Sensory-motor interactions in speech perception, production and imitation: behavioral evidence from close shadowing, perceptuo-motor phonemic organization and imitative changes.

Lucie Scarbel, Denis Beautemps, Jean-Luc Schwartz, Marc Sato

Sensory-motor interactions in speech perception, production and imitation: behavioral evidence from close shadowing, perceptuo-motor phonemic organization and imitative changes.

Lucie Scarbel[1], Denis Beautemps[1], Jean-Luc Schwartz[1], Marc Sato[1]

[1] *Gipsa-lab,Département Parole & Cognition, CNRS & Grenoble Université, Grenoble, France*

lucie.scarbel@gipsa-lab.grenoble-inp.fr, denis.beautemps@gipsa-lab.grenoble-inp.fr, jean-luc.schwartz@gipsa-lab.grenoble-inp.fr, marc.sato@gipsa-lab.grenoble-inp.fr

Abstract

Speech communication can be viewed as an interactive process involving a functional coupling between sensory and motor systems. In the present study, we combined three classical experimental paradigms to further test perceptuo-motor interactions in both speech perception and production. In a first close shadowing experiment, auditory and audio-visual syllable identification led to faster oral than manual responses. In a second experiment, participants were asked to produce and to listen to French vowels, varying from height feature, in order to test perceptuo-motor phonemic organization and idiosyncrasies. In a third experiment, online imitative changes on the fundamental frequency in relation to acoustic vowel targets were observed in a non-interactive situation of communication during both unintentional and voluntary imitative production tasks. Altogether our results appear exquisitely in line with a functional coupling between action and perception speech systems and provide further evidence for a sensory-motor nature of speech representations.

Keywords: speech perception, speech production, sensory motor interaction

1. Introduction

An old and classical debate in the speech communication domain concerns the possible motor implication in speech perception and, more generally, the auditory versus motor nature of the speech code. Auditory theories assume that speech perceptual processing and categorization are based on acoustic cues and auditory representations (Stevens and Blumstein 1978, 1979; Lindblom et al. 1988, 1990) Conversely, the motor theory of speech perception (Liberman et al., 1985) and its direct realist variant (Fowler et al., 1986) claim that there is a crucial role of the motor system in speech perception. More recently, a number of perceptuo-motor theories attempted various kinds of syntheses of arguments by tenants of both auditory and motor theories, proposing that implicit motor knowledge and motor representations are used in relationship with auditory representations and processes to elaborate phonetic decisions (Skipper et al., 2007; Schwartz et al., 2012).

Various experimental settings enable to test and study the relationship between speech perception and action. Let us describe three of them which provide the basis for the present work. First, close-shadowing provides a natural paradigm for testing perceptuo-motor links. Indeed, Porter et al. (1984) and later Fowler et al. (2003) observed very fast reaction times when participants had to shadow a syllable as quickly as possible. Compared to manual responses, oral speech responses were also found quicker than manual ones (Galantucci et al., 2006). This difference was interpreted by the theoretical assumption that perceiving speech is perceiving gestures, and that gesture perception directly controls speech response and makes it faster.

Another way to prove the evidence of a perceptuo-motor linkage is to directly test the existence of a common perceptual and motor phonemic organization. From that view, Bell-Berti (1979) showed that differences between subjects in the perception of the [i] versus [I] contrast in American English seemed to be linked to differences in the articulatory implementation of this contrast. Menard et al. (in press) further showed similar idiosyncrasies in both vowel production and perception, a result suggesting a link between perceptual and motor phonemic prototypes in the human brain.

Finally, the ability to converge and to imitate a listener also attests of a perceptuo-motor coupling. Recently, online unintentional and voluntary imitative changes in relevant acoustic features of vowel targets were observed during speech production in a non-interactive situation of communication (e.g., Garnier et al., 2013; Sato et al., 2013). These results were explained by the possibility that speech production continuously draws on perceptuo-motor learning from the external speech environment and prior listener's sensory-motor knowledge.

In the present study, we further tested sensory-motor interaction in these three paradigms. In a close shadowing experiment (Experiment A), we compared reaction times to auditory and audio-visual speech stimuli from manual and oral responses. We expected to find faster reaction times to oral compared to manual responses, and to audiovisual compared to auditory stimuli. The second experiment (Experiment B) tested perceptuo-motor phonemic organization in vowel production and perception. Our aim was to possibly determine a common phonemic organization in vowel perception and production as well as to test subtle perceptuo-motor idiosyncrasies between participants. Finally, the third experiment (Experiment C) concerned phonetic convergence and voluntary imitative changes in relation to acoustic vowel targets.

2. Methods

2.1. Participants

Three groups of respectively fifteen, twenty-seven and sixteen healthy adults, native French adults, participated in Experiments A, B and C. All participants had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders.

2.2. Stimuli

2.2.1. Experiment A

Multiple utterances of /apa/, /ata/ and /aka/ sequences were individually produced by a male native French speaker (who did not participate in the experiments) in a sound-attenuated room. The corpus was audio-visually recorded with the objective to obtain 4 different occurrences of /apa/, /ata/ and /aka/ with various durations of the initial /a/ vowel (i.e., 0.5s, 1s, 1.5s and 2s) so as to obtain 12 distinct stimuli.

2.2.2. Experiment B

Thirteen acoustic stimuli were used for the vowel perception task of Experiment B. Those stimuli were synthesized from VLAM (Variable Linear Articulatory Model), an articulatory-to-acoustic model of the vocal tract based on Maeda's adult model (Boe and Maeda, 1997; Boe, 1999). Using VLAM, we generated thirteen stimuli distributed regularly within the maximal adult vowel space from high to low front unrounded vowels.

2.2.3. Experiment C

A vowel database was created from /e/, /oe/, /o/ French vowels produced by two male and female speakers. From these stimuli, f0 was artificially shifted by steps of ±5Hz (from 80Hz to 180Hz for the male vowels, and from 150 to 350Hz for the female vowels) using the PSOLA module integrated in Praat software (Boersma and Weenink, 2013).

2.3. Experimental procedure

The three experiments were carried out in a sound-proof room. Participants sat in front of a computer monitor at a distance of approximately 50 cm. The acoustic stimuli were presented at a comfortable sound level through a loudspeaker, with the same sound level set for all participants. The Presentation software (Neurobehavioral Systems, Albany, CA) was used to control the stimulus presentation during all experiments, and to record key responses in Experiment A and B (see below). All participants' productions were recorded for off-line analyses.

2.3.1. Experiment A

The experiment consisted of two categorization tasks: close-shadowing in one case, where the responses were provided orally by repeating as quickly as possible the presented speech sequence; manual decision in the other case, where the responses were provided manually, by pressing as quickly as possible the appropriate key. The stimuli to categorize consisted in /apa/, /ata/ and /aka/ sequences. For each task (oral vs. manual response) and each modality (auditory vs. audiovisual), 16 occurrences of /apa/, /ata/ and /aka/sequences were presented in a fully randomized sequence of 48 trials. The order of task and modality of presentation was fully counterbalanced across participants.

2.3.2. Experiment B

This experiment consisted of two vowel perception and production tasks, counterbalanced across participants. For the production task, participants were asked to produced fifteen repetitions of the 10 oral French vowels /i y u e ø o ɛ œ ɔ a/, according to a visual orthographic target. Target vowels were presented in a fully randomized order. For the perception task, participants had to manually categorize acoustic stimuli among the four front unrounded French vowels /i ɛ e a/. Each stimulus was presented ten times in a fully randomized order.

2.3.3. Experiment C

Experiment C consisted in three vowel production tasks. First participants had to individually produce /e/, /œ/ and /o/ vowels, according to a visual orthographic target. This allows the experimenter to measure participant's f0. In the subsequent task, participants were asked to produce the three vowels according to an acoustic target. Importantly, no instruction to "repeat" or to "imitate" the acoustic targets was given to the participants. Finally, the third task was the same as the second task except that participants were explicitly asked to imitate the acoustic targets. The only indication given to participants was to imitate the voice characteristics of the perceived speaker. Acoustic target for each participant were 27 stimuli selected from the vowel database, with the 9 quantified f0 frequencies varying from -20% to +20% by steps of 5% around his/her own pitch, as measured in the first task.

2.4. Data analysis

All acoustic analyses of participants' productions were performed using Praat software (Boersma and Weenink, 2013).

2.4.1. Experiment A

The proportion of correct responses was determined for each participant and each condition, together with reactions times (RTs) for correct responses. RT in the oral task was estimated from the burst onset of the stop consonant to categorize to the burst onset of the oral response.

2.4.2. Experiment B

For the production task, the mean F1 frequency for /i e ɛ a/ was computed for each participant. In all this study, frequencies are estimated in bark, thanks to to the formula proposed by Schroeder et al. (1979)
For the perception task, the mean F1 frequency of all stimuli categorized respectively as /i/, /e/, /ɛ/ or /a/ was determined for each participant. For both the perception and production tasks, mean normalized bark values for /e/ and /ɛ/ with regard to their distance from /a/ and /i/ was then calculated. Correlation scores between production and perception was finally determined for all participants.

2.4.3. Experiment C

In all tasks of Experiment C, we measured f0 for each produced vowel. In the second and third tasks, correlation analyses between f0 values in the perceived and produced vowels were performed for each participant.

3. Results

3.1. Experiment A - see Figure 1

3.1.1. Reaction times

RTs were entered into an ANOVA with three factors: modality (auditory, audiovisual), response (speech, key) and syllable (/pa/, /ta/, /ka/). Although no significant difference between the auditory and audiovisual stimuli was observed, RTs were shorter for speech responses (240 ms) than for key responses (462 ms) ($F(1,14)=81.8$; $p<0.001$). Interestingly, an interaction between the three factors was found ($F(2,28)=4.6$; $p<0.01$). While, for speech responses, RTs for /pa/ did not differ between the auditory (196ms) and audiovisual (208ms) modalities, for key responses an audiovisual advantage was

observed (audiovisual stimuli: 415 ms, audio stimuli: 442 ms). For /ta/, manual RTs were longer for audiovisual (506 ms) than for auditory (465 ms). For /ka/, no differences were found between the modalities and tasks.

3.1.2. Perceptual recognition

As for RTs, the percentage of correct responses were entered into an ANOVA with three factors: modality (auditory, audiovisual), response (speech, key) and syllable (/pa/, /ta/, /ka/). No difference was observed between auditory (95%) and audiovisual (94%) stimuli. However, participants made significantly fewer errors for key (97%) than for speech responses (93%) ($F(1,14)=13$; $p<0.002$), and fewer errors for /pa/ (98%) than for /ta/ and /ka/ syllables (93%) ($F(2,28)=6.8$; $p<0.004$). In addition, a significant interaction between the modalities and syllables was also observed ($F(2,28)=5.6$; $p<0.01$). For /ta/ and /ka/, more correct responses were observed for key (97% and 97%) than for speech (90% and 89%) responses. For /pa/, no difference was observed.

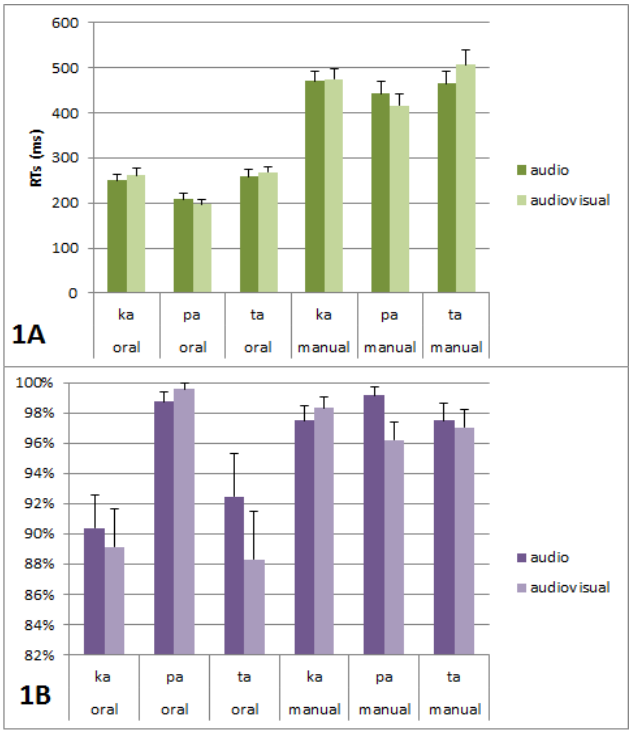

Figure 1: RTs (1A) in ms. and percentage (1B) of correct responses in Experiment A.

3.2. Experiment B - see Figure 2

In the production task, the mean F1 values for /i/, /e/, /ɛ/ and /a/ in barks were respectively 3.1 (range: 2.6-3.6), 4.4 (range: 3.4-4.5), 5.9 (range: 4.4-7.1) and 7.3 (range: 6.2-8.5). Idiosyncrasies were weak for /e/ (normalized distance from /i/ between .19 and .46 bark but with a small standard deviation at .07) and larger for /ɛ/ (normalized distance from /i/ between .35 and .88 bark with a standard deviation at .15). In the perception task, the mean F1 values for /i/, /e/, /ɛ/ and /a/ in barks were respectively 2.8 (range: 2.6-3.9), 4.2 (range: 3.8–4.5), 5.5 (5.3–5.7) and 6.8 (range: 6.6–7.0). Variability in perception was extremely small, showing that no idiosyncrasies were found between participants. From these results, a quasi perfect correlation of acoustic values between produced and perceived vowels is observed (with a mean slope for all participants of .93, range: 0.7-1.3).

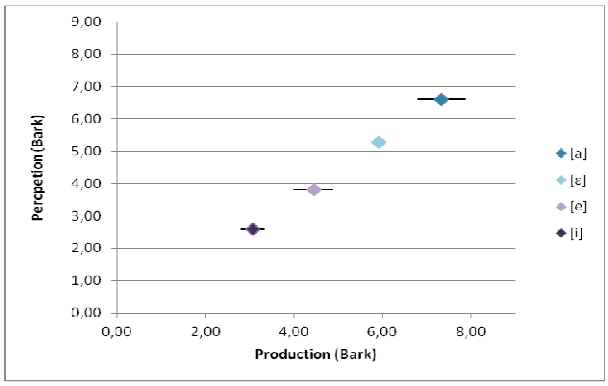

Figure 2: Mean acoustic values (in barks) for each vowel in the perception vs. production task.

3.3. Experiment C - see Figure 3

In Experiment C, imitative changes were observed in both tasks, though stronger in voluntary imitation. Slope coefficients differed significantly from zero in both the production ($t(15)=6.2$; $p<0.001$) and imitation ($t(15)=19.2$; $p<0.001$) tasks. In addition, slope coefficients were higher in the imitation (0.83) compared to the production (0.44) tasks ($t(15)=5.6$; $p<.001$). Similarly, correlation coefficients differed significantly from zero in both the production ($t(15)=8.6$; $p<.001$) and imitation ($t(15)=30.4$; $p<0.001$) tasks, and were higher in the imitation ($r=0.93$) compared to the production ($r=0.63$) tasks ($t(15)=4.2$; $p<0.001$).

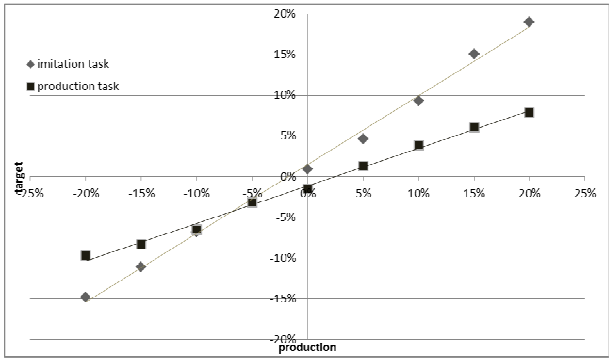

Figure 2: Phonetic convergence and voluntary imitative changes observed in Experiment C.

4. Discussion and Conclusion

4.1. Experiment A

Overall, as in the studies by Fowler et al. (2004) and Porter et al. (1984), orofacial responses were much quicker than manual ones. While no differences were found between auditory and audiovisual modalities in the close shadowing task, quicker response times were however observed in the audiovisual modality for the manual categorization task for bilabial consonants, likely due to the visible anticipatory gesture. The fact that this visual gain was not seen in the orofacial modality is probably due to a floor effect considering the small response time in close shadowing. Although these results do not provide global evidence for faster response times in the audio-visual modality in the close shadowing task, they appear compatible with a sensory-motor framework in which there is a functional connection between action and perception systems.

4.2. Experiment B

Our results for the production task appear partly coherent with those found by Ménard and Schwartz (in press). One important difference, however, is that though our study displays idiosyncrasies in production more or less in line with their study, we did not find almost any idiosyncrasy in the perception task. This difference is likely due to the different experimental factors used in these two studies. While we only tested adults, Ménard and Schwartz tested two groups of 4 and 5 years old children and one group of adults. Moreover, the stimuli used in the perception task for the adults were not the same as ours (with a larger number and type of stimuli, and a more variable distribution in the acoustic space). Given the larger variability of the stimuli used by Ménard and Schwartz (in press), idiosyncrasies are more likely to emerge. Importantly, in line with the maximal dispersion theory of Lindblom (1972) and with a perceptuo-motor coupling of vowel perception and production (Schwartz et al., 2012), we found a near to perfect acoustic equidistance between the centers of vocalic targets both in the production and perception tasks (see Figure 3).

4.3. Experiment C

As in Garnier et al. (2013) and Sato et al. (2013), we found a quasi perfect imitation of vowel targets on f0 in the voluntary imitation task, as well as clear evidence for phonetic convergence in the production task. This latter result suggests that participants tend to converge towards an acoustic speech target even if they don't imitate consciously. Altogether, these results are perfectly compatible with a perceptuo-motor linkage in speech production and perception.

4.4. General discussion

Taken together, the three experiments largely confirmed previous results and strongly suggest a functional perceptuo-motor coupling of speech perception and production systems. They provide further evidence for a sensory-motor nature of speech representations. This series of coupled experimental paradigms for studying the relationship between perceptual and motor processes will now serve as a platform for assessing the recovery of this relationships in hearing impaired subjects after cochlear implantation.

5. Acknowledgements

This work was supported by the French National Research Agency (ANR) through funding for the PlasMody project (Plasticity and Multimodality in Oral Communication for the Deaf).

6. References

Bell-Berti, F., L. J. Raphael, D. B. Pisoni, and J. R. Sawusch (1979). "Some relationships between speech production and perception". In: *Phonetica* 36, pp. 373-383.

Boë, L-J. (1999). "Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults: consequences for ontogenesis and phylogenesis". In: *Proceedings of the International Congress of Phonetic Sciences* 3, pp. 2501-2504.

Boë, L-J. and S. Maeda (1997). "Modélisation de la croissance du conduit vocal. Espace vocalique des nouveau-nés et des adultes. Conséquences pour l'ontogenese et la phylogenese ». In : *Journées d'Etudes Linguistiques : « La voyelle dans tous ses états »*, pp. 98-105.

Boersma, P. and D. Weenink (2013). "Praat: doing phonetics by computer". [Computer program], Version 5.3.42, retrieved 2 March 2013 from http://www.praat.org/.

Fowler, C. A., J.M. Brown, L. Sabadini, and J. Weihing (2003). "Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks". In: *Journal of Memory and Language* 49, pp. 296–314.

Fowler, C. A., and M. Smith (1986). "Speech perception as ''vector analysis": An approach to the problems of segmentation and invariance". In: *Invariance and variability of speech processes, Eds J. Perkell and D. Klatt (Hillsdale, NJ: Lawrence Earlbaum Associates)*, pp. 123-136.

Galantucci, B., C. A. Fowler, and M. T. Turvey (2006). "The motor theory of speech perception reviewed". In: *Psychonomic Bulletin and Review* 13, pp. 361-377.

Garnier, M, Lamalle, L., ansd Sato, M. (2013). "Neural correlates of phonetic convergence and imitation of speech". In: *Frontiers in Psychology* 4(600).

Liberman, A. M. and I. G. Mattingly (1985). "The motor theory of speech perception revised". In: *Cognition* 21 (1), pp. 1–36.

Liljencrants, J. and B. Lindblom (1972). "Numerical simulation of vowel quality systems: the role of perceptualcontrast". In: *Language* 48, pp. 839-62.

Lindblom, B. (1990). "Explaining phonetic variation : a sketch of the HandH theory". In: *Speech production and speech modelling, Eds W.J. Hardcastle. And A. Marchal (Dordrecht, The Netherlands, Kluwer Academic Publishers)*, pp. 403-439.

Ménard, L. and J-L. Schwartz (in press) : "Perceptuo-motor biases in the perceptual organization of the height feature in French vowels". In: *Acta Acustica.*

Porter, R. and F. Castellanos (1980). "Speech production measures of speech perception: Rapid shadowing of VCV syllables". In: *the Journal of the Acoustical Society of America* 67, pp. 1349–1356.

Sato, M., K. Grabski, M. Garnier, L. Granjon, J-L. Schwartz, and N. Nguyen (2013). "Converging to a common speech code: imitative and perceptuo-motor recalibration processes in speech production". In: *Frontiers in Psychology* 4(422).

Schwartz, J.L., A. Basirat, L. Ménard, and M. Sato (2010). "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception". In: *Journal of Neurolinguistics* 25, pp. 336-354.

Skipper, J. I., V. van Wassenhove, H. C. Nusbaum, and S. L. Small (2007). "Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception". In: *Cerebral Cortex* 17, pp. 2387-2399.

Stevens, K. N. and S. E. Blumstein (1978). "Invariant cues for place of articulation in stop consonants". In: *The Journal of the Acoustical Society of America* 64, pp. 1358-68.