Pairing audio speech and various visual displays: binding or not binding ?

Aymeric Devergie[1]*, Frédéric Berthommier*[2] *and Nicolas Grimault*[1]

[1]Laboratoire Neuroscience Sensorielle, Comportement et Cognition, CNRS UMR 5020, Université Lyon 1, Lyon, France
[2]Gipsa-Lab,CNRS UMR 5216, INPG, UJF, Université Stendhal, Grenoble, France
{adevergi,ngrimault}@olfac.univ-lyon1.fr, berthommier@gipsa-lab.inpg.fr

Abstract

Recent findings demonstrate that audiovisual fusion during speech perception may involve pre-phonetic processing. The aim of the current experiment is to investigate this hypothesis using a pairing task between auditory sequences of vowels and non speech visual cues. The audio sequences are composed of 6 auditory French vowels alternating in pitch (or not) in order to build 2 interleaved streams of 3 vowels each. Various elementary visual displays are mounted in synchrony with one vowel stream out of the two. Our hypothesis is that, in a forced choice pairing task, the AV synchronized vowels will be found more frequently if such a perceptual binding operates. We show that the most efficient visual feature increasing pairing performance is the movement.

Surprisingly, some features we manipulated do not provide the increase in pairing performances. The visual cue of contrast variation is not correctly paired with the synchronized auditory vowels. Moreover, the auditory segregation, based on the pitch difference between the vowels streams, has no additional effect on pairing. In addition, the modulation of the auditory envelop, synchronized with the variation of the visual cue, has also no effect. Finally, when we introduce a phonetic cue in the visual display, pairing increases in comparison with non specific visual cues. The relative contribution of perceptual binding and late phonetic fusion is discussed.

Index Terms: Audiovisual fusion, perceptual binding, multimodal phonetic processing

1 Introduction

In speech, fusion of audio and visual inputs has been widely investigated through intelligibility tasks. It has been assumed that late phonetic fusion occurs during the perception of speech [1]. Only recently, other hypothesis arose, assuming that audio and visual inputs could interact at a pre-phonetic level. Intelligibility tasks are not appropriate to test this hypothesis because it necessarily involves 'lip reading' of the stimuli.

In order to focus on the lower level of processing involved in audiovisual fusion, some studies proposed a detection paradigm. When presenting a visual cue related to the auditory speech, it enhances detection of speech in noise [2] [3]. This observation argues in favour of a fusion at a more pre-phonetic level.

Findings about the ventriloquism effect showed that we are able to pair A and V inputs, even if they are not spatially coherent. This suggests that an underlying binding mechanism based on the temporal coherence is involved. The role of the temporal coherence in speech has been investigated with asynchrony detection tasks. In speech, despite the introduction of an offset asynchrony (e.g. with audio lag) between audio and visual inputs, a multimodal event could be perceived as coherent. This corresponds to a quite large temporal window of integration of about 250ms as described in [4].

Some electrophysiological studies proposed sequences of non speech auditory events grouped in several configurations [5]. Adding an elementary visual cue has been demonstrated to affect the perception of the auditory sequence and facilitated access to particular events in the auditory stream. The facilitation would suggest that audio and visual may be bound at a more prephonetic level.

From that point, the question of audiovisual binding in speech should be addressed. Our current study aimed to focus on prephonetic level of audiovisual fusion. As in [6], the experimental visual material we built consisted in elementary display varying in contrast or in movement. Such basic visual features would prevent phonetic processing to occur. The auditory material consisted in sequences of six French vowels alternating in pitch. The contrast or movement feature of visual display varied in synchrony with the auditory vowels. We defined a 'open-state' and a 'close state' for the two visual displays. The open-state for the contrast cue was the white disk and the close-state was the black disk. Open-state for the movement display corresponded to the large visual display and the close-state to the small display. In every sequences, one group of three vowels was synchronous with the 'open state' and the other group of three vowels was synchronous with the 'close state' (see figure 1 for a representation of the different states). The task consisted in pairing the group of three vowels which was synchronous with a particular state of the visual display. It is important to notice that in our experimental design, the phonetic identification of the auditory vowels would not have helped participants to perform the pairing task because the auditory vowels were not phonetically correlated to the visual displays.

Since it has been demonstrated that speech was tolerant to asynchrony [4], this probably suggests that the underlying binding process is relatively robust. Thus, in order to weaken the strength of binding, we introduced some temporal ambiguity in the experimental material. The vowel's duration was shorter than an open-close cycle of the visual cue. One open-close cycle overlapped parts of two successive vowels. Thus, pairing of A and V inputs was difficult even if variations of the visual cue were synchronous with the center of auditory vowels. We hypothesized that only the strength of binding would help participants to pair the correct audio vowels with the visual display.

Finally, in the current study, we have also investigated the role

of phonetic cues in pairing by introducing some phonetic features in the visual displays. Since, pairing of audio and visual probably involved either perceptual and phonetic aspects, it became interesting to design a kind of continuum between pure non speech and speech visual cues.

2 Movement and contrast features

2.1 Material and Methods

Sequences of six French vowels without overlap and silence were built. Each sequence was repeated in loop. Fundamental frequency of the vowels had two possible values: 100 or 134Hz. Two auditory patterns were proposed: one with constant f0 set to 100Hz and one with an alternating f0 (between 100 and 134Hz). In the latter condition, the f0 difference lead to the clear perception of two separate streams. Figure 1 represents the different visual cues. Three shapes and two visual features were proposed. In the contrast condition, shapes varied from black to white contrast on a black background frame synchronously with 1 out of 2 vowels. In the movement condition, shapes varied from open to close position with the same temporal pattern. Vowels were generated with [Klatt algorithm (1980)]. Stimuli were played using a SIGMATEL internal sound card and Sennheiser HD 250 Linear II headphones. Output level was set to 70dB SPL with RMS-value adjustment. Video display was achieved using a Samsung SyncMaster 540N TFT 17" display with a video frame rate set at 60Hz. Figure 7 shows a schematic view of one sequence represented on the timeline for the two kinds of visual displays.

2.2 Procedure

Twenty participants aged between 18 and 30 years took part in the experiment. They had to choose the triplet of vowels synchronized with a particular state of the visual display: in the contrast condition, they had to identify the group of three vowels synchronized with the open state (white disk) and in the movement condition the group of three vowels synchronized with the close state. This was made in accordance with pilot observations revealing a better detection of the close state for the movement condition. The temporal evolution of these two visual conditions were represented on figure 7.

The test was divided into 3 sessions. Participants were seated comfortably in a double-walled sound booth. The first session was a presentation session. All combinations of visual shapes and visual conditions were presented randomly. Then, the adaptation session began. In this session, they performed the pairing task with a set of 36 different runs. Each combination of shape, visual condition and f0 difference was repeated four times. Each sequence lasted 10 seconds. At the end of each sequence, the two triplets of vowels are displayed in the lower part of the screen. Participants have to choose the triplet synchronized with the target visual display. After this adaptation session, the test session started. It consisted in 240 runs divided into 4 blocks. In each block, all combinations were repeated five times each. The whole experiment lasted 30 minutes.

2.3 Results

On Figure 2, correct pairing of the target vowel triplet synchronized with the target visual state were averaged for each visual condition and f0 difference for all participants. A repeated-measure ANOVA with factors f0 difference and visual condition grouped by visual shape was performed. Visual condition has a significant effect on performances [$F(1,18)=10.86$; $p<0.01$]. F0 difference has no effect [F(1.18)=0.001; p=0.97]. No interaction between these two factors has been found [$F(1.18)=0.057$; $p=0.81$]. Performances for each combination (Visual type and Frequency) were compared to chance level. T-tests were performed and revealed only significant difference to chance level (50% correct) for the movement cue [Movement / Same F0: $t(19)=3.26$; $p<0.01$, Movement / different F0: $t(19)=2.67$; $p=0.015$]. Correct responses were grouped by visual display type in figure 3 for movement cues and in figure 4 for contrast cues. Responses were significantly different from chance level only for the vertical and horizontal bars in the movement condition.

The first experiment revealed that the most salient visual cue allowing perception of the temporal synchrony between auditory speech and non speech visual cue was the movement (Figure 2). Since the performance was at chance level with the Contrast visual display, this cue did not support to perceive the synchrony detection between the inputs.

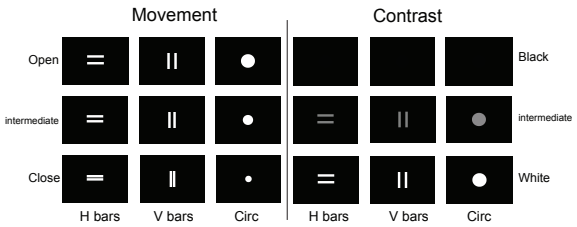

Figure 1: The different visual cues are represented in their key states (open, close and intermediate). The movement display varied from open to close. The contrast display varied from white to black. They also varied along orientation.

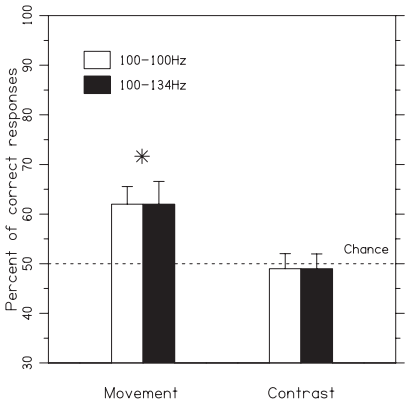

Figure 2: Percent of correct identification of the vowel triplet synchronized with white disk in the Contrast condition and close state in the Movement condition depending on Visual cue type and f0 difference between auditory vowels. Only the movement cue was significantly different from chance. Chance level was equal to 50%

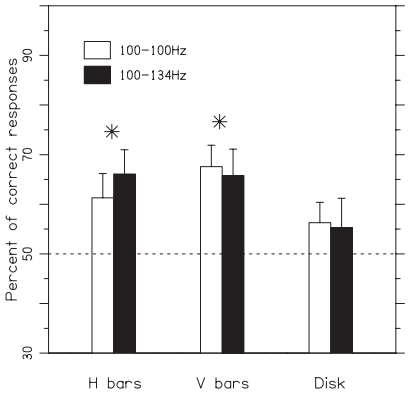

Figure 3: Percent of correct identification of the vowels triplet synchronized with the visual shape for each f0 difference grouped by visual display for the movement condition. Performance for vertical bars and horizontal bars were significantly better from chance.

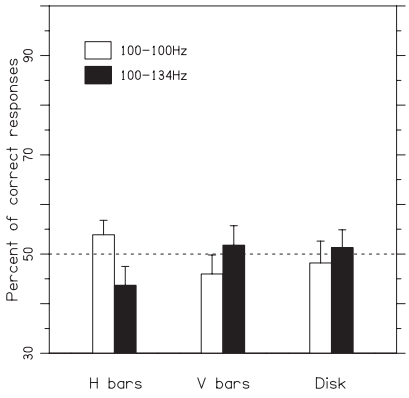

Figure 4: Percent of correct identification of the vowels triplet synchronized with the visual shape for each f0 difference grouped by visual display for the contrast condition. No visual cue elicited identification better than chance level.

3 Auditory envelop modulation

In the second experiment, we hypothesized that perception of AV synchrony would be facilitated if the modulation of the visual parameter is matched with a coherent envelop modulation of the auditory signal. The purpose of the Experiment 2 was to enhance detection of AV synchrony, particularly for the contrast condition.

3.1 Material and Methods

Ten participants took part in this experiment. Stimuli were similar to those used in Experiment 1. The number of visual shapes was reduced to two: the horizontal bars varying in movement and the disk varying in contrast. The two f0 conditions were maintained. We introduced modulation of the auditory envelop. In the 'No modulation' condition, the auditory envelop remained flat. In the 'Modulation condition', the level of each vowel was modulated with a triangular window in synchrony with the variation of the visual parameter. Global RMS-levels of the two modulation conditions were equalized. The experimental design was the same as in Experiment 1.

3.2 Results

A repeated-measure ANOVA with factors, visual condition, frequency difference and envelop modulation was performed. Figure 5 showed percent of correct pairing for all participants averaged for each visual condition, and envelop modulation for the f0 condition '100-100Hz' in the left panel and for f0 condition '100-134Hz' in the right panel. Visual condition had a significant effect on performances [$F(1,9)=12.61$; $p<0.01$]. F0 difference alone had no effect on performances [$F(1,9)=1.94$; $p=0.19$]. Theses results are consistent with the experiment 1. Results showed that envelop modulation had no effect on performances [$F(1,9)=0.11$; $p=0.74$]. Concerning the interactions between the three factors, no interaction was found between f0 difference and visual condition [$F(1,9)=0.15$; $p=0.70$] nor between f0 difference and envelop modulation [$F(1,9)=1.32$; $p=0.28$]. Right panel of the Figure 5 suggested that an interaction between visual condition and envelop modulation occured [$F(1,9)=10.31$; $p=0.011$]. T-tests were performed for each visual condition and did not revealed any significant difference between envelop modulation condition (Movement display: $t(18)=0.54$; $p=0.59$, Contrast display: $t(18)$; $p=0.10$). In addition, the performances in the 'modulation' condition for the contrast display was not significantly different from chance level [$t(9)=1.95$; $p=0.08$]

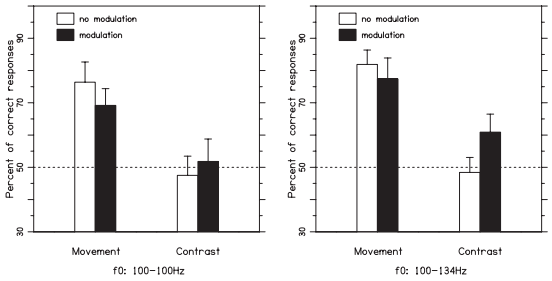

Figure 5: Percent of correct identification of the target triplet for each visual condition (x-axis) grouped by envelop modulation (white bar: no modulation, black bar: modulation). The left panel shows the performances for the f0 condition '100-100Hz'. The right panel shows the performances for the f0 condition '100-134Hz'.

4 Phonetic processing

Altogether, Experiments 1 and 2 demonstrated that the only salient cue allowing pairing between A and V cue is the movement. The experiment 3 included the following features. The auditory envelop modulation was maintained. The movement cue was further investigated by introducing the natural vertical extend of the lips. Using natural lips movement for building the visual display was also expected to introduce some phonetic features.

4.1 Material and Methods

Ten participants took part in this experiment. Design was the same as in Experiment 1 and 2. New visual conditions were introduced. Three visual cues were proposed. The first one consisted in the disk varying in size (as in Experiment 1). Dynamics of the visual parameter defining the size of the shape was extracted from video records of natural lip movements. The vertical extend of natural lips, as referred as 'A parameter' in the literature, controlled the variation of the radius of the disk varying in size. Two visual cues were derived from the horizontal bars cue present in Experiment 1 and 2. A first one, called '1Dsym' (one dimension - symmetric) had its vertical extend defined by the 'A parameter' and had its global movement centered on the vertical axis. A second one, called '1D' (one dimension) had its vertical extend defined by the 'A parameter' and the upper bars had the same vertical coordinate than the video-recorded upper lip. A single f0 difference was used in this design because it had been demonstrated (in Experiment 2) that f0 difference has no effect. Two conditions of modulation were used here: 'modulation' and 'no-modulation'. The experimental procedure was the same as in Experiment 1 and Experiment 2.

4.2 Results

A repeated-measure ANOVA with factors visual condition and envelop modulation was performed. Figure 6 shows percent of correct pairing for all participants averaged for each visual condition, and envelop modulation. Visual condition has no significant effect on performances [$F(2,12)=1.91$; $p=0.18$]. On overall, envelop modulation has no effect on performances [$F(1,6)=0.0057$; $p=0.94$]. No interaction is found between envelop modulation and visual condition [$F(2,12)=0.0124$; $p=0.98$]. Comparisons to similar conditions presented in Experiment 1 and in Experiment 3 reveal differences on averaged performances. Percent of correct identification significantly increases. The only difference introduced between Exp 1 and Exp 3 is the adding of phonetic cues. This significant difference could only be attributed to this adding.

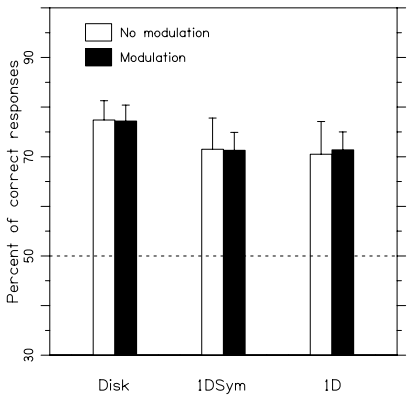

Figure 6: Percent of correct identification across the different visual cues with the two modulation conditions. All condition were significantly different from chance. Compared with similar condition in Experiment 1 (disk varying in size), identification performances increases significantly.

5 Discussion

The results found in the three experiment showed that the ability of pairing could be a consequence of an underlying binding process allowing detection of audiovisual synchrony. First surprisingly, the contrast visual display did not enabled participants to pair correctly the audio and visual stimuli even if physical synchrony was ensured. Second, the movement feature was the most relevant visual feature allowing pairing of auditory speech with non speech visual cue. In the experiment 1, we asked participants to detect synchrony between audio and the close state of the visual display. This was in contradiction with the natural lip movement, which would have been related to the open state of our visual displays. Moreover, the vertical range of the bars remained the same for all the vowels. As a consequence, no account for a speech specific processing could explain the better performance of pairing. When we provided more audiovisual coherence, thanks to the auditory envelop modulation, the detection of synchrony was not enhanced. Moreover the auditory stream organization induced by the difference of fundamental frequency between the two vowel streams did not impact the pairing. In sum, the features influencing the organization of the auditory input did not affect the pairing processing and thus the underlying binding. Finally, the effect of the phonetic discrimination across the different vowels was relevant. The same visual feature (disk varying in size), which contains or not this phonetic cue, provided better pairing performances when the phonetic parameter was present. The underlying perceptual binding could have been overruled by a phonetic processing which could have enhanced pairing.

6 Acknowledgments

This work was supported by Grants from the Region Rhones-Alpes Auvergne 'Cluster HVN 2007' and the Agence Nationale de Recherche (ANR-08-BLAN-0167-01). Special thanks to running participants.

References

[1] D. W. Massaro and M. M. Cohen, "Evaluation and integration of visual and auditory information in speech perception." *J Exp Psychol Hum Percept Perform*, vol. 9, no. 5, pp. 753–771, Oct 1983.

[2] J. Kim and C. Davis, "Hearing foreign voices: does knowing what is said affect visual-masked-speech detection?" *Perception*, vol. 32, no. 1, pp. 111–120, 2003.

[3] L. E. Bernstein, E. T. A. Jr., and S. Takayanagi, "Auditory speech detection in noise enhanced by lipreading," *Speech Communication*, vol. 44, pp. 5–18, 2004.

[4] D. W. Massaro, M. M. Cohen, and P. M. Smeele, "Perception of asynchronous and conflicting visual and auditory speech." *J Acoust Soc Am*, vol. 100, no. 3, pp. 1777–1786, Sep 1996.

[5] T. Rahne, M. Beckmann, 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, May 2007.

[6] Schwartz, Berthommier, and Savariaux, "Auditory syllabic identification enhanced by non-informative visible speech," in *Audio Visual Speech Perception*, 2003.

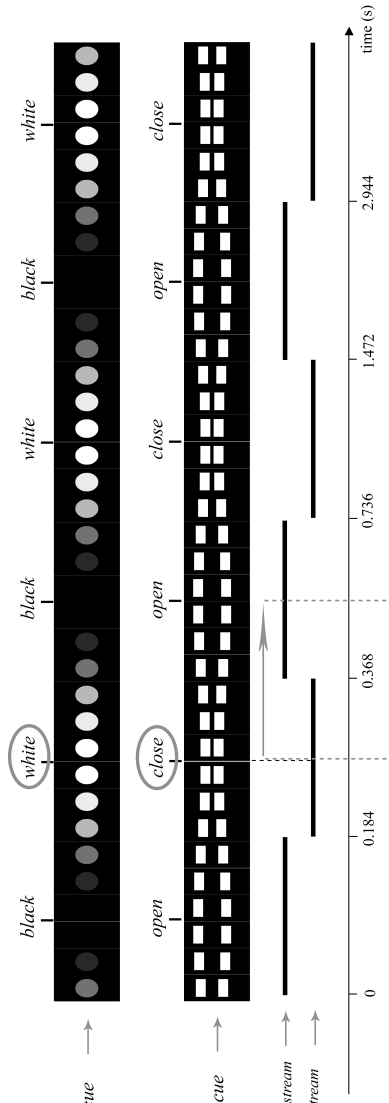

Figure 7: Representation of the stimuli on the temporal axis. The 2 different types of visual cue are represented. Frames shows the variation of the visual cues over time. Each initial and ending state of the cue is tagged. Duration of the temporal window of integration was defined equal to 250ms. Since the duration of the integration window was longer than duration of the vowels, uncertainty in binding could have appeared. For example, the second close state in the movement cue could either have been paired with the vowel starting at 0.184ms or with the vowel starting at 0.368ms because each vowel dropped in the span of the temporal window of integration. This hypothesis could account for the perceptual ambiguity.