Orofacial somatosensory inputs enhance speech intelligibility in noisy environments

Rintaro Ogane[1], Jean-Luc Schwartz[1], Takayuki Ito[1,2]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
[2] Haskins Laboratories, CT, USA
[*] Institute of Engineering Univ. Grenoble Alpes
{rintaro.ogane, jean-luc.schwartz, takayuki.ito}@gipsa-lab.grenoble-inp.fr

Abstract

Noise in speech communication reduces intelligibility and makes it more difficult for the listener to detect the talker's utterances. In such noisy environments, other sensory inputs coming with auditory inputs can help to increase speech sound intelligibility. For example, seeing the speaker's facial movements aids the perception of speech sounds in noise (Grant and Seitz, 2000; Kim and Davis, 2004; Sumby and Pollack, 1954). Recent findings have demonstrated that somatosensory information associated with facial skin deformation also intervenes in speech perception (Ito et al., 2009; Ogane et al., 2019, 2020). While the effect of somatosensory stimulation was only assessed in quiet environments, somatosensory inputs might also increase the intelligibility of speech sounds in noisy environments. The current experiment examined whether orofacial somatosensory inputs facilitate the detection of speech sounds in noise. We carried out a test to evaluate the detection threshold of speech sounds in noise and examined whether this threshold was decreased when the sound was accompanied with somatosensory stimulation. Moreover, we examined whether somatosensory stimulation provides just a temporal clue for accurate detection or includes more specific articulatory information related to auditory stimulation. For this aim, we compared different types of auditory stimuli, varying in terms of articulatory compatibility with the somatosensory stimulation.

Twenty-eight native French speakers participated in the experiment. We focused on two French speech sounds, /pa/ and /py/ respectively associated with vertical (jaw opening) and horizontal (lip rounding) articulatory gestures. Both stimuli were recorded by a male native French speaker. The intensity levels for both stimuli were adjusted to be equal. Each speech sound was tested in a separate group. The participants were randomly assigned to either of the two groups. During the test, a 1-s white noise sound was presented twice in sequence with an inter-stimulus interval of 250 ms. The speech stimulus (/pa/ or /py/ depending on the group) was embedded inside either of the two noise stimuli. Participants were asked to identify which noise interval included the speech sound by pressing a key as quickly as possible. The amplitude of the noise was fixed at 80 dB SPL. We tested 8 signal-to-noise ratio levels by modifying the amplitude of the target speech sound from -8 dB to -15 dB for /pa/, and from -10 dB to -17 dB for /py/ (values selected after a pilot experiment). The onset of the speech sound in the corresponding noise interval was randomly set to either 200 or 600 ms after noise onset. The auditory stimulation was presented through headphones. Somatosensory stimulation associated with facial skin deformation was produced using a robotic device in a vertical direction with a 6 Hz half-sinusoidal pattern providing a 167 ms stimulation duration. The peak timing of the somatosensory stimulation was adjusted at the peak amplitude of the target speech sound. The stimulus was applied in both noise intervals whatever the interval containing the speech sound to detect. We tested two experimental conditions: a pure auditory condition and a condition with somatosensory stimulation. These two conditions were alternated every 8 trials. In total, 320 stimuli (eight SNR levels × 20 occurrences per SNR level × two experimental conditions) were presented in a pseudo-randomized order. For data analysis, the percentage of correct detection response was obtained at each SNR level. We compared the average correct detection score across SNR levels between the two experimental conditions. One-way ANOVA with repeated-measures was applied to each participant group separately since the two groups displayed clearly different variances.

For /pa/, there was a significant 3% difference between auditory alone and auditory-somatosensory conditions (0.73 ± 0.01 vs. 0.76 ± 0.01 on average ± standard error, $F(1, 13) = 5.44$, $p < 0.04$). On

contrary, there was no difference between them for /py/ (0.63 ± 0.02 in both cases, $F(1, 13) = 0.07$, $p > 0.8$). The results indicate that somatosensory inputs may indeed increase speech intelligibility in noise. This is consistent with audio-visual processing (Bernstein et al., 2004; Schwartz et al., 2004; Sumby and Pollack, 1954) and audio-tactile integration (Derrick et al., 2019) showing that additional sensory inputs may increase intelligibility of speech sounds in noisy environment. Importantly, it appears that the effect varies depending on speech utterances, being displayed only when the somatosensory stimulation is compatible with the articulatory nature of the corresponding speech sound. These results support the idea that somatosensory information does intervene in the speech perception process, in a way related to its underlying articulatory/motor content.

Acknowledgements
This work was supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant Agreement no. 339152) and the National Institute on Deafness and Other Communication Disorders R01DC017439. We thank Gurvan Quiniou for data collection and analysis, and Silvain Gerber for statistical analysis. We also thank Coriandre Vilain for his technical support.

References

Bernstein, L. E., Auer, E. T., and Takayanagi, S. (**2004**). "Auditory speech detection in noise enhanced by lipreading," Speech Commun., **44**, 5–18. doi:10.1016/j.specom.2004.10.011

Derrick, D., Hansmann, D., and Theys, C. (**2019**). "Tri-modal speech: Audio-visual-tactile integration in speech perception," J. Acoust. Soc. Am., **146**, 3495–3504. doi:10.1121/1.5134064

Grant, K. W., and Seitz, P.-F. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am., **108**, 1197. doi:10.1121/1.1288668

Ito, T., Tiede, M., and Ostry, D. J. (**2009**). "Somatosensory function in speech perception," Proc. Natl. Acad. Sci., **106**, 1245–1248. doi:10.1073/pnas.0810063106

Kim, J., and Davis, C. (**2004**). "Investigating the audio-visual speech detection advantage," Speech Commun., **44**, 19–30. doi:10.1016/j.specom.2004.09.008

Ogane, R., Schwartz, J.-L., and Ito, T. (**2019**). "Orofacial Somatosensory Effects for the Word Segmentation Judgement," Proc. Int. Congr. Phonetic Sci. 2019, Melbourne, Australia.

Ogane, R., Schwartz, J.-L., and Ito, T. (**2020**). "Orofacial somatosensory inputs modulate word segmentation in lexical decision," Cognition, **197**, 104163. doi:10.1016/j.cognition.2019.104163

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (**2004**). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," Cognition, **93**, 69–78. doi:10.1016/j.cognition.2004.01.006

Sumby, W. H., and Pollack, I. (**1954**). "Visual Contribution to Speech Intelligibility in Noise," J. Acoust. Soc. Am., **26**, 212–215. doi:10.1121/1.1907309