Being looked at in the face attenuates articulatory movements in both sighted and blind speakers

Haruka Saito[a], Paméla Trudeau-Fisette, Camille Vidou, Lucie Ménard[b]

Université du Québec à Montréal
saito.haruka@courrier.uqam.ca[a], menard.lucie@uqam.ca[b]

ABSTRACT

Many studies have shown that looking at the speaker's face improves speech perception, whereas what being looked at does to the speaker's speech production remains unclear. The current study asked both sighted and congenital blind speakers to produce vowels in face-to-face and audio-only conditions (both in a quiet and noisy background) while measuring the tongue and lip movements using electromagnetic articulography. We found that sighted speakers showed attenuated movements of the lips in the face-to-face conditions, while blind speakers showed attenuated movements of the tongue in the same conditions. These results suggest that (a) being looked at in the face may lead to attenuated, rather than enhanced, articulatory movements, and (b) this phenomenon may not only be associated with the speaker's attempt to adjust visual speech cues, but also with other communicative factors.

Keywords: Speech production, audio-visual, electromagnetic articulography, blind speakers

1. INTRODUCTION

Visual information is an integral part of speech communication. Prior literature has consistently found that looking at the interlocutor's face, along with their speech, improves auditory detection [1] and speech intelligibility in noise both for listeners with normal hearing [2] and with hearing impairment [3], as well as enhances learning of foreign sounds [4].

What remains unclear is to what extent speakers are actively aware and make use of visual speech cues to optimize intelligibility. Considering that auditory and visual cues are usually in sync—a clearly pronounced audio tends to be produced by clearly defined articulatory movements, and vice versa—it is possible that the benefits of visual cues are a fortunate by-product of speech production for listeners, but not an active strategy of speakers. A few recent studies, however, suggested that speakers may be capable of adapting their auditory and visual speech cues separately in a face-to-face condition by selectively enhancing visible articulatory movements: speakers showed increased lip movements when performing a face-to-face communicative task in noise, despite that their auditory productions in the same condition were somewhat attenuated (i.e., decreased intensity and/or F0) compared to an audio-only condition [5, 6].

To further investigate whether and how speakers adapt their speech output when they know that visual modality is available for their interlocutor, the current study asked both sighted and congenital blind speakers to produce sentences in both face-to-face and audio-only conditions. Because blind speakers have never been exposed to the visual domain of speech, any changes that they would make in a face-to-face situation are less likely associated with their attempt to adapt visual cues. By comparing them and sighted speakers, we aimed to investigate whether speakers' speech adaptations are more likely to be associated with visual adaptation strategies or with other factors specific to a face-to-face context. In this paper, we focus on articulatory data and will not present acoustic data, which is to be discussed elsewhere [7].

2. METHODS

2.1. Participants

Nine sighted (age M=33.2 years; Male 6, Female 4) and ten congenitally blind (age M=37.6; Male 6, Female 3) native speakers of Canadian French were recruited in Montreal, Canada. All speakers had normal hearing and reported no speech or language impairments. The blind speakers never had any visual perception of light or movement, classified as class 3, 4, or 5 in the International Disease Classification of the World Health Organization (WHO).

2.2. Task and procedures

Participants were asked to produce six Canadian French vowels /a, i, u, y, ɛ, e/ (note that only the first four were used in the current analysis) embedded in a carrier sentence, 'Monsieur /pVp/ est parti' (*Mister /pVp/ is gone*), in four different conditions: audio-only, audio-only with noise, face-to-face, face-to-face with noise. In all conditions, participants were asked to produce the sentences so that an interlocuter (= experimenter) could understand them, with the only difference being that in the audio-only conditions, the interlocuter was standing behind the participant

(about one meter behind their left shoulder) and the participant was explicitly told that the person could not see their face; in the face-to-face conditions, their interlocuter was standing in front of them (about 1.5 meters away) and the participant was explicitly told that the person was looking at them. In addition, in the conditions with noise, the participant wore earbuds to listen to white noise (60 dB) and was informed that their perceiver was listening to the same noise. All participants produced 192 vowel samples in total (6 vowels x 8 repetitions x 4 conditions).

2.3. Data recordings

Articulatory data was obtained using Carstens AG501 (electromagnetic articulography: EMA) with a sampling rate of 250 Hz. Sensors were placed on upper lip (UL), lower lip (LL), lower incisors (JAW), tongue tip (TT), tongue blade (TB) and tongue dorsum (TD), in addition to three reference sensors to track head movements (two mastoids and upper incisors). We also recorded each participant's bite plane using a solid plate with three sensors on it. Simultaneously, acoustic data was recorded via an Audio-Technica microphone (BP892) at 44100 Hz.

2.4. Data analysis

The EMA position data was transformed to account for head movements using Matlab scripts (courtesy of Mark Tiede, Haskins Laboratories) with reference to three reference sensors and the bite plane. The position data was then extracted at the middle point of each target vowel, determined based on landmarks the acoustic signal. To normalize speakers' physiological differences in size, all sensor positions were z-scored within speaker (across all six experimental sensors and across all three dimensions).

Statistical analysis was performed using the lme4 [8] and emmeans [9] packages in R [10].

2.4.1. Tongue between-vowel distances

To examine whether speakers used a wider, or narrower, 'vowel space' of the tongue as a function of face-to-face and/or noise conditions, between-vowel distances were calculated using the following procedure: first, for each participant, the 'center' of each of the three corner vowel categories (/a/, /i/, and /u/) were calculated by averaging all the samples for the vowel category for each of the three dimensions (x, y, z); the Euclidean distance in the 3D space between each sample of the corner vowels and the 'center' of the other two vowel categories were then calculated (ex., if the sample is /a/, the distance between the x, y, z coordinates of that particular sample of /a/ and the 'center' of the speaker's /i/ category was calculated, as well as between that particular /a/ and the 'center' of the speaker's /u/ category); the two distances were then averaged to have one distance for each sample. This way, we could evaluate how far apart each sample of the corner vowels was from the other two corner vowels, with longer distances representing a wider vowel space. This analysis was done for TT, TB and TD sensors separately, although this paper will only present the results for TD.

2.4.2. Lip aperture

Vertical lip aperture, the Euclidian distance in the 3D space between UL and LL, was calculated for the low vowel /a/, assuming that larger lip aperture represents more salient visual cues for the vowel. One sighted participant was excluded from this analysis due to a measurement failure of UL and LL.

2.4.3. Lip protrusion

Lip protrusion for the rounded vowels /u/ and /y/ was measured as the Euclidian distance in the 3D space between LL and JAW. One sighted participant was excluded from this analysis due to a measurement failure of LL.

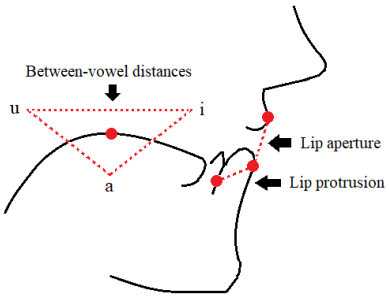

Figure 1: Illustration for EMA measurements

3. RESULTS

3.1. Tongue between-vowel distances

Figure 1 shows the average Euclidian distances between corner vowels /a, i, u/. A linear mixed effects model was built with Face (Audio or Face, sum-coded), Noise (Quiet or Noise, sum-coded), and Group (Sighted or Blind, sum-coded) and a three-way and all possible two-way interactions among the three as fixed effects and random intercepts by Participant and by Vowel. A significant main effect of Face (β=-.008, t=-3.23, p=.001) and an interaction between Face and Group (β=-.012, t=-4.71, p<.001) were found. Tukey's post-hoc test revealed that the blind group showed less tongue movements in the face-to-face condition compared to the audio-only condition

(p<.001 for the quiet, p=.003 for the noisy background). There was no such difference for the sighted group.

Another significant effect found was a main effect of Noise (β=-0.014, t=5.79, p<.001), which suggests that both groups used a larger vowel space when exposed to noise. No significant interactions between Noise and other fixed effects were found.

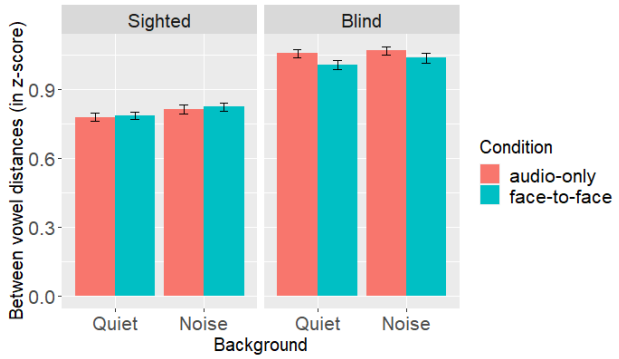

Figure 2: Tongue between-vowel distances of TD for the corner vowels /a, i, u/

3.2. Lip aperture

Figure 3 shows the average vertical lip aperture for /a/ for both groups. A linear mixed model was built using the same fixed and random effects as 3.1. The results show both a significant main effect of Face (β=-.013, t=-2.59, p=.012) and an interaction between Face and Group (β=-.023, t=4.18, p<.001). Tukey's post-hoc test revealed that the sighted group used smaller lip aperture when face-to-face compared to the audio-only condition in the quiet background (p<.001) but not in the noisy background (p=.44). There was no other significant difference with respect to the face-to-face vs. audio-only condition.

As for the effect of Noise, we found a significant main effect of Noise (β=.039, t=-7.22, p<.000) and no interaction with other fixed effects. Again, both groups tended to use larger lip aperture in a noisy background.

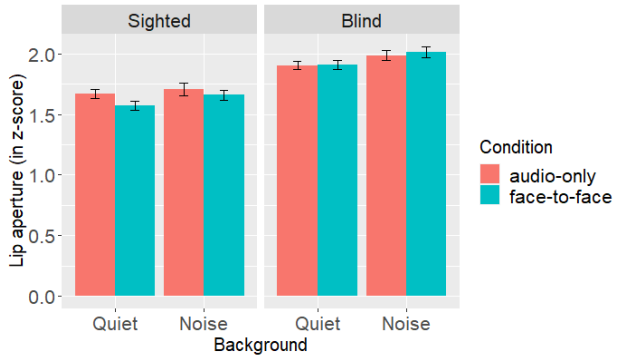

Figure 3: Lip aperture for the low vowel /a/

3.3. Lip protrusion

Figure 4 illustrates the average lip protrusion for /u/ and /y/ for both groups. A linear mixed model was built using the same fixed and random effects as 3.1 and 3.2. Again, we found both a significant main effect of Face (β=-.275, t=-8.19, p<.000) and an interaction between Face and Group (β=.254, t=7.57, p<.000). Tukey's post-hoc test found that, both in the quiet and noisy background, the sighted group showed less lip protrusion in the face-to-face condition compared to the audio-only condition (p<.001 for both the quiet and the noisy background), whereas the same difference was not found for the blind group.

As for Noise, we did not find either main effect of Noise or interaction between Noise and Group, which suggests that participants did not change the extent of lip protrusion as a function of noise.

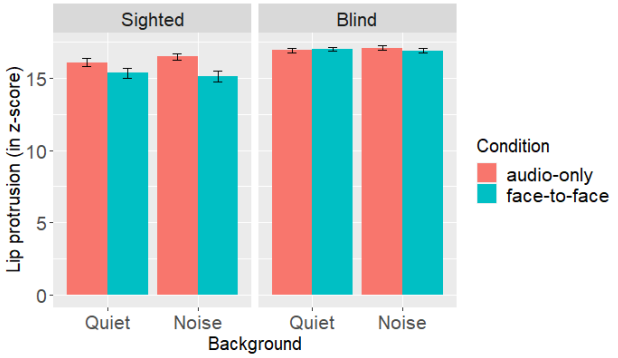

Figure 4: Lip protrusion for the rounded vowels /u, y/

4. DISCUSSION

The current study investigated how sighted and blind speakers may adapt their articulatory movements when they were aware that their interlocuter could utilize visual speech cues. We found that sighted speakers tended to use smaller, rather than larger, visual cues (lip aperture and protrusion) in face-to-face conditions. Blind speakers were also affected by face-to-face conditions: they showed less distinct vowel contrasts in their tongue movements.

Previous studies found that sighted speakers tended to enhance visible articulatory movements in face-to-face conditions and argued that (at least some) speakers could make active use of visual modality to enhance their speech intelligibility [5, 6]. Our findings were opposite: our sighted speakers attenuated the visual cues precisely when they knew their interlocuter could watch their face. Adding one more (i.e., visual) modality to speech communication means that listeners would have an additional source of information, which would in turn decrease the informational burden on one modality. Thus, it seems plausible if speakers make less effort in a face-to-face

situation than in an audio-only situation, since they are probably more confident that their speech would be understood by their listeners due to additional visual information.

In previous studies, speakers did show attenuated speech in the acoustic domain, but enhanced in the visual domain [5, 6] in a face-to-face situation. We speculate that this was because the task used in these studies, where participants must convey minimal word pairs to the interlocutor under the pressure of clarification in a noisy background, had a higher communicative load than the current study. These studies have shown that speakers *can* enhance visual speech cues under a certain condition specifically designed to make speakers rely more on visual modality. We presume that, without that kind of specific condition, simply being looked at in the face more likely attenuates, rather than enhances, speech production in the articulatory aspect as well.

Furthermore, the finding that blind speakers, despite their complete inexperience in the visual domain of speech, also attenuated their speech production in face-to-face conditions led us to consider *why* speakers may make less effort when being looked at during speech production. In the above discussion, we speculated that it was because sighted speakers were aware that additional visual modality would help them understood by their interlocuter, but this does not entirely apply to blind speakers. For this group of speakers, a face-to-face situation should be less defined by the presence of visual modality, but more by other communicative factors. For example, knowing that they have the interlocuter's gaze signals that they most likely have the person's full attention, which would decrease the need for enhanced speech. Another factor in a face-to-face situation might be the need to compose facial expressions (and blind individuals are known to share some universal, albeit different in some contexts, facial expressions as sighted individuals [11]), which would constrain certain articulatory movements as speakers need to use part of muscles to control facial features. These factors are not directly associated with the speaker's attempt to adjust visual speech cues, and if any of these might be part of the motivation for blind speakers to adapt their speech production, why are they not for sighted speakers?

In the current study, sighted and blind speakers differed in terms of which articulator they adapted: sighted speakers decreased visible (lip) movements while blind speakers less visible (tongue) movements. One therefore can argue that the adaptation strategy used by the two groups of speakers were fundamentally different, the former attempting to control visible speech cues and the latter motivated by other communicative factors. However, it has been reported that sighted and blind speakers use different articulators to produce similar acoustic results: the sighted use more lip contrasts and the blind use more tongue contrasts when they produce contrastive or clear speech [12, 13]. Thus, we speculate that the two groups in the current study may have aimed for similar acoustic results (i.e., attenuated speech) due to similar reasons—be it due to having the interlocutor's attention, composing facial expressions, or other factors specific to the face-to-face condition—and even sighted speakers may not have actively attempted to adapt visual speech cues in the face-to-face conditions.

Finally, although the above discussion describes that speakers "attenuated" articulatory movements compared to the baseline of the audio-only condition, one can argue that it was the other way round: speakers may have made an extra effort in audio-only situations considering, in reality, speech production should occur face-to-face more often. We chose to use the audio-only condition as a reference because it is arguably the most common way to collect speech production data in this field of study (i.e., experimenters either do not look at the speaker or do not report where their gaze was during a recording session). This leads to the question of whether non face-to-face conditions should be the standard for data collection, and this study suggested that the awareness of gaze does affect speech production.

5. CONCLUSION

Visual speech cues are a significant source of information for listeners. This, however, does not necessarily mean that speakers enhance their speech cues whenever visual modality is available. The current study found that sighted speakers tended to attenuate, rather than enhance, visible articulatory movements in face-to-face situations. Considering that we observed attenuated articulatory movements in congenital blind speakers as well (albeit it was in an invisible articulator), we assume that this attenuating effect may be more related to general factors in face-to-face situations, rather than specifically to the attempt by sighted speakers to actively control visual cues.

6. REFERENCES

[1] Grant, KW., Seitz, PF, 2000. The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. 108(3), 1197–1208.

[2] Schwartz, JL., Berthommier, F., Savariaux, C. 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93(2), B69–B78.

[3] Bergeson, TR., Pisoni, DB., Davis, RAO. 2003. A Longitudinal Study of Audiovisual Speech Perception by Children with Hearing Loss Who have Cochlear Implants. Volta Rev 103(4). 347–370.
[4] Hazan, V., Sennema, A., Iba, M., Faulkner, A. 2005. Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. Speech Communication 47(3), 360–378.
[5] Fitzpatrick, M., Kim, J. Davis, C. 2015. The effect of seeing the interlocutor on auditory and visual speech production in noise. Speech Communication 74, 37–51.
[6] Garnier, M., Ménard, L., Alexandre, B. 2018. Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? J. Acoust. Soc. Am. 144(2), 1059–1074.
[7] Trudeau-Fisette, P., Vidou, C., Uribe, C., Ménard, L., in Preparation. Acoustic impacts of congenital blindness on speech production strategies linked to Lombard speech and audiovisual interaction.
[8] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-23. Retrieved from http://CRAN.R-project.org/package=lme4
[9] Length, RV. 2021. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.2. Retrieved from https://CRAN.R-project.org/package=emmeans
[10] R Core Team. 2015. R: A Language and Environment for Statistical Computing (Version 4.0.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/
[11] Valente, D., Theurel, A., Gentaz, E. 2018. The role of visual experience in the production of emotional facial expressions by blind people: a review. Psychonomic Bulletin & Review 25, 483–497.
[12] Ménard, L., Leclerc, A., Tiede, M. 2014. “Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults,” J. Speech Lang. Hear. Res. 57, 793–804.
[13] Ménard, L., Trudeau-Fisette, P., Côté, D., Turgeon, C. 2016. Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind speakers,” PloS One 11, e0160088.