovisual speech fusion: Follow-up experiments on a new paradigm

Olha Nahorna, Frédéric Berthommier, Jean-Luc Schwartz

GIPSA-Lab – DPC, ICP
UMR 5216 –CNRS Uiversité de Grenoble
Olha.Nahorna, Jean-Luc.Schwartz, Frederic.Berthommier@gipsa-lab.grenoble-inp.fr
http://www.gipsa-lab.grenoble-inp.fr

ABSTRACT

The McGurk effect demonstrates the existence of a fusion process in audiovisual speech perception: the combination of the sound "ba" with the face of a speaker who pronounces "ga" is frequently perceived as "da". We assume that in the upstream of this phonetic fusion process, there is a “binding” process, which controls the combination of image and sound, and can block or reduce it in the case of audiovisual incoherencies (conditional binding process), as in the case of a dubbed film. To test and explore this binding hypothesis, we designed various experiments in which a coherent or incoherent audiovisual context is placed before McGurk stimuli, and we show that the incoherent contextual stimulus can significantly reduce the McGurk effect.

Keywords: McGurk effect, binding, multisensory fusion, audiovisual speech perception, audiovisual scene analysis.

1. INTRODUCTION

In the McGurk effect, fusion between the sound and the image has long been considered as automatic (Massaro, 1987) (Soto-Faraco, Navarra, & Alsius, 2004). This is now questioned in recent experiments showing that imposing high demands on the attention system decreases the amount of audiovisual fusion (Alsius, Navarra, Campbell, & Soto-Faraco, 2005) (Alsius, Navarra, & Soto-Faraco, 2007 .

While evidence for the non-automaticity of the fusion mechanism stays compatible with a one-stage architecture, some data suggest that audiovisual interactions could intervene at various stages in the speech decoding process (Grant & Seitz, 2000) (Kim & Davis, 2004) (van Wassenhove, Grant, & Poeppel, 2005) (Bernstein, Auer, & Moore, 2004) (Bernstein, Auer, Wagner, & Ponton, 2008) (Bernstein, Takayanagi, & Auer, 2004). Actually, audiovisual fusion could be conceived as a two-stage process, beginning by *binding* together the appropriate pieces of audio and video information, followed by integration per se (Berthommier, 2004). The binding stage would occur early in the audiovisual speech processing chain enabling the listener to extract and group together the adequate cues in the auditory and visual streams, exploiting coherence in the dynamics of the sound and sight of the speech input.

2. PRELIMINARY EXPERIMENTS

In a preliminary experiment presented in AVSP 2010 (Nahorna, Berthommier, & Schwartz, 2010), we proposed an original paradigm to test this idea and attempt to show the existence of a binding process able to modulate audiovisual fusion. In this paradigm, incongruent “McGurk” targets (A/ba/ + V/ga/) or congruent “ba” (A/ba/ + V/ba/) targets are preceded by coherent or incoherent audiovisual contexts (Fig. 1).

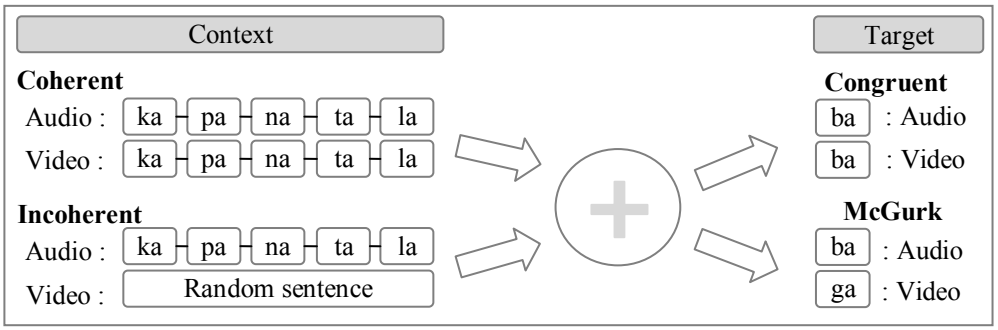

Figure 1: Experimental paradigm

Two contextual audiovisual stimuli (either coherent or not) precede two target audiovisual stimuli (a congruent audiovisual “ ba ” or a McGurk stimulus combining an audio “ ba ” with a visual “ ga ”). The coherent context consists of a sequence of 5, 10, 15 or 20 syllables randomly selected within {“ pa ”, “ ta ”, “ va ”, “ fa ”, “ za ”, “ sa ”, “ ka ”, “ ra ”, “ la ”, “ ja ”, “ cha ”, “ ma ”, “ na ”}. In the incoherent context, the auditory content is the same, but the visual content is replaced by a series of sentences matched in global duration.

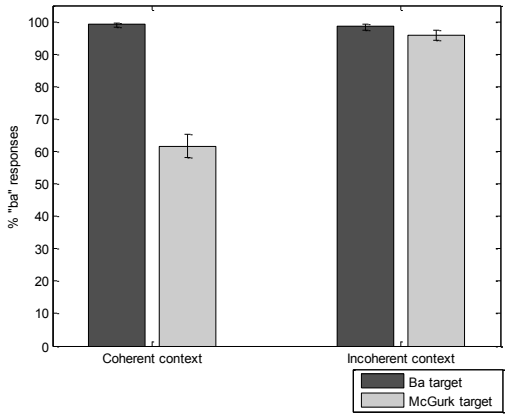

Figure 2: Results of Experiment 1
Percentage of " ba " responses for " ba " (in dark grey) and " McGurk " (in light grey) stimuli, in the coherent (left) vs. incoherent (right) contexts.

The results showed the almost complete elimination of the McGurk effect in the case of incoherent contexts (Fig. 2).

A second experiment aimed to check if a short 200-ms audiovisual alert introduced in the coherent or incoherent contexts just before (280 ms) "McGurk" targets could result in focusing the attention of the subject and hence remove the effect of the incoherent context and reset the McGurk effect at its initial stage. It appeared that the alert had actually no effect at all: the effect of the incoherent context remained exactly the same whether the alert was present or not (Fig. 3).

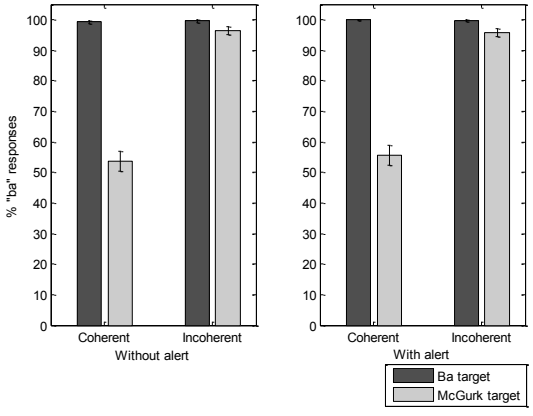

Figure 3: Results of Experiment 2
(a) For each box, percentage of " ba " responses for " ba " (in dark grey) and " McGurk " (in light grey) stimuli, in the coherent (left) vs. incoherent (right) contexts. Left box: without alert stimulus. Right box: with alert stimulus.

3. CONTROL EXPERIMENTS AND VALIDATION OF THE EFFECT

In the preparation of Experiments 1 and 2 we controlled congruent "ba" and incongruent "McGurk" stimuli in terms of labial dynamics and acoustical intensity, but we had not checked by perception tests if there could exist fine differences in stimuli whether they were extracted from the coherent or incoherent material. Particularly, the visual "ga" stimulus incorporated in the "McGurk" targets was extracted, in the incoherent context case, from sentence material, and in the coherent context case, from sequences of syllables. Therefore we designed a third experiment assessing the perceptual content of pure targets without context. The congruent "ba" and incongruent "McGurk" targets used in experiment 1 and 2, trimmed 80ms before burst onset, were presented to the subjects in random order. It appeared that though "McGurk" targets produced a strong McGurk effect whatever the stimuli (either coming from the coherent, or from the incoherent context), there were small but significant differences between the targets recorded after coherent or incoherent contexts. Indeed, the McGurk effect was slightly less with targets coming from the incoherent context material, which could have influenced the validity of our conclusions after Experiments 1 and 2.

Therefore we prepared a fourth experiment, where we used only targets recorded with coherent contexts. For this aim, we exploited the finding of Experiment 2 showing that the context effect seemed to resist an alert pointing to the target location in time. In Experiment 4, we used exactly the same contextual stimuli as in Experiment 1, but targets consisted of "ba" and "McGurk" stimuli extracted only from the coherent context (that is, corresponding in the visual stream to sequences of syllables, as in the audio stream). For this purpose, a fixed set of target stimuli (comprising "ba" and "McGurk" stimuli) was cut and placed at the end of the coherent and incoherent context sequences of Experiment 1. Since there was no more continuity between the end of the context stimulus and the onset of the target stimulus, we introduced a 200-ms transition stimulus between context and target, considering that this transition, possibly providing a visible "alert", would probably not remove the potential unbinding effect with incoherent contexts. The same set of target stimuli was used for both contexts and for all context durations.

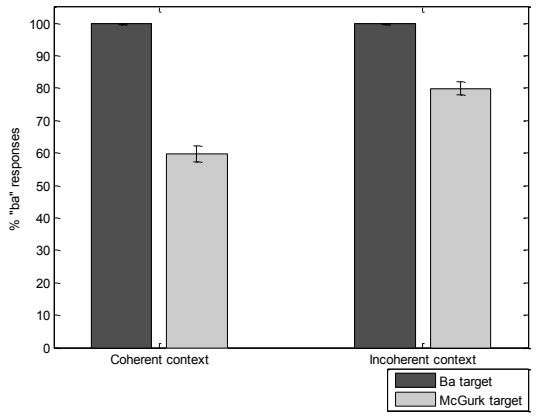

Figure 4: Results of Experiment 4
Percentage of " ba " responses for " ba " (in dark grey) and " McGurk " (in light grey) stimuli, in the coherent (left) vs. incoherent (right) contexts.

The results confirmed that an incoherent context reduces the McGurk effect (Fig. 4), though reduction was not complete in this case, either because of a better control of the targets, or because of the mounting procedure, which could, in spite of the results of Experiment 2, have slightly decreased the "unbinding" effect produced by the incoherent context.

4. FURTHER EXPERIMENTS ON THE ROLE OF CONTEXT

Now it seems clear that an incoherent context decreases the amplitude of the McGurk effect. In a last experiment, we attempted to better understand what kind of incoherence was necessary to produce this reduction in the McGurk effect. For this aim, we tested smaller kinds of incoherencies, in which the audio and visual contents of incoherent contexts were both made of syllables (while the visual content of the incoherent context was made of sentences in Experiments 1, 2 and 3). We kept intact the visual track, and we applied various modifications on the audio track in the context material.

In a first manipulation (phonetically inconsistent context, P), we permutated the audio content from one syllable to the other. To maximize the chance that the audio-visual incoherence would indeed be perceivable for each context syllable, syllables were firstly organised in five groups known to be visually rather distinguishable (visemes), then the audio content of each syllable was permutated with the content of a syllable from a different group.

In a second manipulation (temporally incoherent context, T), we slightly advanced or delayed each audio syllable at random from 30ms audio lead to 170ms audio lag. This was aimed at staying within an "integration window" (Van Wassenhove, Grant, & Poeppel, 2007) in which the McGurk effect has been shown to hardly vary. In the last incoherent context (PT), both phonetic and temporal manipulations were applied, in exactly the same way respectively as the two previous contexts.

The results are displayed in Fig. 5. They show a reduction in the McGurk effect in the P context and to a lesser extent in the T context, and an even larger reduction in the PT context. A three-factor ANOVA was computed with the factors "subject", "target" and "context". The three factors produce highly significant effects ($p<0.0001$). Focussing on the stimuli of interest that are the McGurk targets (the congruent ones being only a control), a second three-factor ANOVA was computed with the factors "subject", "phonetic incoherence" and "temporal incoherence", considering that the four contexts (coherent, P, T and PT) could be decomposed into these two factors. The results show that the three factors are significant, and particularly the phonetic incoherence ($p<0.0001$) and the temporal incoherence ($p<0.02$), with no interaction between the two factors ($p>0.8$).

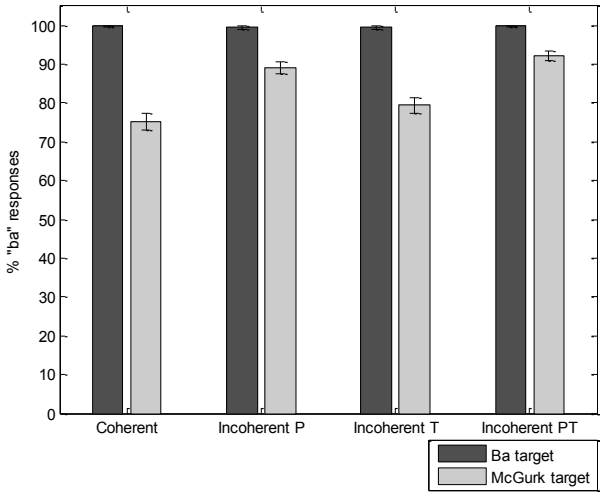

Figure 5: Results of Experiment 5
Percentage of "ba" responses for "ba" (in dark grey) and "McGurk" (in light grey) stimuli, in the coherent (C) vs. phonetically incoherent (P), temporally incoherent (T) and phonetically and temporally incoherent (PT) contexts.

5. DISCUSSION

All these experiments converge to show that McGurk fusion depends on the previous audiovisual context. This suggests that the coherence vs. incoherence of the audio and video streams could lead the subject to selectively increase vs. decrease the role of the visual input in the fusion process.

The existence of a two-stage process has long been introduced in auditory perception through "Auditory Scene Analysis", with a first binding stage grouping together the auditory components of a given acoustic source, before categorisation processes could be applied on this source (Bregman, 1990). The present paper extends this idea towards "Audiovisual Speech Scene Analysis". It is classically considered that the Auditory Scene Analysis process involves a default grouping stage followed by a possible build-up of auditory segregation. The present data are consistent with the hypothesis of a default state of the binding mechanism in which audio and video components are fused together (leading to the McGurk effect), followed by an "unbinding" process when evidence for different auditory and visual sources accumulates.

Let us come back to the models of audiovisual fusion available in the literature. One-stage models consider that phonetic decision operates at a given representational stage, and produces an integrated percept combining auditory and visual cues in a given way, possibly mediated by general attentional mechanisms (Fig. 6a). The present data suggest that an additional computational stage should be incorporated before decision operates (Berthommier, 2004). This involves online computation of some assessment of the coherence/incoherence of the auditory and visual inputs (C in Fig. 6b). Local coherence may also help the subject to better process the auditory

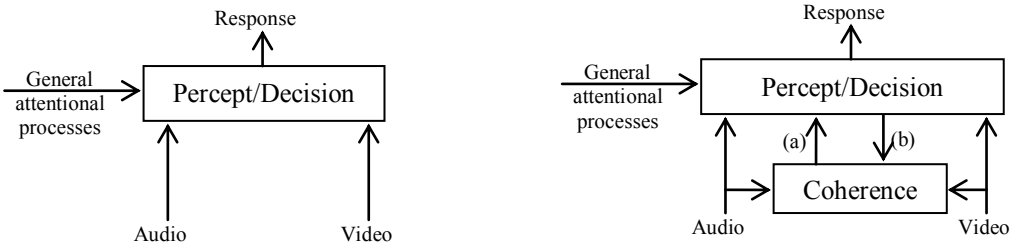

Figure 6: One-stage vs. two-stage model for audiovisual fusion in speech perception
(a) Left: A possible one-stage model
(b) Right: A possible two-stage model

and visual streams and extract adequate information (Schwartz, Berthommier, & Savariaux, 2004). Though instantaneous evidence for incoherence does not suffice to unbind the auditory and visual inputs, as displayed by the McGurk effect, accumulation of such evidence may modulate the decision process. This is displayed in Fig. 6b by a bottom-up arrow (a). The effect of phonetic incoherence, displayed in Experiment 2, suggests that the decision stage itself could intervene in the computation of the coherence measure: this motivates the top-down arrow (b). The coherence evaluation C could result in decreasing the weight of the visual stream in the decision output if it suggests that the audio and video streams are incoherent, as happens in the various cases of incoherent contexts explored in this paper.

6. ACKNOWLEDGEMENTS

This work was supported by the French National Research Agency (ANR) through funding of the MULTISTAP project (MULTISTability and binding in Audition and sPeech: ANR-08-BLAN-0167 MULTISTAP)

7. BIBLIOGRAPHY

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

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

Bernstein, L. E., Takayanagi, S., & Auer, E. T. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Comm. 44* , 5-18.

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

Bernstein, L., Auer, E., Wagner, M., & Ponton, C. (2008). Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage 39* , 423–435.

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

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

Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am. 108* , 1197–1208.

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

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

Nahorna, O., Berthommier, F., & Schwartz, J. (2010). Binding and unbinding in audiovisual speech fusion: Removing the McGurk effect by an incoherent preceding audiovisual context. *AVSP2010 - International Conference on Auditory-Visual Speech Processing.* Hakone, Kanagawa, Japan Japon.

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

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

van Wassenhove, V., Grant, K., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *PNAS 102* , 1181–1186.

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