Effect of being seen on the production of visible speech cues. A pilot study on Lombard speech

Maëva Garnier[1], *Lucie Ménard*[2], *Gabrielle Richard*[2]

[1] Speech and Cognition Department, GIPSA-Lab, UMR CNRS 5216 & Grenoble Universités, France
[2] Laboratoire de phonétique, Université du Québec à Montréal, Canada

maeva.garnier@gipsa-lab.grenoble-inp.fr, menard.lucie@uqam.ca

Abstract

Speech produced in noise (or Lombard speech) is characterized by increased vocal effort, but also by amplified lip gestures. The current study examines whether this enhancement of visible speech cues may be sought by the speaker, even unconsciously, in order to improve his visual intelligibility. One subject played an interactive game in a quiet situation and then in 85dB of cocktail-party noise, for three conditions of interaction: without interaction, in face-to-face interaction, and in a situation of audio interaction only. The audio signal was recorded simultaneously with articulatory movements, using 3D electromagnetic articulography.

The results showed that acoustic modifications of speech in noise were greater when the interlocutor could not see the speaker. Furthermore, tongue movements that are hardly visible were not particularly amplified in noise. Lip movements that are very visible were not more enhanced in noise when the interlocutors could see each other. Actually, they were more enhanced in the situation of audio interaction only. These results support the idea that this speaker did not make use of the visual channel to improve his intelligibility, and that his hyper-articulation was just an indirect correlate of increased vocal effort.

Index Terms: Lombard speech, hyper-articulation, audiovisual intelligibility, multimodality

1. Introduction

On one hand, it is now well known that seeing speech improves its perception, especially when speech is degraded by a noisy background [1]. On the other hand, some studies have shown that speakers adapt their speech production in noisy conditions. This adaptation, also called the « Lombard effect », mainly consists in talking louder and at higher pitch [2-4]. It is also accompanied by other speech modifications, such as increased amplitude and speed of lip articulation [5-7].

This raises the question of whether the hyper-articulation of Lombard speech can be considered as a communicative strategy to improve visual intelligibility.

A first element of answer comes from the fact that not only jaw movements are amplified in noise but also other articulatory movements that are not as related to the increase of vocal intensity, such as lip closure and spreading, and lip protrusion (in some speakers only) [5-6]. A second argument is that the gain in intelligibility from an auditory-only to an audiovisual perception of utterances is weaker in Lombard speech, compared to normal speech [8]. On the contrary, vowels produced in noise are in average more easily recognized in visual-only and audiovisual modalities, as compared to vowels produced in silence [9].

This study aims at bringing a third element of answer, by examining whether, in noise:

- speakers enhance significantly more their visible articulatory movements when their speech partner can see them compared to when the partner can only hear them.
- all the articulatory movements are enhanced similarly, or if the most visible ones (lips) are more enhanced than the others (tongue).

2. Material and Methods

A French Canadian speaker was recorded while speaking in a quiet environment and in a cocktail-party noise of 85 dB [10] played over loudspeakers. Three conditions of interaction were examined: (NI) No Interaction: The speaker read sentences aloud. (AO) Audio Only: The speaker gave instructions to the experimenter who was standing at a writing board placed 2m in front of him and who was turning the back to him. (AV) Audio Visual: The experimenter was standing at the same place as in the AO condition, this time facing the speaker. Seven target-words (/pap/, /pip/, /pup/, /pɛp/, /map/, /tap/, /nap/) were produced in the carrying sentence « le mot ___ me plaît » (*I like the word* ___) and repeated ten times in each condition. The speaker chose freely the order of production of the 70 sentences, so that the experimenter could not predict the target-word.

The Audio signal was recorded with a microphone (Shure SM58) placed 10cm away from the lips, then digitized at a rate of 44.1kHz. Noise was removed from the acoustic signal using the method designed by Ternstrom et al. [11]. The mean intensity and the mean frequency of the first two formants were measured with Praat from the central 50 ms of each target vowel /a/, /i/, /u/ and /ɛ/.

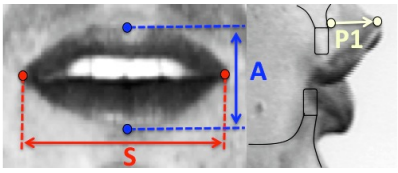

Figure 1. *Articulatory parameters: lip aperture (A), lip spreading (S) and protrusion of the upper lip (P1).*

The 3D movements of the lips, the jaw and the tongue were recorded synchronously with the audio signal, using 3D electromagnetic articulography (Carstens AG 500), at a rate of 200 Hz. The experimental setup is similar to the one used in Thibeault et al. (2011) [12]. Four coils of reference were placed behind each ear and just above the two upper incisors, in order to

consider all the articulatory movements in the fixed referential of the head. One coil was placed just under the lower incisors in order to examine jaw movements. Four other coils were placed on the external contour of the lips in order to measure lip aperture (A), lip spreading (S) and the protrusion of the upper lip (P1) (see Figure 1). The last three coils were placed on the central line of the tongue, approximately 1.5cm, 2.5 and 3.5cm away from the tip of the tongue. The coils attached to the tongue were found to move almost in a plane, with a mean distance of 0.8mm to it. This plane was estimated from the 420 sentences produced during the experiment and was then considered as the sagittal plane in order to analyze the tongue movements. The mean value of each articulatory descriptor was measured on the target a/, /i/, /u/ and /ɛ/, over a 50ms interval that was centered on the local maximum of that descriptor (if there was one), or by default, on the time of maximum jaw aperture.

3. Results

3.1. Acoustic modifications

3.1.1. Vocal intensity

Figure 2 shows the average increase of vowel intensity from the quiet to the noisy situation, for the 3 conditions of speech production. As expected, vocal intensity increased with noise exposure in the 3 conditions. In agreement with our previous study [6], this increase was greater in the interactive situations (AV and AO) than in the non-interactive one (NI). Like Fitzpatrick et al. [13], we also observed that the Lombard effect was affected by the sensory modality of interaction: For the same levels of noise exposure, the increase of vocal intensity was greater when speakers could only hear each other (ΔI_{AO}=16.1±1.8 dB), compared to when they could both hear and see each other (ΔI_{AV}=11.9±1.7 dB).

3.1.2. Formant frequencies

Figure 3 summarizes the acoustic modification of the vowels /a/, /ɛ/, /i/ and /u/ in the F1*F2 plane.

Similar tendencies of vowel modification with noise exposure were observed in the 3 conditions of speech production: the frequency of the first formant increased with noise exposure for the 4 vowels examined. However, F1 increased more for the vowels /i/ and /u/ ($\Delta F1_{i,u}$=131±37 Hz) than for the vowel /a/ ($\Delta F1_a$=80±54 Hz) so that the acoustic contrast in vowel height was rather reduced in Lombard speech. Furthermore, the frequency of the second formant increased with noise exposure for the vowel /u/ ($\Delta F2_u$=164±93 Hz) and tended to decrease for the vowel /i/ ($\Delta F2_i$=−48±34 Hz), so that the acoustic contrast between front and back vowels was also reduced in Lombard speech.

How did the modality of interaction modulate this modification of the vowel system? The shift towards higher F1 frequencies was greater in the AO condition of interaction ($\Delta F1_{AO}$=171±25 Hz), compared to the NI condition ($\Delta F1_{NI}$=98±33 Hz). So was the increase of F2 on the vowel /u/ ($\Delta F2_{AO}$=159 Hz and $\Delta F2_{NI}$=125 Hz). However, the communicative interaction lost its effect when the speakers could see each other: no difference was observed in the modification of F1 and F2 between the NI and the AV conditions.

The audible contrast along the F1 dimension between open and close vowels was almost preserved in the AO interactive condition (-31 Hz) whereas it was more reduced in the AV and NI condition (-51Hz and -70 Hz respectively). On the contrary, the audible contrast along the F2 dimension between front-spread vowels and back-rounded vowels, was altered in the condition of AO interaction (-333Hz) and in the NI condition (-230Hz) whereas it was less affected in the AV condition (-97Hz).

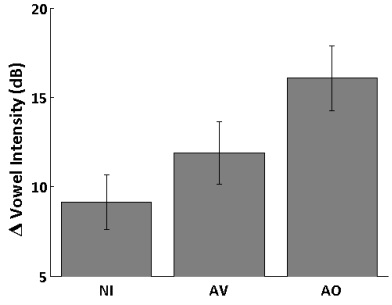

Figure 2: *Increase of mean intensity of vowels with noise exposure, for a non interactive condition of speech production (NI) and two conditions of Audio Only (AO) and Audio Visual (AV) interaction.*

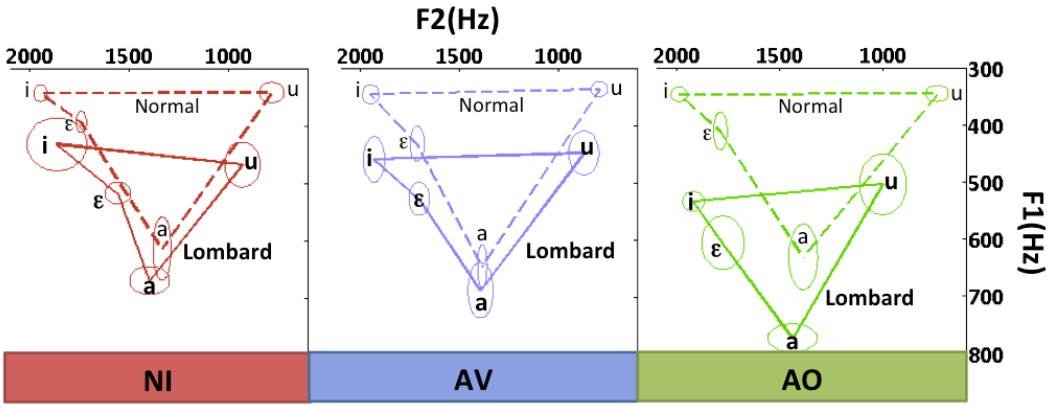

Figure 3: *Modification of the first two formant frequencies of the vowels /a/, /e/, i/ and /u/ between normal and Lombard speech, for a non-interactive condition of speech production (NI) and two conditions of Audio Only (AO) and Audio Visual (AV) interaction.*

3.2. Articulatory modifications

3.2.1. Lip articulation (visible)

Figure 4 shows how the different dimensions of lip articulation were modified with noise exposure for the vowels /a/, /ɛ/, /i/ and /u/ in the 3 conditions of speech production.

Unlike in previous experiments involving non-interactive tasks [6-7], the speaker did not amplify his lip movements in noise, compared to silence, in the NI condition.

A very slight increase of lip aperture was observed for all vowels in the condition of AV interaction (ΔA_{AV}=1.6±1.6 mm). However, in that condition, no enhancement of lip spreading was observed for the spread vowels /ɛ/ and /i/. No clear change in lip spreading and protrusion was observed for the rounded vowel /u/ either. At least, these visible cues were not degraded in Lombard speech, in comparison to normal speech.

The condition of AO interaction showed the greatest modification of lip articulation. The vowels /a/, /ɛ/ and /i/ showed an increase of lip aperture (ΔA_{AO}=5.3±1.8 mm) and lip spreading (ΔS_{AO}=5.5±1.6 mm), and a decrease of lip protrusion ($\Delta P1_{AO}$=-1.9±0.7 mm). For each of these 3 parameters, the greatest modification was observed for /a/, then for /ɛ/ and finally for /i/. Lip articulation did not change for the vowel /u/.

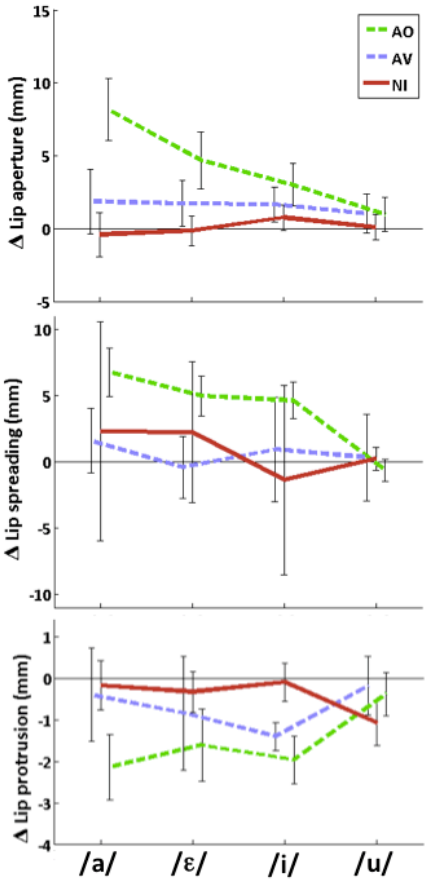

Figure 4: *Modification of visible lip movements with noise exposure, for a non interactive condition of speech production (NI) and two conditions of Audio Only (AO) and Audio Visual (AV) interaction.*

As a consequence, the visible contrast between the 4 vowels was enhanced for Lombard speech in the situation of AO interaction.

3.2.2. Tongue articulation (less visible)

Figure 5 shows how tongue articulation was modified with noise exposure for the vowels /a/, /ɛ/, /i/ and /u/ in the 3 conditions of speech production.

In the NI condition, changes weaker than 1 mm were observed in the tongue position for all the vowels.

In the AV condition, no change in tongue position was observed for the vowels /a/, /ɛ/ and /i/ either, although the jaw was lowered by 2.1±1.7mm in average. However, a displacement of the tongue downwards was observed in Lombard speech for the vowel /u/ ($\Delta Height_{u,AV}$=-5.0±2.2mm for the most forward coil), accompanying a lowering of jaw by 1.1±1.3mm for the vowel [u].

An even greater displacement of the tongue was observed again in the AO interactive condition, this time for the 4 vowels examined ($\Delta Height_{AO}$= -8.7±2.0 mm, -7.2±1.8 mm, -4.0±2.2 mm and -11.1±4.3 mm for respectively /a/, /ɛ/, /i/ and /u/). As an indication, the jaw was lowered by 11.3±2.0 mm, 8.8±2.9 mm, 6.7±1.4 mm and 6.6±3.2 mm for these same respective vowels.

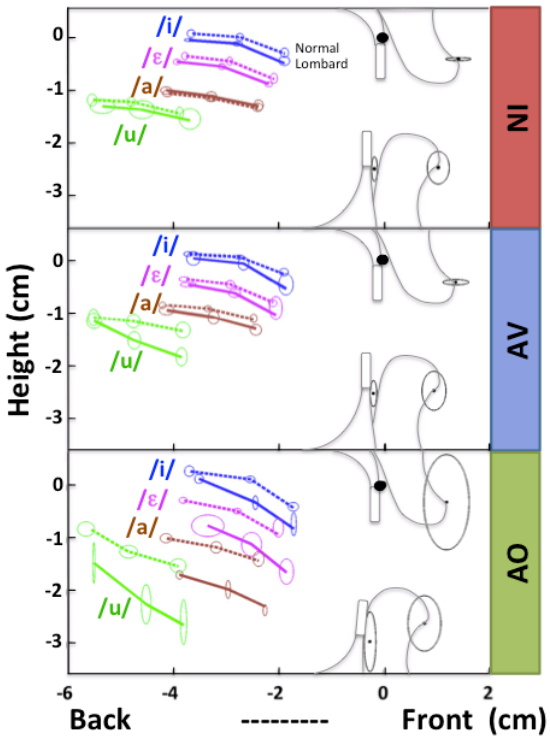

Figure 5: *Modification of the tongue position between normal speech (dotted lines) and Lombard speech (plain lines), for a non interactive condition of speech production (NI) and two conditions of Audio Only (AO) and Audio Visual (AV) interaction. The big black dot indicates the position of the right upper incisor, which defined the origin of the sagittal plane. The charts also represent the movement of the coils attached to the lower incisors, the lower lip and the upper lip, as well as a schematic representation of the teeth and the lips.*

Furthermore, in that AO condition, the lowering of the tongue was accompanied by a shift of the tongue forward for the vowels /ɛ/, /a/ and /u/ ($\Delta Forward_{AO}$=3.0±2.2 mm, 2.0±1.1 mm and 1.9±1.7 mm respectively, in average over the 3 coils).

4. Discussion

The results confirmed previous observations of Lombard speech: the speaker increased vocal intensity, F1 and lip aperture in noise [2-7]. In Lombard speech, the vowel contrast was reduced along both F1 and F2 dimensions [2-5]. The visible contrast in lip aperture, spreading and rounding between the different vowels was enhanced in noise [5]. In addition, this study also brought new information on the modification of tongue movements in noise: the increase in vocal effort and jaw aperture was accompanied by a global rotation downwards of the tongue and by a more forward position of all types of vowels. This observation is consistent with higher values of F2 for the vowels /u/ and /a/ produced in noise. On the contrary, this appears in contradiction with the decreased values of F2 for the vowels /i/ and /ɛ/. Furthermore, the displacement of the tongue is consistent with the lowering of the jaw, so it seems that the modifications of tongue height were directly related to the increase in jaw aperture.

The results confirmed, following [6], that the increase of vocal intensity and lip aperture, from a quiet to a noisy situation, was greater when the speaker interacted with a speech partner (AO and AV conditions), compared to when he only produced sentences on his own (NI condition). However, when it dealt with the modification of formants and other lip and tongue movements, communicative interaction had an effect only in the case of an AO interaction. In other words, noise exposure had a comparable effect on the modification of vowels in the NI condition and in the condition of AV interaction, which was not the case in [6].

How does the modality of interaction affect the Lombard effect? As expected, and in agreement with [13], acoustic modifications (increase of vocal intensity, modifications of F1 and F2) were greater in a condition of AO interaction relative to a condition of AV interaction. However, contrary to our hypothesis, very visible movements such as lip aperture, spreading, closure and protrusion, were not further enhanced in noise in AV interaction than in AO interaction. On the contrary, they were more enhanced in AO interaction, in correlation with the increase of vocal intensity. On the other hand, less visible tongue movements did not seem to be amplified in noise, whatever the modality of interaction. It just seems as if the tongue position followed directly the global increase of jaw aperture in noise, but there did not seem to be any enhanced articulatory contrast in tongue height or along the front/back dimension.

5. Conclusion

The results obtained from this speaker do not support the hypothesis that speakers modulate their production of visible cues in adaptation to the perceptual modalities of interaction. Instead, these results support the idea that all articulatory movements, regardless of their visibility, are enhanced similarly when speaking in noisy conditions, and that this enhancement is primarily related to the increase of intensity. To compensate for the perturbation of intelligibility – which is greater in AO interaction than in AV interaction –, “expanding sonority” (i.e. increasing vocal intensity) appears to be the main strategy of this speaker, instead of expanding the space of vowel realizations. In some extent, such a strategy can be compared to that observed in the production of prosodic focus [14]. As a finer strategy, this speaker did not seem to play on the visual channel to improve their intelligibility. The current investigation of five additional speakers will enable us to determine if these results can be widespread.

6. References

[1] Sumby, H. and Pollack, I. W., "Visual Contribution to Speech Intelligibility in Noise", Journal of the Acoustic Society of America 26: 212-215, 1954.

[2] Junqua, J., "The Lombard reflex and it role on human listener and automatic speech recognizers", Journal of the Acoustic Society of America 93(1): 510-524, 1993.

[3] Van Summers, W., Pisoni, D. B. et al., "Effects of noise on speech production: Acoustic and perceptual analyses", Journal of the Acoustic Society of America 84(3): 917-928, 1988.

[4] Castellanos, A., Benedi, J. M. and Casacuberta, F., "An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect", Speech Communication 20: 23-35, 1996.

[5] Garnier, M., "May speech modifications in noise contribute to enhance audio-visible cues to segment perception?", in Proc. AVSP, Moreton Island, 2008.

[6] Garnier, M., Henrich, N. and Dubois, D., "Influence of Sound Immersion and Communicative Interaction on the Lombard Effect", Journal of Speech, Language and Hearing Research 53(3): 588-608, 2010.

[7] Kim, J., Davis, C. et al., "A visual concomitant of the Lombard reflex", In Proc. AVSP, Vancouver, 2005.

[8] Davis, C., Kim, J. et al., “Lombard speech: Auditory(A), Visual(V) and AV effects”, in Proc. Speech Prosody, Dresden, 2006.

[9] Garnier, M. “Audio, Visual and Audiovisual intelligibility of vowels produced in noise”, submitted to ICSLP 2012.

[10] Zeiliger, J., Serignat, J. F. et al., "BD_Bruit, une base de données de parole de locuteurs soumis à du bruit", in Proc. JEP, Trégastel, 287-290, 1994.

[11] Ternström, S., Sodersten, M. and Bohman, M., "Cancellation of simulated environmental noise as a tool for measuring vocal performance during noise exposure", Journal of Voice, 16(2): 195-206, 2002.

[12] Thiebault, M., Ménard, L., et al., "Articulatory movements during speech adaptation to palatal perturbation", Journal of the Acoustical Society of America, 129(4), 2112-2120, 2011.

[13] Fitzpatrick, M., Kim, J. and Davis, C., "The effect of seeing the interlocutor on speech production in different noise types", in Proc. ICSLP, Florence, 2828-2832, 2011.

[14] Beckman, M. E., Edwards, J., and Fletcher, J., “Prosodic structure and tempo in a sonority model of articulatory dynamics”, In Docherty, G. J. and Ladd, D. R. [Eds], Papers in Laboratory Phonology II: Segment, Gesture, Prosody, 68-86, Cambridge University Press, 1992.