Benefit of audiovisual presentation in close shadowing task

Denis Beautemps, Marie-Agnès Cathiard and Yvon Le Borgne
Institut de la Communication Parlée, CNRS UMR5009 / INPG/ Université Stendhal
Grenoble, France
E-mail: beautemps@icp.inpg.fr, cathiard@icp.inpg.fr

ABSTRACT

Close shadowing paradigm consists in repetition of speech as soon as the speaker produces it. This paradigm was used for the evaluation of time co-perception and production process. We used this paradigm in order to evaluate the benefit of visual lip information in case of audiovisual presentation of stimuli with a perfectly audible signal. We presented French [VCVsa] sequences made of [a, i, u, y] vowels and [p, t, k] consonants pronounced by a male French speaker for both normally articulated and hyper articulated speech. Four subjects were asked to repeat *on line* syllables for stimuli presented in the two conditions : auditory alone and audiovisually. The average reaction times measured between stimulus onset acoustic closure of the consonant and the corresponding event on the response vary between 227 to 314 ms. We found no advantage for the hyper articulated production. But a significant advantage of audiovisual presentation with a visual gain of 35 ms and 49 ms (i.e. 13.4% and 14.6%) for two of the four subjects.

1. INTRODUCTION

Shadowing definition

"Speech shadowing is an experimental task in which the subject is required to repeat (shadow) speech as he hears it. When the shadower is presented with a sentence, he will start to repeat it before he has heard all of it. The response latency to each word of a sentence can therefore be measured" ([1] p. 252). This paradigm is used to study co-perception and production processing. Reaction time (*RT*) (i.e. response delay) generally observed for isolated words or non-sense syllables were as small as 200 to 250 ms [1]. In a close shadowing task of a 300-words passage of normal prose, Marslen-Wilson ([1]) obtained, for the closest subjects, *RT* ranged from 254 to 287 ms. The analysis of the errors showed that close shadowers process in a same way the syntactic and semantic structure of speech than the distant shadowers, thus revealing the natural processing of speech in shadowing task.

V-V syllable shadowing

Porter and Lubker ([2]) presented Vowel-Vowel synthesized stimuli as [ao], [ai] and [aæ] to four subjects. The first vowel was lengthened up to 1100 - 1500 ms. The authors compared reaction time in three conditions. (i) Subject under phonation of the first vowel was asked to pronounce the second vowel as soon as he perceived it (in average 147.6 ms for *RT*). (ii) Subject in phonation of the first vowel was asked to pronounce the [o] vowel as he perceived the change in vowel (162.4 ms in average for *RT*). (iii) A third one similar to the previous one where no preliminary phonation was used before the [o] response to the vowel change (in average 187.6 ms for *RT*).

In the first condition, subjects had to repeat the second vowel initially not known contrarily to the other conditions where a simple known response to the vowel change was asked. Taking into account that reaction time of the first condition is in average not much different from the others (147.6 ms *vs* 162.4 ms or 187.6 ms), these results showed that the auditory phase needed in the identification of the second vowel (first condition) did not delayed the response production phase. The authors concluded on the following noticeable result that motor control mechanisms involved in the response largely overhead the auditory process.

VCV syllable shadowing

Similar experiments on VCV sequences were performed by Porter and Castellanos [3]. Reaction time to three conditions were tested in the presentation of VCV auditory stimuli made of [a] vowel lengthened to 2 to 5 seconds and [p, b, m, k, g] consonants as for example in [apa] sequence. (i) A shadowing task where subjects were asked to repeat in line (*RT* = 223 ms in average), (ii) subjects under phonation of the first vowel were asked to answer [ba] as they perceived the consonant (*RT* = 171 ms in average), (iii) same as the precedent condition but with no initial phonation (*RT* = 296 ms in average). The second condition where no identification was needed (corresponding to a simple reaction time) gave the quickest reaction times. Once more the authors explained the slight elevation of reaction time (223 ms *vs* 171 ms) observed in the shadowing task by a consequent overhead of the perception process needed in the first condition with the response production mechanism.

Vision and shadowing

In a series of shadowing experiments, Reisberg et al.

ISBN 1-876346-48-5

([4]) evaluated the gain with vision in identification of intact auditory stimuli, i.e. not acoustically degraded. In the repetition of 10 phrases made of 110 words each in foreign languages (French stimuli and English subjects with fair level in French) the authors obtained a gain of 15 % on identification scores with the audiovisual presentation (AV) in comparison to the auditory alone (A). This gain was computed as following: $\left|\frac{AV-A}{A}\right|$. A supplementary shadowing task made of German phrases and English subjects with poor German level was performed to dismiss the hypothesis where the gain with vision is due to a better concentration of the subjects. In the audio condition, the view of the face with mouth and chin being masked was added. Results showed a gain for the audiovisual presentation higher than 21.5% in comparison with the audio alone condition where subjects could not see neither the lips, neither the chin. In a third experiment where English phrases pronounced by a Belgium speaker were repeated by English subjects, the authors obtained a gain of 4% with the vision. In a latter experiment, English students were asked to repeat in English a part of the famous philosophic "Critique de la raison pure" text of Kant, complex for comprehension. The authors also observed a gain with the vision (8%). These latter experiments showed that a shadowing experiment even with familiar vocabulary and syntax leads to better results in audiovisual condition. In this ensemble of experiments, the gain of vision for speech perception was evidenced even if the audio signal is not degraded by noise, nor in conflict with vision. However in these results, this benefit of vision was not based on reaction time measurements but simply on scores of correctly repeated words.

More recently, Davis and Kim [5] tested the visual gain obtained for correct repetition of shadow sentences and for words memorization in a foreign language. They asked to 10 English subjects to repeat Corean sentences after 3 successive presentations in 2 audiovisual conditions: one with only the high part of the face visible and another one with the low part of the face. Note that it is not a close shadowing task but a differed shadowing one since subjects repeat sentences after complete audition. The authors evidenced a benefit of the view of the low part of the face (lip, mandible, teeth and tongue) both for production with 100 ms of advance in *RT* and for memorization scores.

Hyper-articulated speech

In a series of work devoted to speech adaptability, Lindblom [6] explained that speech variability is controlled through negotiation between the listener and the speaker. In his hyper-hypo theory, the information carried by the acoustic signal depends inversely on context: if the context is sufficiently rich, the information in the signal can be poor. In the contrary, when the context gives no information, the speaker can hyper-articulate to increase information in the signal. Beautemps et al. [7] quantified the effect of hyper-articulation. In the identification of plosive consonants acoustically degraded, the authors compared the hyper condition with normal speech. They observed that identification scores decreased with the signal to noise ratio but with a better solidity for the hyper condition. In the hyper-articulation production, it was observed on the acoustic signal a clear control of the frequency distribution of the energy at the release instant in addition to a global volume effect [7, 8].

The present work aimed at evaluation of the benefit of vision and hyper-articulation on reaction times in shadowing task of French unvoiced plosive consonants [p, t, k], for sequences with perfect audible acoustic signal.

2. EXPERIMENTAL SETUP

Speech material:

A French speaker uttering a set of 28 /VCVsV/ sequences was audiovisually recorded in quiet environmental conditions and with a 80 dB SPL white noise presented at both ears for a corpus made of the French [a, i, u, y] vowels and [p, t, k] plosives. The noisy environmental condition was used to encourage the speaker to produce an auditory vocal effort (the so-called « Lombard reflex ») with a natural hyper-articulation. For each of the two environmental conditions the speaker was asked to pronounce the sequence and to repeat it with an emphasis on the first consonant. The perfect identification of the consonant from the quiet and hyper-articulated + emphasis conditions was reported in Beautemps et al. [7, 8].

The set of data involved in the shadowing experiment was made of with the 12 audiovisual stimuli of the quiet condition and normal production (so-called normal speech) and with the 12 audiovisual stimuli of the noisy environmental condition with emphasis on the consonant (so-called hyper-articulated speech). In order to dismiss the possibility to predict the consonant from the initial vowel duration (Reisberg et al. [4]), the initial vowel of each sequence was lengthened a multiple of 40 ms (i.e. image rate) up to successively 2, 3 and 4s. A few periods of the acoustic signal at the vowel center was duplicated using a TDPSOLA technique. The image preceding the cut off point was duplicated as to complete the video sequences. We thus obtained a set of 144 [VCVsV] stimuli (12 sequences x 2 speech conditions x 3 durations of the initial vowel x 2 speech modes (audio *vs* audiovisual)).

Finally, the energy of the acoustic signal of stimuli of the same vowel were normalized to the hyper-condition one.

Stimuli presentation

Four subjects with no known hearing damage and visual injuries were submitted to shadowing of [VCVsV] sequences with audio and audiovisual stimuli. They where asked to repeat as early as possible the [CVsV] part as they were listening and uttering the first vowel. For example, in

case of listening [apasa] sequence with the first [a] lengthened up to 2s, 3 or 4s, the subject had to maintain [a] in phonation and to produce [pasa] sequence as soon as he identified the consonant. Stimuli were grouped by vowel and presented randomly in a same session, the information on the vowel being given to the subject. Thus duration of the initial vowel, consonant and speech articulation condition (normal *vs* hyper-articulated) were the unknown parameters.

A monitor placed one meter in front of the subject was used for presentation of the speaker face bottom part (nose, mouth, chin, larynx) in the audiovisual condition. In the audio alone condition, the monitor was switched off. In both conditions, the audio stimuli was presented at the subject through headphones and simultaneously recorded on the stereo first line of an audio DAT tape. Subject response to the stimulus was recorded on the synchronised second line of the same audio DAT tape, thus allowing a further reaction time calculation.

Subjects were preliminary trained with presentation of twelve assorted audiovisual stimuli.

Reaction time calculation

The recorded stimuli and the corresponding audio responses were digitalized in stereo. Reaction Time was derived from the duration between the instant of consonant acoustic closure of the stimulus and the instant of the corresponding acoustic event in the response. The acoustic closure was automatically marked at the point where the energy (integrated through a 10ms window) attained 20% of the energy in the preceding vowel center.

3. RESULTS

Error analysis

Responses for which the consonant was different from the stimulus one or with reaction time superior to 400 ms or containing interruption were considered as errors for the shadowing task and were not considered in reaction time analysis. The average rate error of 6.8 % observed for the responses was much less than the 23 % obtained by Porter & Catellanos [3]. Curiously the audiovisual condition gave much more errors (9 %) than with auditory alone (4 %). The error on the consonant was increased (9 %) in context of the rounded vowels [u] and [y] - context well known for masking visual consonant effect ([9])- in comparison to 3 % for [a] context and 6 % for [i]. For bilabial [p] and dental [t] the error is in average 4 %, in comparison to 12.5 % for the alveolar [k]. Surprisingly, the error attained 9 % in the hyper-articulated speech condition (5 % for the normal speech condition). Subjects did not take benefit of the focus on the consonant. Finally error rates in relation to the duration of the initial vowel attained 10 % in case of 4s against 5 % for 2 and 3s.

Average Reaction time

The average reaction times measured between stimulus onset acoustic closure of the consonant and the corresponding event on the response vary between 227.4 to 314.19 ms. These reaction times are somewhat larger than the 170 to 278 ms ones observed by Porter and Castellanos [3]: our subjects were less trained to the task (12 vs. 30 to 60 stimuli of their training phase). On the other hand we obtained much less identification errors (6.8 % vs. 23 %).

Analysis by subject

For each of the four subjects, we performed a three way ANOVA with 3 factors (consonant factor: [p, t, k]; speech articulation factor: normal and hyper-articulated and presentation: audio and audiovisual). Note that when a stimulus was not correctly repeated (error in the response), the responses were not considered both for the audio and audiovisual conditions.

No significant effect was observed for two subjects. Their average reaction time (RT) was 227.4 ms for CL subject and 296,91 ms for NL subject.

For the slower subject (314.19 ms for reaction time), only the presentation factor was significant ($F(1,57) = 10.19$, $p = 0.0023$), i.e. the audiovisual presentation (287.46 ms) allows faster *RT* than the audio presentation (336.75 ms). Following Reisberg et al. [4] formula, the visual gain was calculated: $\left|\frac{287.46-336.75}{336.75}\right| = 14.6\%$

For subject VA (average *RT* of 244.7 ms), the ANOVA showed the significant effect of presentation ($F(1,57) = 13.54$, $p = 0.0005$) with a significant consonant–presentation interaction $F(2,57) = 4.03$, $p = 0.023$). The *RT* in audiovisual presentation (225.8 ms) are shorter than audio *RT* (260.65 ms). The visual gain was: $\left|\frac{225.8-260.65}{260.65}\right| = 13.4\%$

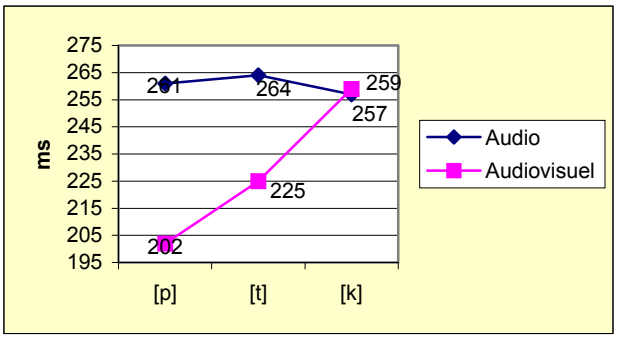

Figure 1: Average *RT* obtained in audiovisual and audio presentations of [p, t, k] consonants for VA subject.

Concerning interaction effect, post-hoc comparisons (Newman-Keuls test) showed an audiovisual benefit for [p] consonant ($p < 0.01$) (Figure 1). We can also observe that

ISBN 1-876346-48-5

the audiovisual *RT* is shorter than audio *RT* for the consonant [t] but this tendency is not significant.

4. CONCLUSION

In this shadowing experiment, contribution of two main factors were evaluated: the hyper vs. normal articulation and the audiovisual vs. audio presentation effect.

No benefit of hyper articulation was obtained. It appeared that articulatory variations reflected in burst energy of the consonants were not exploited by our subjects during the shadowing task. It shall be recalled that the perceptual benefit observed in no shadowing task in beautemps et al. [7, 8] was related to a greater intelligibility robustness in noise. Since our shadowing task was performed with perfectly audible signals, it is possible that these variations could not processed fast enough to be taken into account.

It was observed an advantage of the audiovisual presentation for two of our four subjects with visual gains of 13.4 % and 14.6 %. These gains are relatively high and similar to those of Reisberg et al. [4] (0 to 21.5 %). The originality of these results is that our gain are obtained from reaction time in close shadowing task and not from intelligibility scores. The specific gain obtained for audiovisual [p], by one of our subjects, can be explained by the great visibility of the labial occlusion. Owens & Blazek [9] showed that labial plosives obtained the higher lipreading scores, whatever the vocalic context. We can explain in the same way the tendency to better identify the consonant [t] vs. [k]. However, the effectiveness of audiovisual presentation remains to be confirmed by further experiments with more trained subjects.

More generally, this paradigm of close shadowing seems particularly adapted to understand the cognitive processes implied in the linguistic self-monitoring *vs.* the monitoring of the speech of the other speaker. Many studies begin to show that the execution of an action or the observation of this action executed by another activated the same cortical regions, in frontal and posterior parietal lobes [10, 11]. Thus a common coding of self-action and other action can be postulated. However, recently, a study of Decety [12] exploring the reciprocal imitation in motor tasks, confirmed the role of the *left* inferior parietal in the imitation of the other by the self and the role of the *right* homologous region when one monitors the imitation of oneself by the other. They concluded : "This region may play a fundamental role in agency, i.e. in attributing the source of the action to self or other" ([12] p. 271). Will speech monitoring be as differentiate in hemisphere dominance when our rapid shadowers (hence left) were contrasted with the monitoring of another speaker as a follower of oneself ?

Acknowledgments : This research was supported by a "Jeune équipe" project of the CNRS (French National Research Center).

REFERENCES

[1] W. Marslen-Wilson, "Linguistic structure and speech shadowing at very short latencies," *Nature*, vol. 244, pp. **522-523,** 1973.

[2] R.J. Porter and J.F. Lubker, "Rapid reproduction of vowel-vowel sequences: evidence for a fast and direct acoustic-motoric linkage in speech", Journal of Speech and Hearing Research, 23, pp. 593-602, 1980.

[3] R.J. Porter and Jr. Castellanos, "Speech production measures of speech perception: rapid shadowing of VCV syllables" Journal of the Acoustical Society of America, 57(4), pp. 1349-1356, 1980.

[4] D. Reisberg, J. McLean and A. Goldfield, "Easy to hear but hard to understand: A lip reading advantage with intact auditory stimuli", in B. Dodd and R. Campbell (Eds), Hearing by Eye: The psychology of lip-reading, London, Lawrence Erlbaum Associates, pp. 97–113, 1987.

[5] C. Davis and J. Kim, "Repeating and remembering foreign language words: Does seing help?", In D. Burnham, J. Robert-Ribès and E. Vatikiotis-Bateson (Eds.), AVSP'98: International Conference on Audio-Visual Speech Processing, 121-125, Terrigal, Australia, 4-7 Dec. 1998.

[6] B. Lindblom, "Adaptive variability and absolute constancy in speech signals: Two themes in the quest for phonetic invariance", in PERILUS, 5, pp. 2-20, 1986.

[7] D. Beautemps, P. Borel, S. Manolios, "Hyper-articulated speech", in proceedings *of the 6th European Conference on Speech Communication and Technology* (Budapest), Vol.1, 109-112, 1999.

[8] D. Beautemps. « Parole hyper-articulée : données et analyses acoustiques pour des plosives en français » *Actes des XXIIIèmes Journées d'Etude sur la Parole*, 437-440, 2000.

[9] Owens E. & Blazek B., "Visemes observed by hearing-impaired and normal-hearing adult viewers", *Journal of Speech and Hearing Research*, 28, 381-393, 1985.

[10] J. Decety, N. Grèzes, D. Costes, M. Perani, E. Jeannerod, F. Procyk, F. Grassi and F. Fazio, "Brain activity during observation of actions. Influence of action content and subject's strategy", Brain, 20, 1763-1777, 1997.

[11] G. Rizolatti, L. Fadiga, M. Matelli, V. Bettinardi, E. Paulesu, D. Perani and F. Fazio, "Localization of grasp representations in humans by PET. 1. Observation versus execution", Experimental brain Research, 111, 246-252, 1996.

[12] J. Decety, T. Chaminade, J. Grèzes and A.N. Meltzoff, "A PET exploration of the neural mechanisms involved in reciprocal imitation", NeuroImage, 15, 265-272, 2002.