Audiovisual streaming in voicing perception: new evidence for a low-level interaction between audio and visual modalities

Frédéric Berthommier, Jean-Luc Schwartz

Audiovisual streaming in voicing perception: new evidence for a low-level interaction between audio and visual modalities

Frédéric Berthommier, Jean-Luc Schwartz

GIPSA-Lab, UMR CNRS 5216 – Grenoble University, France

frederic.berthommier, jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

Abstract

Speech Audio-visual (AV) interaction has been considered for redundancy and complementary properties at the phonetic level but a few experiments have shown a significant role in early auditory analysis. A new paradigm is proposed which uses the pre-voicing component (PVC) excised from a true /b/. When the so called target PVC is added up to a /p/ this leads to the clear perception of /b/. Moreover, the amplitude variation of the target PVC allows building of a perceptual continuum between /p/ when amplitude is set at 0 and /b/ at original amplitude. In the audio channel, adding a series of PVC at fixed low amplitude before and after the target allows the creation of a stream of regular sounds, which are not related to visible events. On the contrary, the bilabial aperture of the /p/ is a specific speech gesture visible in the video channel. The target PVC and the visible gesture are also not redundant events. Then, depending on its intensity level, the target PVC added to an audio /p/ could be either embedded to a stream of other PVCs or phonetically fused to perceive /b/. To study the competition between these two alternatives and the role of the AV interaction, we use a 2*2 factorial design to contrast Clear/Stream and Audio/AV conditions with a control of the amplitude of the target PVC. There is no stream of PVCs in the "Clear" condition for providing the baseline. The streaming effect by itself is significant in the audio condition, but the novelty is that we find a strong AV interaction. When a stream of PVCs is present, in the "AV" condition, the rate of perceived /p/ is higher than in the "Audio" condition, suggesting that the video lip opening gestures increases the trend to isolate the formant trajectory towards the vowel from the PVC, hence increasing the perception of unvoiced stimuli. We conclude that the process of low level audio streaming is reinforced when the visual information is not redundant, and that, in this case, the phonetic fusion of the voicing cue is disadvantaged by visual information.

Index Terms: auditory streaming, voicing perception, multisensory fusion, AV speech perception, scene analysis

1. Introduction

The voicing cue has been used for showing an early interaction between audition and vision (Schwartz et al., 2004) since this is a non visible feature. When added to crowd noise (irregular, unpredictable, composed of many overlapped speeches), the effect of vision of the lips was to favor the detection of the pre-voicing component (PVC) and then the grouping of the feature with the following consonant, then perceived as voiced. This effect due to the vision was attributed to the anticipatory movement of the lips, allowing focusing on the target PVC and the identification of the voiced consonant, instead of the unvoiced counterpart (e.g. /d/ vs. /t/). Moreover, the interfering noise carries little confusion with the target speech and no redundancy at all with visible movements of lips. The task was to discriminate speech and crowd noise similar to non speech. In another experiment using two concurrent speeches, it has been shown by Driver (1996) that the intelligibility of the background speech can be increased by the vision of the foreground speaker. The confusion between foreground and background is also greatly decreased by the contribution of visual information and the segmentation of the two concurrent speeches is improved. This effect is rather uncontrolled and this is likely a consequence of phonetic integration. Recent experiments (Devergie et al., 2011) show an enhancement of the auditory streaming of vowel sequences by visible lip movement. The new paradigm we propose (Fig. 1) is based on a careful control of the confusion/intersect between foreground and background, as well as of the speech content, in order to analyze how this intersection is processed under the influence of vision thanks to the exclusive allocation rule (Bregman, 1990): the target PVC will be assigned to the background stream or to the phoneme to form a /b/. New evidence in favor of low-level functional interactions between audition and vision are provided in the following.

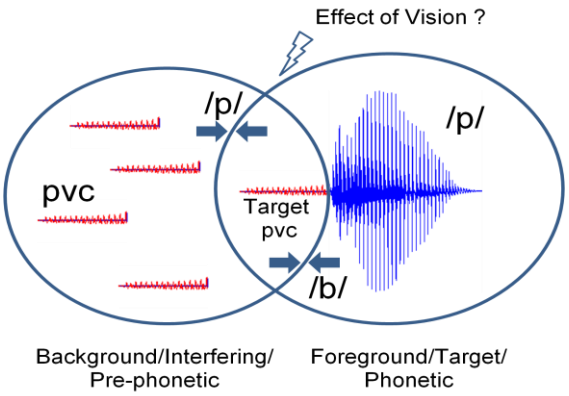

Figure 1: *Aim of the new paradigm.*

2. Experiment procedure

2.1. Material

First, a continuum between /p/ and /b/ is realized by varying the amplitude of a unique pre-voicing segment obtained from a /b/. This target PVC is spliced with a /p/ at four different log-scaled amplitude levels [0.125 0.25 0.5 1] relative to the original level and the first element is the original /p/ (Fig. 2). Since this is log-scaled, the five indexes, from 1 to 5, are taken as quantitative support of the PVC amplitude (PVCA). Amplitude is set at 0 for PVCA=1 and at 1 for PVCA=5. Firstly, an audio-visual sequence of 8 /pa/ and 8 /ba/ items, randomized to avoid any visual bias, was recorded by a male speaker at a regular rhythm. The total duration is 49.2s and the time of occurrence of the 16 plosives was measured at the time of release. The audio content was generated by insertion of the same /pa/ sample at each of these 16 mounting points, so as to preserve the audio-visual synchrony. Then, the target PVC was added at the same points with a randomly selected level within the five amplitudes defined before. The syllable rate was about 0.3 Hz, allowing large pauses in which 34 isolated PVCs with constant amplitude at 0.125 (PVCA=2) have been inserted regularly in the "stream" condition so that the PVC rate including the target PVC was about 1Hz (Fig. 3). Using this principle, we prepared a 2*2 factorial design, Clear/Stream and Audio/AV, by concatenation of 20 sequences of 16 syllables, allowing a total of 320 stimuli. In the "Clear" condition there is no supplementary stream of PVCs. In the "Audio" condition, the generated audio is mounted with a static image of lips at intermediate aperture. In the "AV" condition this is mounted with the original video track, including lip-opening gestures. The 5 levels are randomly distributed across the "Clear" and the "Stream" condition separately so that the 2 distributions are paired between A and AV in order to avoid a bias. The 4 conditions are also randomized across the 20 sequences.

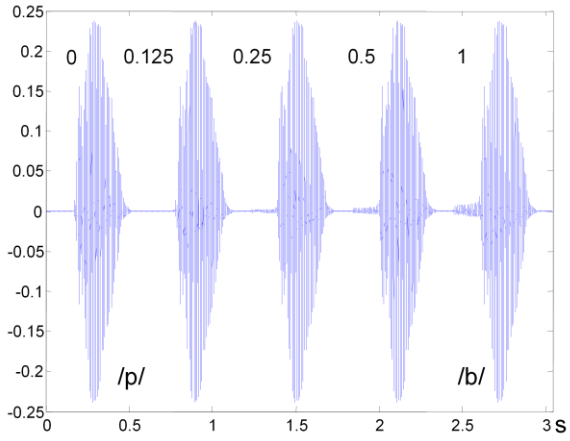

Figure 2: *The /p/-/b/ continuum with PVC amplitude.*

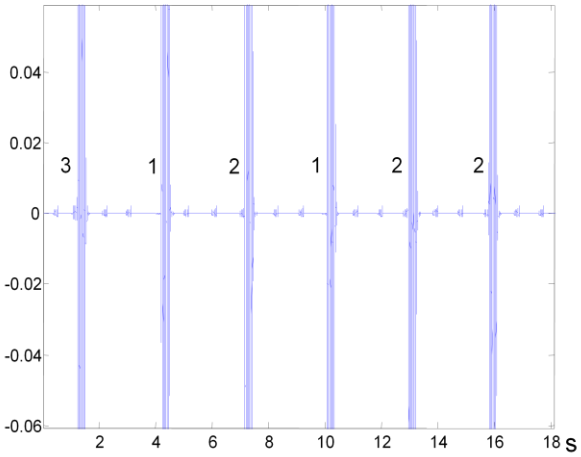

Figure 3: *Sample.wav "Stream" sequence with PVCA noted.*

2.2. Experiment and data processing

The 16*4*5=320 target stimuli are presented with headphones in a soundproof room, in a same long movie having 16.4 mn duration, using the *Presentation* software with a 2 forced-choice response paradigm. The task was to press online a left arrow key for /p/ identification and a right arrow key for /b/, for each occurrence. Eleven subjects (5 males and 6 females) performed the experiment. They were not previously informed about the existence of the background stream so that they didn't pay too much attention to it because its amplitude is weak. The response time was also recorded and used for identification of missing responses as well as for further analysis. The response delay was calculated according to the mounting points described before. When 2 responses were given during the interval separating two consecutive mounting points, or when no response was given, the response was considered as missing. Only a few responses (15 altogether for all subjects) were missing.

3. Results and discussion

3.1. Analysis of identification responses

The average response rate for /p/ according to the PVC amplitude is displayed on Fig. 4. The /p/ response rate decreases rapidly when the amplitude of the target PVC increases for the "Clear" condition, for all conditions, but with shifts of the global curves between conditions. The shift of "Stream-Audio" (S-A) relative to "Clear-Audio" (C-A) is about one point of PVCA. Basically, this means that the target PVC needs to have an amplitude twice larger than the amplitude of the PVC competitors set at (PVCA=2) to "escape" from the stream and lead to a /b/ response at the same rate. Decrease is much slower with a sustained /p/ response rate in the S-AV condition, indicating a strong interaction with the visual input. It has been checked that visual /pa/ and /ba/ have the same response rates.

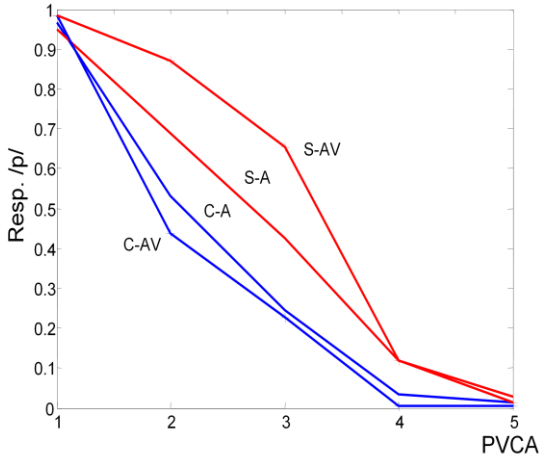

Figure 4: *Average results of identification responses.*

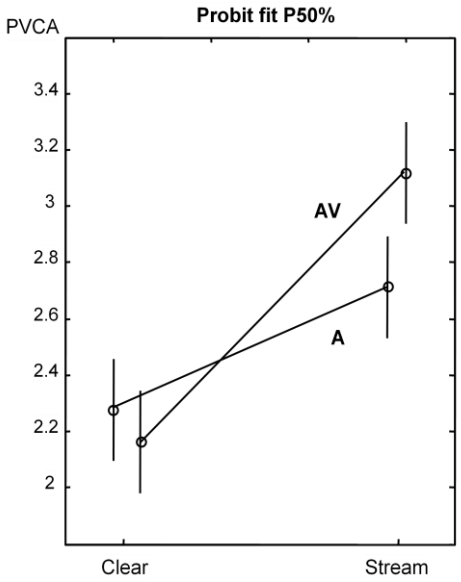

Figure 5: *Interaction between the 2 factors.*

The statistical data analysis is performed using a per-subject *probit* fit for each of the four conditions, with PVCA as a controlled variable. For each subject, abscissa of intersect between the *probit* fit and the 50% response point (P50%) is measured for each of the 4 conditions. A three ways ANOVA is applied on these data with subject included as a factor to remove across-subject variability. The effect of the "Stream" factor is highly significant ($p<0.001$) and the interaction between A/AV and stream factor is also significant ($p=0.011$). The post-hoc analysis with error bars in Fig. 5 shows that S-A is significantly different from Clear conditions as well as from S-AV. The two clear conditions are not significantly different. This indicates that streaming has an effect in the audio modality, greatly reinforced in the audio-visual modality.

To explain the decrease of the /b/ response rate in the "Stream" condition, an adaptation mechanism could be invoked. With a high rate of voicing cue, the units responsive for /b/ could be saturated, leading to an increase in the rate of /p/ responses. Adaptation could occur either at a pre-phonetic level with units specific for the voicing features or at a phonetic level by adaptation of the units responsive for the voiced consonants, including the /b/. Adaptation at a phonetic level is discarded because there is no reason to observe a supplementary adaptation of the /b/ unit in the AV condition. On the other hand, the streaming of sounds according to their regularity is a well-known phenomenon (van Noorden, 1975, Bregman, 1990) and the target PVC could be embedded in the stream of the other PVCs because it is well interleaved (see Fig. 3). In this case, this is a general mechanism available for all sounds and the level of streaming could be the primary auditory cortex or lower auditory areas (Micheyl et al., 2007). The effect of vision is either due to an increase of the embedding of the target PVC in the primary levels and by its exclusion at the higher phonetic levels. The presence of the target PVC in the speech stream is not apparent in the visual stream, which could favor its rejection because it is already integrated in the audio background in the primary levels. Whatever the underlying mechanism, this is a demonstration of interplay between a low-level auditory mechanism and multimodal integration for feature assignment.

3.2. Analysis of response time

The average response time over all subjects are plotted according to the PVCA in Figure 6. We observe that the curves have their minimum at PVCA=1 and 5 and a maximum in between. For each condition, the position of the maximum appears to be related to the P50% point present in Figure 4. In other words, the maximum of response time is related to the maximum of uncertainty.

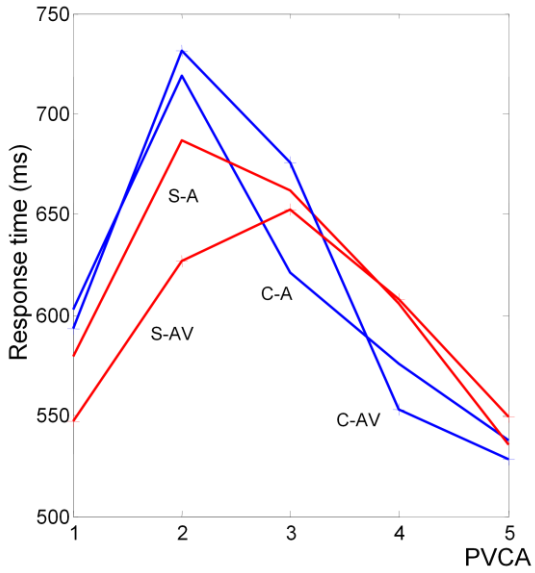

Figure 6: *Response time vs. PVCA.*

To assess this relationship, abscissa of the maximum (Max RT) has been evaluated subject by subject for each condition after a fine *spline* interpolation. This

allows evaluating continuous values of PVCA after a local processing of the peaks. The plotting of the 4*11 maximal RT positions *vs.* the P50% points shows a strong linear relationship (Fig. 7). The correlation is 0.88 after removal of a unique spurious point (circled in Fig. 7). This high correlation shows that the response time is mainly explained by the degree of uncertainty and not by the involvement of supplementary delays, e.g. due to audiovisual integration. There is no difference between C-A and C-AV, and the difference between S-A and S-AV is related to the shift of the identification curve, not due to audiovisual integration.

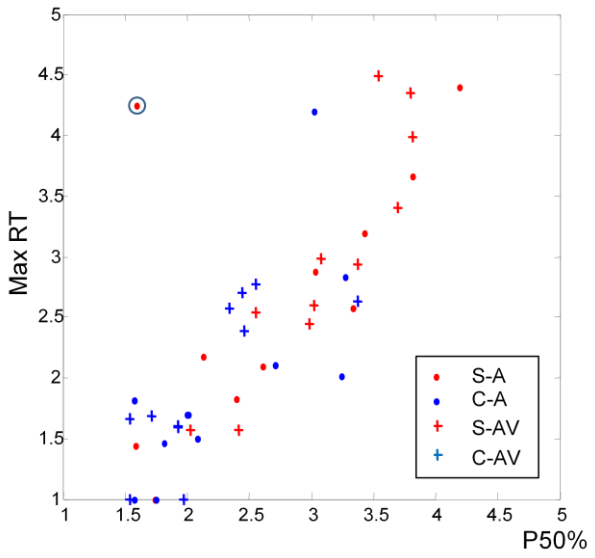

Figure 7: *Correlation between Max RT and P50%*

4. Conclusion

We have found a strong interaction between the vision of the lips movements and the auditory streaming of the pre-voicing component. When a stream of PVCs is presented in the background, the target PVC is significantly embedded in this stream in the audio-only condition, and a greater proportion of /p/ is perceived, due to its exclusive allocation to the stream. The vision of the lips does not favor the speech specific grouping of the target PVC together with the simultaneous /p/ to form a /b/ but on the contrary aids to discard it from the foreground speech and to favor its embedding in the audio background stream (Fig. 8). Remarkably, the sequential grouping of the PVCs is favored against the simultaneous grouping of "target PVC" and /p/. This is because the /p/ only is visible and redundant (both sound and video carry information about the /p/). Then, the grouping of non-visible *vs.* visible components is improved, rather than the grouping of speech *vs.* background components, which is the alternative solution. This could be related to direct influences coming from multimodal regions on primary areas as well as the result of higher-level mechanisms. The target PVC might be simply gated in speech specific/phonetic levels under the influence of vision and then rejected to primary levels, without any *per se* interaction with primary auditory levels.

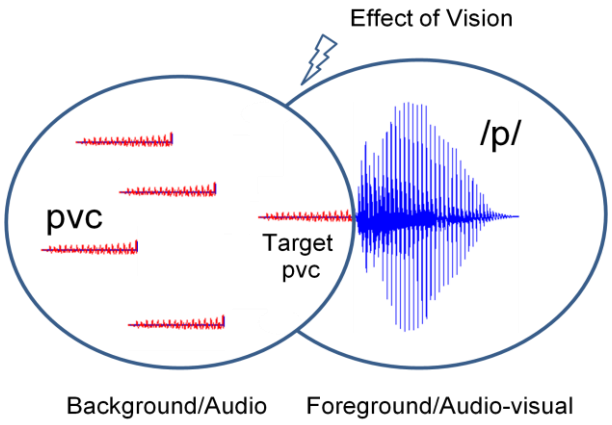

Figure 8: *Effect of vision of the lips movement.*

5. 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).

6. References

[1] Bregman, A. S., "Auditory scene analysis", MIT Press: Cambridge, MA, 1990.

[2] Devergie, A., Grimault, N., Gaudrain, E., Healy, E., and Berthommier, F., "The effect of lip-reading on primary stream segregation", *J. Acoust. Soc. Am.* (in press).

[3] Driver, J., "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading", *Nature* 381: 6577, 66-68, 1996.

[4] Micheyl, C., Carlyon, R.P., Gutschalk, A., Melcher, J.R.M., Oxenham, A.J., Rauschecker, J.P., Tian, B. and Wilson, E.C., "The role of auditory cortex in the formation of auditory streams", *Hearing Research* 229, 1161-131, 2007.

[5] van Noorden, L., "Temporal coherence in the perception of tone sequences", Unpublished doctoral dissertation, Technische Hogeschool Eindhovern, Eindhoven, The Netherlands, 1975.

[6] Schwartz, J.L., Berthommier, F., and Savariaux, C., "Seeing to hear better: Evidence for early audio-visual interactions in speech identification", *Cognition* 93, B69–B78, 2004.