

Perception as a (Shaped) Mirror of Action: It Seems Easier to Lipread One's Own Speech Gestures than those of Somebody Else

Laureline Arnaud, Jean-Luc Schwartz, H el ene L oevenbruck & Christophe Savariaux

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

laureline.arnaud@wanadoo.fr; {Jean-Luc.Schwartz; Helene.Loevenbruck;
Christophe.Savariaux }@gipsa-lab.inpg.fr

INTRODUCTION

Since the first papers by Liberman and colleagues suggesting and then claiming that speech perception is based on the recovery of articulatory gestures from the sound and sight of a speaking partner (Liberman *et al.*, 1952, 1967, Liberman and Mattingly, 1985, Liberman and Whalen, 2000), the debate has been animated, between “auditory” and “motor” theorists. The discovery of “mirror neurons” in the premotor and parietal cortices of the monkey, responding both to the execution and observation of an action, has greatly enhanced and renewed the appeal of motor theories (Gallese *et al.*, 1996; Rizzolatti *et al.*, 1996). Further evidence of motor responses to speech percepts in the frontal and parietal cortices in humans has strengthened the likeliness of the involvement of motor representations in speech perception (Fadiga *et al.* 1995, Iacoboni *et al.* 1999). However, direct evidence in favour of a *functional* role of motor areas and representations in the processing of speech sounds and sights remains rare and controversial. Outside the speech communication domain, Viviani introduced the concept of “motor procedural knowledge” in movement perception. He displayed convincing evidence that motor procedural knowledge related to hand movements could well play a role in the processing of perceived hand trajectories, both in terms of kinematics and geometry (Viviani and Stucchi, 1992). Viviani and colleagues also mention a very interesting experiment by Beardworth and Buckner (1981) in which the authors studied the ability to recognise one's own versus somebody else's movements from a recorded point-light display of walking movements. The subjects in the experiment were a group of college students who knew each other quite well. It appears that certain subjects were better at recognising themselves than at recognising their friends. The crucial point in the interpretation is that they had never seen themselves walking from an external point of view, while they had seen their friends walking every day. Therefore, this result is interpreted by the authors as suggesting “some sort of kinesthetic-visual cross-modal transfer” (p. 19).

EXPERIMENTAL PARADIGM

Beardworth and Buckner's experiment seems to provide a strong and direct evidence that action may guide or complete perception in some cases. When thinking about speech, we thought that it could be interesting to apply the same kind of idea to the problem of speech recognition. Of course, studying the *auditory* identification of one's own vs. somebody else's utterances does not seem very judicious, since one hears his/her own voice much more than other voices, even if the auditory pathway is partly different for one's own voice and for the other voices. But *visual* identification should not suffer from the same objection: one almost never has the occasion to lipread his/her own gestures. A few years ago, we prepared a first experiment on the lipreading of self vs. other speech gestures, with the idea that if a benefit of seeing one's own gestures could be demonstrated, it would provide a nice and strong cue in favour of the role of “motor procedural knowledge” in the elaboration of speech perceptual representation. The experiment involved lipreading French spoken digits uttered by 6 speakers and identified by the same 6 subjects. The result was unsuccessful (Schwartz *et al.* 2001). Indeed, there was a slight advantage of “self” over “other” lipreading scores, but the gain was not significant. However, this experiment had been done on a corpus which had not been specifically conceived for this purpose. A number of recent findings suggest that the involvement of the motor system could depend on the complexity of the task (*e.g.* Zekveld *et al.*, 2006, Skipper *et al.*, 2007). This led us to conceive a new experiment focussed on a perception task in which the role of motor

procedural knowledge could be expected as critical. Perception of stimuli involving vowel reduction is such a task. In a series of production and perception experiments, Loevenbruck and colleagues had shown that in a sequence such as [iaɪ] vs. [iɛɪ], varying from a close to an open then back to a close vowel, the most open configuration, which should have distinguished the target /a/ vs. /ɛ/, led in fact to poor classification because of reduction (mainly controlled by speech rate). However, the dynamic trajectory could play a crucial role in disentangling one target from the other. Presenting the entire vowel trajectory improved the perceptual scores, and articulatory modelling and inversion techniques confirmed the role of speech dynamics to recover the hidden motor target (Loevenbruck and Perrier, 1996; Perrier *et al.*, 1996). The present experiment capitalised on this idea.

METHODS AND RESULTS

We recorded 4 French speakers producing the sentence “Nacim y a immédiatement parlé” (“Nacim has immediately spoken there”) or “Nacim y est immédiatement parti” (“Nacim has immediately gone there”), contrasting the sequences [iaɪ] vs. [iɛɪ] in a similar environment, embedded in the same left and right [m] context imposing a lip closure. The lips were made up in blue, to be able to use the ICP lip-measuring system based on chroma-key (Liptrack, see Lallouache, 1991). The sentences were uttered in four conditions: two rates (“normal” vs. “fast”) and two focus conditions (broad focus vs. narrow focus on the target vowel (the auxiliary verb), the latter obtained by a correction procedure: “Nacim y a immédiatement” or “Nacim y est immédiatement”). Ten repetitions of each sequence were recorded. The video sampling frequency was 50 Hz (with non-interleaved frames).

All sequences were then carefully analysed in terms of lip dynamics, which enabled to select one occurrence of each sequence, in which the trajectory seemed clearly visible. A gating paradigm was then conceived: we prepared a series of movies all starting from the initial closed lips in [m] including the following [i] and extending to various portions of the opening trajectory towards the [a] or the [ɛ], until the most open configuration was reached. Since the duration of the opening phase was variable, the number of 20-ms steps per sequence was also variable, between 6 and 16, but care was taken to obtain altogether a fixed number of 80 movies per speaker (2 vowels, 2 rates, 2 focus conditions, mean number of 10 steps per sequence). For each sequence, the image with the most open configuration was also selected for a static perception experiment (8 images per speaker).

The speakers were associated by pairs. A pair of subjects had to visually identify the vowels produced by him/herself and by the associated speaker in the pair. For each subject, the experiment included two parts: (1) a “static” perception experiment, with the 8 images per speaker, presented 10 times each, in a random order in which the two speakers appeared randomly to prevent attentional biases (8 image per speaker, 2 speakers, 10 repetitions, 160 stimuli altogether); (2) a “dynamic” experiment in which the 80 movies per speaker were displayed 5 times each in a random order in which the two speakers appeared randomly. The static condition always preceded the dynamic one, and they were analysed separately.

Altogether, a significant difference appears in favour of the “self” recognition in both cases. The difference is significant for static images (probabilities of correct responses, averaged over vowel, rate, focus, and subject: 78% for self vs. 69% for other, $\chi^2(1)=6.76$, $p<0.01$). It is not significant for dynamic stimuli considered gating step by gating step, but the advantage for “self” over “other” recognition is displayed for a large part of the trajectory. Moreover once integrated over the 6 final gating steps (120 ms) it becomes significant (probabilities of correct responses, averaged over vowel, rate, focus, step and subject: 63% for self vs. 58% for other, $\chi^2(1)=6.35$, $p<0.025$).

DISCUSSION AND CONCLUSION

These results are very encouraging. They suggest that motor procedural knowledge could indeed intervene in speech perception. The fact that a gain appears for static as well as dynamic stimuli is a bit puzzling, particularly in light of the debate about static vs. dynamic perception of vowels (see *e.g.* Strange, 1999), and of possible differences in processing in the human brain (Calvert and Campbell, 2003). It suggests that articulatory *postures* can also engage the listener in a mental recovery of articulatory strategies. This pilot experiment has to be strengthened by further experiments with more subjects. If confirmed, this would suggest that face to face interaction in speech communication involves a subtle and complex combination of perceptual processing and motor recovery processes (Schwartz *et al.*, 2002, 2007).