INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Audiovisual speech scene analysis in the context of competing sources

Attigodu C. Ganesh, Frédéric Berthommier & Jean-Luc Schwartz

CNRS, GIPSA-Lab, F-38000 Grenoble, France
Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

Abstract

Audiovisual fusion in speech perception is generally conceived as a process independent from scene analysis, which is supposed to occur separately in the auditory and visual domain. On the contrary, we have been proposing in the last years that scene analysis such as what takes place in the cocktail party effect was an audiovisual process. We review here a series of experiments illustrating how audiovisual speech scene analysis occurs in the context of competing sources. Indeed, we show that a short contextual audiovisual stimulus made of competing auditory and visual sources modifies the perception of a following McGurk target. We interpret this in terms of binding, unbinding and rebinding processes, and we show how these processes depend on audiovisual correlations in time, attentional processes and differences between junior and senior participants.

Index Terms: audiovisual fusion, McGurk effect, scene analysis, attention, seniors

1. Introduction

The classical cocktail party effect and the problems it raises for speech perception in adverse conditions [1] has generated two series of theoretical and experimental developments, which remain surprisingly separate in the years. On the one hand, Auditory Scene Analysis (ASA) puts at its agenda the search for auditory mechanisms enabling to group together auditory cues into auditory primitives, based on principles such as temporal synchrony, correlations in time or space, and more globally common fate [2]. On the other hand, Audiovisual Speech Perception (AVSP) is focused on the cognitive processes enabling to fuse together the auditory and the visual input corresponding to the speaker's utterances (e.g. [3-5]), while assuming implicitly or explicitly that scene analysis occurs independently in the auditory and the visual domain.

As a matter of fact, some experimental studies display evidence where unimodal perceptual grouping precedes multisensory integration (e.g. [6-8]). However, several recent behavioral and neurophysiological studies have suggested that the presentation of a visual stream can affect primary auditory streaming by enhancing segregation or integration ([9-13]). This suggests that audiovisual speech perception likely incorporates a stage of audiovisual scene analysis before fusion might operate [14-15].

This led our group propose since a number of years that the ASA and AVSP frameworks should be combined within a single framework, that we tentatively called "audiovisual speech scene analysis" (AVSSA). In this framework, it is assumed that audiovisual speech perception is a two-stage process (as is auditory perception in the ASA framework), beginning by a scene analysis process where auditory and visual cues are bound within "audiovisual primitives", before audiovisual fusion for decision at a later stage. Fusion would hence operate on a set of audiovisual cues – an audiovisual source – selected at the first stage, and the result of the fusion process would depend on the coherence of the auditory and visual inputs at the first stage [16].

The two-stage model of audiovisual speech perception in the framework of AVSSA has been elaborated in the last years, taking advantage of a specific paradigm that we developed to demonstrate that audiovisual fusion does indeed depend on the coherence of the unisensory inputs. This paradigm is based on the classical "McGurk effect", in which a visual input incongruent with an auditory input may change the perception of the sound: typically a visual "ga" dubbed on an auditory "ba" leads to the perception of "da" or "tha" [17]. In two series of experiments [18-19], we were able to show that if such a McGurk stimulus is preceded by a period of audiovisual material displaying incoherent variations, the participants are lead to consider that the auditory and visual inputs should not be bound together, hence the McGurk effect decreases. This is what we called "unbinding", which can be possibly followed by a period of audiovisual coherence leading to "rebinding".

The binding/unbinding/rebinding processes should in our view play a crucial role in audiovisual speech perception in adverse but ecological conditions where various speakers discuss altogether: typically the cocktail party situation, where the listener should be able to attend to a source and adequately select the corresponding pieces of audio and video information before being able to fuse and understand the message. This is why we recently attempted to evaluate the binding/unbinding/rebinding paradigm with mixtures of sources, to assess whether the AVSSA framework would be able to account for such kinds of configurations. In this paper we will report the major results of this series of experiments about audiovisual speech scene analysis in the context of competing sources.

In Section 2 we present the binding/unbinding/rebinding paradigm together with the previously obtained results [18-19]. Then, Section 3 reports the results of two series of experiments dealing with AVSSA with a mixture of audiovisual sources. Section 4 will be devoted to results obtained on a population of older subjects, showing how the binding abilities vary with age.

http://dx.doi.org/10.21437/Interspeech.2016-62

2. Binding, unbinding, rebinding

The experimental paradigm developed previously [18-19] and which will be adapted in some of the present experiments is displayed in Fig. 1. The principle is the following.

A pure McGurk "target" made of an auditory "ba" and a visual "ga" is not 100% perceived as "ba" because of the McGurk effect leading a number of participants to perceive it as "da": this is "binding" (Fig. 1, left). The assumption is that listeners have a "default binding stage", in which they consider that in lack of any further evidence, the auditory and visual inputs should be bound together for audiovisual fusion.

If the target is preceded by a given duration of incoherent audiovisual material (context), the amount of fusion decreases, hence the score of "ba" responses increases: this is unbinding (Fig. 1, middle).

If a "reset" stimulus made of audiovisual coherent material is presented after the incoherent context and before the McGurk target, subjects recover the original McGurk effect, hence the percentage of "ba" responses decreases down to its level for a pure McGurk target: this is rebinding (Fig 1, right).

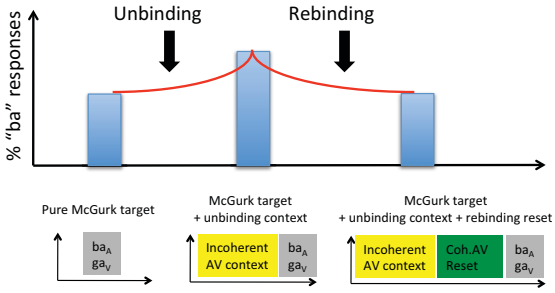

Figure 1: *The unbinding/rebinding paradigm*

The material presented in all the previous experiments [18-19] and in the present paper were prepared from audiovisual material produced by a French male speaker, JLS, with lips painted in blue to allow precise video analysis of lip movements [20]. The videos consisted of the entire speaker's face, keeping natural colors apart from the blue make-up. Recordings were digitized at an acoustic sampling frequency of 44.1 kHz and a video sampling frequency of 50 Hz (25 images per second with two frames per image).

The target was either a congruent audiovisual "ba" syllable, or an incongruent McGurk stimulus with an audio "ba" dubbed on a video "ga" with precise temporal synchronization at the plosive burst. The focus was actually on McGurk targets and the congruent "ba" targets were only presented as controls.

The context and reset stimuli in all the experiments were prepared from various kinds of mixtures (that will be described later) of two types of audiovisual material. The first type, called "syllables", was made of random sequences of audiovisual syllables in the set: "pa", "ta", "va", "fa", "za", "sa", "ka", "ra", "la", "ja", "cha", "ma" or "na". The syllable rhythm was about 1.5 Hz, hence for sequences of 2 or 4 syllables that were used in the experiments presented here, the context duration varied between 1.3 and 2.7 s depending on the number of uttered syllables. The second type, called "sentences", consisted of excerpts from sentences invented online by the speaker at the recording stage.

3. Unbinding and rebinding with competing sources

The princeps experiments [18-19] involved either coherent audiovisual contexts made of the coherent auditory and visual content of syllable sequences (A and V syllables in the following), or incoherent contexts in which A *syllables* were mixed with V *sentences* with the adequate duration.

Of course, the incoherent context material actually corresponded to two differing sources, one syllabic source presented in the auditory modality and one sentence source presented in the visual modality. However, there was no competition of sources in individual modalities.

The objective in present experiments was to go towards more realistic situations involving a competition of sources inside the auditory modality – apart from possible incoherence between modalities.

3.1. Adding noise in the audiovisual context (Exp. 1)

In a first step (Exp. 1), we considered acoustic noise added to the acoustic source in the context part. Therefore, we exploited the "unbinding/rebinding" paradigm presented in Fig. 1, though with an important variant, that is the addition of acoustic noise in the context before the McGurk target [21].

The experiment involved two types of context before the two targets (see Fig. 2). Firstly, a coherent context made of coherent A and V syllables (Fig. 2, bottom), provided a baseline. Secondly, an incoherent context made of 2 or 4 A syllables dubbed on excerpts from V sentences with the adequate duration was followed by a coherent reset made of a variable number (0, 1, 2 or 3) of coherent AV syllables (Fig. 2, top). This should produce unbinding (by the incoherent context) and a given amount of rebinding (depending on the number of syllables in the reset). The (context + reset) portion was presented either in clear or in acoustic noise (Gaussian white noise at 0 dB SNR), with no noise on the target.

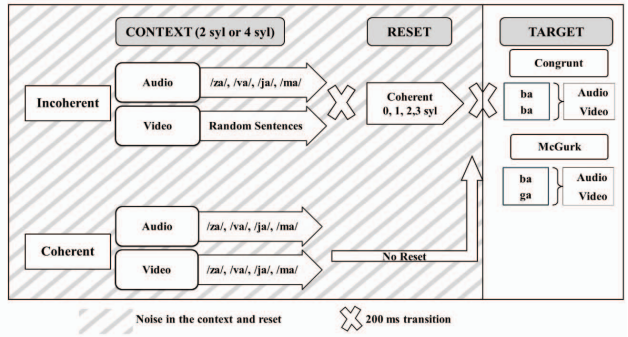

Figure 2: *Material for Experiment 1*

Thirty-one participants (22 women and 9 men; 30 right-handed and 1 left-handed; mean age=31.7 years; SD=11.7 years) took part in this experiment. The task consisted in monitoring online for the perception of either "ba" or "da" syllables – reminding the reader that there were in fact no "da" syllables all along the experiment. The participants had to signal the perception of either "ba" or "da" by pressing as soon as possible an adequate response button. Responses were only considered within a given amount of time after the target onset defined as the burst onset (between 200 ms and 1200 ms, see [18] for more explanations on the experimental procedure).

The results are displayed on Fig. 3 (averaged over the two context durations, 2 and 4 syllables). Fig. 3a displays the results obtained with no noise in the context period. They basically replicate the results in the original study [19], displayed in a schematic way on Fig. 1: while the amount of McGurk responses is large with a coherent context (context "1" in the figure, with less than 30% "ba" responses, hence more than 70% McGurk responses), this amount decreases with an incoherent context (context "2", with more than 50% "ba" responses) and comes back to its original value with a coherent reset made of 3 coherent audiovisual syllables (context "3"). This is the unbinding-rebinding phenomenon described earlier.

Fig. 3b displays how results change when there is noise on the context – though NOT in the target. The amount of McGurk effect largely increases for all contexts – hence the percentage of "ba" responses decreases by 13% to 30% depending on context. There remains unbinding and rebinding, since the scores still vary from the coherent to the incoherent context and back to the incoherent + coherent reset context. However, the fluctuations are largely decreased in noise.

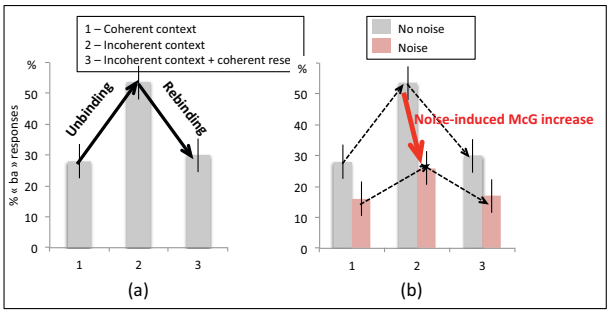

Figure 3: *Results for Experiment 1*

The interpretation we suggest is that noise in the context period is used by the subjects as an indication that the auditory sensor is less reliable and should hence be considered with a smaller weight in the fusion process. This is in line with various proposals about a "weighted fusion process", according to which fusion would depend globally on the weight each subject would attribute to the auditory and visual channels, possibly varying with subject, language, noise and a number of other factors (e.g. [22-24]).

3.2. Adding a competing source in the audiovisual context (Exp. 2)

In a second step, we aimed to further explore the possibility that a scene analysis process would take place in the course of AV fusion. For this aim, we presented contexts made of a mixture of sources. More precisely, we prepared mixtures of two audio sources associated with one video input, being coherent with the one of the two audio sources [25].

We predicted that the binding stage in AVSSA would now serve two roles: 1) compute partial correlations, which could enable the system to select the audio source coherent with the video input, and 2) assess the binding state modulating AV fusion. To test these predictions, we mixed two audio sources which have very different properties over time, and which are hence likely to lead to very different correlations with their corresponding video counterpart: a syllable stream and a sentence stream (see Fig. 4). Indeed, syllables correspond to stronger AV modulations in time and hence stronger AV coherence than sentences. Therefore, the association between the visual input and the corresponding auditory input should be stronger for syllables than for sentences. Hence, we predicted that the coherence of the AV context would be stronger for visual syllables, and would lead to a larger visual weight and more McGurk effect than with visual sentences.

Experiment 2 actually consisted of two parts. Firstly (Exp. 2a) the subjects performed the same task as before – that is monitored for the presence of "ba" or "da" targets" – with one or the other context, to assess whether the "video syllables" context would produce more "da" responses than the "video sentences" context. Secondly (Exp. 2b), the same subjects performed the same task though with an additional demand: to attend selectively either to the syllables or to the sentences (with a counterbalanced order of the "attention to syllables" and "attention to sentences" conditions). The experiment was passed by twenty-nine French participants without hearing or vision problems (22 women and 7 men; 27 right-handed and 2 left-handed; mean age= 29.2 years; SD=10. 4 years).

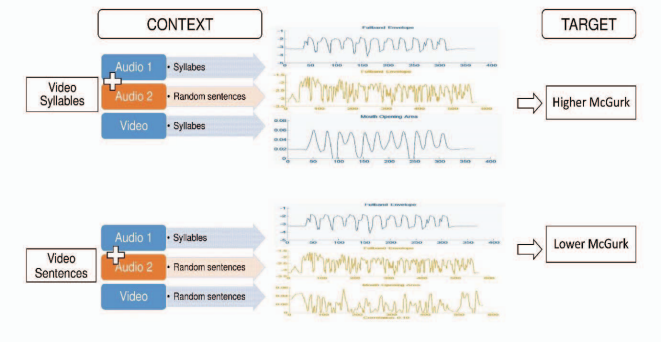

Figure 4: *Material for Experiment 2*

In the first part with no specific attentional demand (Exp. 2a), the results were actually in line with our prediction, with a 10% higher amount of McGurk responses (10% lower amount of "ba" scores) with video syllables. This is due in our view to a larger audiovisual correlation with syllables, hence more binding and more McGurk responses.

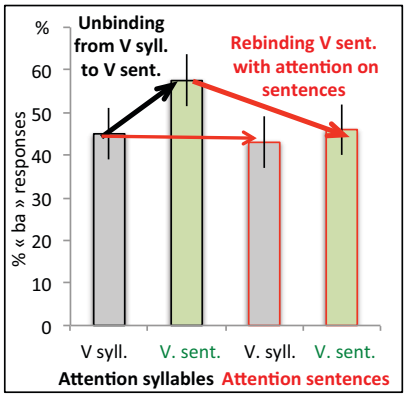

Figure 5: *Results for Experiment 2b*

The results obtained in the second part with an attentional focus on syllables or sentences (Exp. 2b) are displayed on Fig. 5. For the attentional focus on syllables ("Attention syllables") we replicate the results with no attention: there is unbinding from the video syllables (V. syll.) to the video sentences (V. sent.) context, hence more "ba" responses in the second case. The fact that attention is put on syllables does not modify the

results with no attentional focus, possibly because video syllables are so correlated that they "pop out" automatically as coherent in the binding process. However, attention put on sentences ("Attention sentences") seems to "rebind" the auditory and visual streams for the visual sentences context (V. sent.), hence removing the difference between the two contexts "V. syll." and "V. sent." This could be due to top-down processes based on audiovisual schemas, according to which the intrinsically low binding associated with audiovisual sentences would be enhanced thanks to the attentional process.

4. Assessing binding in seniors

Our last set of experiments consisted in assessing whether binding could vary with age. A series of recent papers have attempted to establish whether audiovisual integration was different in senior (typically above 65 years of age) contrasted to junior adults. The conclusions are not fully convergent. Altogether, while seniors display a degradation in their auditory perception but less so in their visual perception of speech, their integration abilities seem to stay stable, with actually a trend for a higher role of the visual input in speech perception in seniors (see a review in [26], and recent evidence in [27] that the visual influence is greater in older adults compared with younger ones not only with equal SNRs but also with SNRs calibrated to equalize unisensory performance).

The question we decided to explore in this final study was hence if binding would operate in a different way in seniors compared to juniors [28]. Indeed, since the visual modality is of particular importance for seniors, it could be envisioned that the visual input would be exploited and fused even in case of incoherence, hence a decrease in unbinding in older adults. However, since the audiovisual scenes used in our experimental paradigm are rather complex, it could be suggested quite in the contrary that unbinding would be larger with a larger modulation by context and attention.

To test these two inverse predictions, we exploited two of the previous experimental paradigms in this study that are Experiment 1 without noise (the primary "unbinding-rebinding" paradigm without noise developed in [19]) and Experiment 2b (unbinding with context including two competing audio streams, and with attentional focus either on syllables or on sentences).

Twenty-five native French speaking older adults participated in the experiments (2 women and 23 men; from 60 to 75 years, 21 right-handed and 4 left-handed, mean age= 65.3 years; SD=3.9 years). None of them reported any hearing, vision (after correction) or neurological disorders.

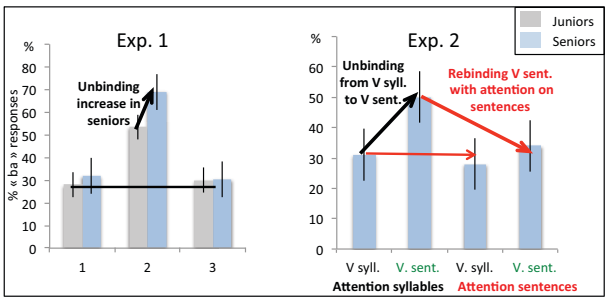

Figure 6: *Results for Experiments 1 and 2 on seniors*

The results are displayed on Fig. 6. They compare the results obtained with our senior group with those obtained with the younger group reported previously. They display two very interesting results. Firstly, in Experiment 1, while the average McGurk score was typically the same in our two groups of subjects with a coherent context (no significant difference in the percentage of "ba" responses), there was a strong and significant increase in unbinding in seniors (more than 15% increase in "ba" responses from juniors to seniors with the incoherent context). Rebinding resulted in recovering the same amount of fusion in the two populations. Secondly, in Experiment 2b, the results were qualitatively similar to those obtained with juniors, with effects of both context and attention as described in Section 3.2, without obvious differences in the size of the effects.

The lessons of this set of experiments on seniors are rather clear. Firstly, globally, they provide a replication of the results with juniors, confirming (i) that context matters in fusion; (ii) that unbinding-rebinding processes modulate the McGurk process; (iii) and that within competing sources syllables produce more binding but attention may rebind in the case of audiovisual sentences.

Secondly, they suggest that seniors unbind more than juniors. The larger effect of unbinding in older subjects could be related to the fact that under cognitive load, integration reduces (see [29, 30]). Indeed, it could be assumed that in the case of incoherence, a certain amount of attention is required for keeping audition and vision bound together and hence produce binding. If the ability to maintain this amount of attention is decreased in seniors, this would result in less fusion and more unbinding, which is actually what happens here.

5. Conclusions

All the data presented in the present paper confirm that audiovisual context may modulate audiovisual fusion displayed by the McGurk effect. They are all in good agreement with the two-stage "binding-and-fusion" model that we are developing in our team, in the framework of Audio-Visual Speech Scene Analysis (AVSSA). They converge towards a model in which the binding stage has actually various roles, such as audiovisual estimation of the reliability of the sources (Exp. 1), and selection of the adequate unisensory sources or pieces of sources for construction of the audiovisual source to process (Exp. 2; see also [11]). Interestingly, senior data also suggest that binding could be a computationally costly process, which may suffer some limitations in case of increased cognitive load (see also [29, 30]). The binding paradigm exploited in the present study once again appears as a very efficient tool for studying the cognitive processes likely to take place in Audio-Visual Speech Scene Analysis.

6. Acknowledgements

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. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s", J.-L. Schwartz PI).

7. References

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *J Acoust Soc Am*, vol. 25, pp. 975-979, 1953.

[2] A. S. Bregman, *Auditory scene analysis*. MIT Press, Cambridge, MA, 1990.

[3] D. W. Massaro, *Speech Perception by Ear and Eye*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[4] Q. Summerfield, "Some preliminaries to a comprehensive account of audiovisual speech perception," in *Hearing by eye: The psychology of lipreading*, edited by B. Dodd, and R. Campbell (NJ: Lawrence Erlbaum Associates, Hillsdale), pp. 3-51, 1987.

[5] J.-L. Schwartz, J. Robert-Ribes, and P. Escudier, P., "Ten years after Summerfield. A taxonomy of models for audiovisual fusion in speech perception," in *Hearing by Eye II. Perspectives and Directions in Research on Audiovisual Aspects of Language Processing*, edited by R. Campbell, B. Dodd, and D. Burnham (Psychology Press, Hove), pp. 85–108, 1998.

[6] D. Sanabria, S. Soto-Faraco, J. Chan, and C. Spence, "Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task," *Neurosci Lett*, vol. 377, pp. 59-64, 2005.

[7] M. Keetels, J. Stekelenburg, and Vroomen, J., "Auditory grouping occurs prior to intersensory pairing: evidence from temporal ventriloquism," *Exp Brain Res*, vol. 180, pp. 449-456, 2007.

[8] K. G. Munhall, M. W. ten Hove, M. Brammer, and M. Paré, "Audiovisual Integration of Speech in a Bistable Illusion," *Curr Biol*, vol. 19, pp. 735-739, 2009.

[9] T. Rahne, M. Bockmann, H. von Specht, and E. S. Sussman, "Visual cues can modulate integration and segregation of objects in auditory scene analysis," *Brain Res*, vol. 1144, pp. 127-135, 2007.

[10] J. Marozeau, H. Innes-Brown, D. B. Grayden, A. N. Burkitt, and P. J. Blamey, "The Effect of Visual Cues on Auditory Stream Segregation in Musicians and Non-Musicians," *PLoS One*, vol. 5, e11297, 2010.

[11] F. Berthommier, and J.-L. Schwartz, "Audiovisual streaming in voicing perception: new evidence for a low-level interaction between audio and visual modalities", 10th International Conference on Auditory-Visual Speech Processing (AVSP 2011), pp. 77-80, 2011.

[12] A. Devergie, N. Grimault, E. Gaudrain, E. W. Healy, and F. Berthommier, "The effect of lip-reading on primary stream segregation," *J Acoust Soc Am*, vol. 130, pp. 283-291, 2011.

[13] R. K. Maddox, H. Atilgan, J. K. Bizley, and A. K. Lee, "Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners," *eLife* 2015, vol. 4, e04995, 2015.

[14] T. S. Andersen, K. Tiippana, J. Laarni, I. Kojo, and M. Sams, "The role of visual spatial attention in audiovisual speech perception," *Speech Commun*, vol. 51, pp. 184-193, 2009.

[15] A. Alsius, and S. Soto-Faraco, "Searching for audiovisual correspondence in multiple speaker scenarios," Exp Brain Res, vol. 213, pp. 175-183, 2011.

[16] F. Berthommier "A phonetically neutral model of the low-level audio-visual interaction," *Speech Commun*, vol. 44, pp. 31-41, 2004.

[17] H. McGurk, and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[18] O. Nahorna, F. Berthommier, F., and J.-L. Schwartz, "Binding and unbinding the auditory and visual streams in the McGurk effect," *J Acoust Soc Am*, vol. 132, pp. 1061-1077, 2012.

[19] O. Nahorna, F. Berthommier, F., and J.-L. Schwartz, "Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect," *J Acoust Soc Am*, vol. 137, pp. 362-377, 2015.

[20] M. T. Lallouache, "Un poste 'visage-parole.' Acquisition et traitement de contours labiaux (A 'face-speech' workstation. Acquisition and processing of labial contours)," in *Proceedings XVIII Journées d'Etudes sur la Parole* (Montréal), pp. 282–286, 1990.

[21] A. C. Ganesh, F. Berthommier, O. Nahorna, and J.L. Schwartz, "Effect of context, re- binding and noise, on audiovisual speech fusion," in *Proc. 14th Annual Conference of the International Speech Communication Association* (Interspeech 2013), Lyon, France, 2013.

[22] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process*. 2002, pp. 1260-1273, 2002.

[23] J.-L. Schwartz, "A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent," *J Acoust Soc Am* , vol. 127, pp. 1584-1594, 2010.

[24] A. Huyse, F. Berthommier, and J. Leybaert, "Degradation of labial information modifies audiovisual speech perception in cochlear-implanted children," *Ear Hear*, vol. 34, pp. 110-121, 2013.

[25] A. C. Ganesh, F. Berthommier, and J.L. Schwartz, "Audio Visual integration with competing sources in the framework of Audio Visual Speech Scene Analysis", in *International Symposium in Hearing*, E. Gaudrain (Ed.), to appear.

[26] J. L. Mozolic, C. E. Hugenschmidt, A. M. Peiffer, and P. J. Laurienti, "Multisensory Integration and Aging," in *The Neural Bases of Multisensory Processes*, edited by M. M. Murray, and M. T. Wallace (CRC Press/Taylor & Francis Llc., Boca Raton (FL)), 2012.

[27] K. Sekiyama, T. Soshi, and S. Sakamoto, "Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults," *Front Psychol*, vol. 5, pp. 323, 2014.

[28] A. C. Ganesh, F. Berthommier, and J.L. Schwartz, "Audiovisual Binding in Elderly Population," in *Proceedings of the Conference on Auditory-Visual Speech Processing* (AVSP2015), Vienna, Austria, 2015.

[29] A. Alsius, J. Navarra, R. Campbell, and S. Soto-Faraco, "Audiovisual integration of speech falters under high attention demands," *Curr Biol*, vol. 15, pp. 839-843, 2005.

[30] A. Alsius, J. Navarra, and S. Soto-Faraco, "Attention to touch weakens audiovisual speech integration," *Exp Brain Res*, vol. 183, pp. 399-404, 2007.