Orofacial somatosensory inputs improve speech sound detection in noisy environments

Rintaro Ogane, Jean-Luc Schwartz, Takayuki Ito

Orofacial somatosensory inputs improve speech sound detection in noisy environments

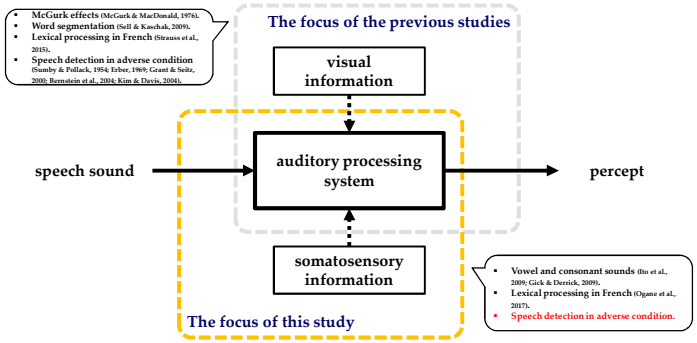
Rintaro Ogane[1], Jean-Luc Schwartz[1], Takayuki Ito[1,2]

[1] GIPSA-lab, CNRS, Univ. Grenoble Alpes, Grenoble INP, Grenoble, France, [2] Haskins Laboratories, New Haven, USA

Contact
rintaro.ogane@gipsa-lab.grenoble-inp.fr

Introduction

- Speech perception is an interactive process with multiple modes and probably some perceptuo(multisensory)-motor connections (Schwartz et. al., 2012).

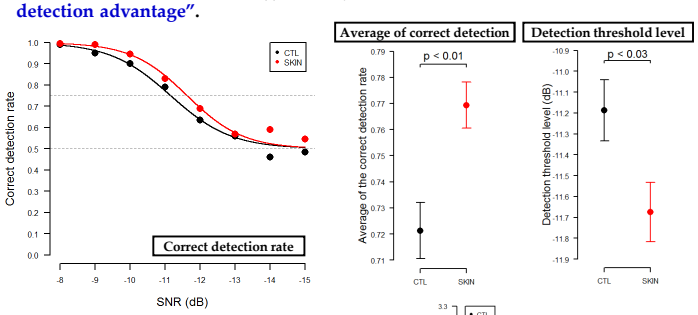


- The aim of this study.
  - To investigate a functional role of somatosensory information on the perception of speech.
- The hypothesis of the current experiment.
  - Somatosensory information associated with facial skin deformation facilitates correct detection of speech sounds in low signal to noise ratio (SNR) environment.

Results

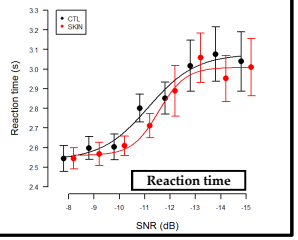
Correct detection rate & detection threshold level

- The correct detection rate for the SKIN condition was significantly greater than for the CTL condition ($t_{(9)} = 3.83$, $p < 0.01$).
- The detection threshold level for the SKIN condition was significantly smaller than for the CTL condition ($t_{(9)} = -2.29$, $p < 0.03$). → "audio-somatosensory detection advantage".



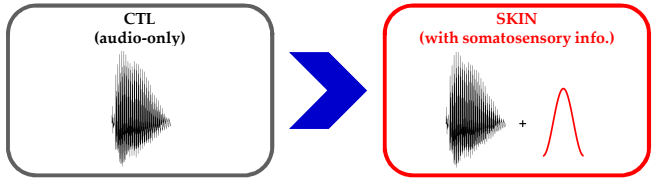
Reaction time

- In both conditions, the reaction time at lower SNRs was longer than at higher SNRs ($F_{(7, 63)} = 14.53$, $p < 0.01$).
- There was no significant difference in the experiment condition ($F_{(1, 9)} = 1.07$, $p > 0.3$), nor an interaction effect ($F_{(7, 63)} = 1.04$, $p > 0.4$).



Discussion

- The detection threshold level for CTL and SKIN conditions:



This result is consistent with the audio-visual speech perception literature.
- ❑ The speaker's facial movements enhance the perception of speech (Sumby & Pollack, 1954).
- ❑ Lipreading in noise enhances the detection of speech (Bernstein et. al., 2004, Schwartz et. al., 2004, Tjan et. al., 2014).

The reduction of detection threshold level → 0.49 dB in average.
↕
cf. audio-visual speech perception.
- ❑ The mean threshold level was lowered about 3 dB (Bernstein et. al., 2004), 1.56 dB (Grant & Seitz, 2000).

This suggests that somatosensory information also provides a rich source to detect the embedded target utterance in noisy environments as well as visual information.

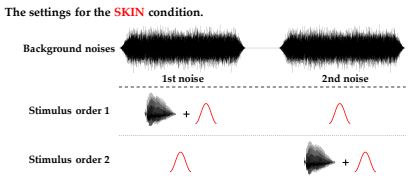
- The relationship between the timings of somatosensory stimulation and output of the target speech sound.

Presented stimulus is "Matched" (congruent)	Presented stimulus is "Unmatched" (incongruent)
↓	↓
The detection threshold became lower compared to "unmatched" or audio-only cond. (Grant & Seitz, 2000).	It does not facilitate the perception of speech (Bicevskis et. al., 2016; Magnotti et. al., 2017; Venezia et. al., 2016).

- ❑ Our result is consistent with the one in "Matched" condition.
- ❑ This suggests that timing of stimulus is important to detect the speech utterance in noise, and we expect that the facilitation would not arise in the unmatched condition.

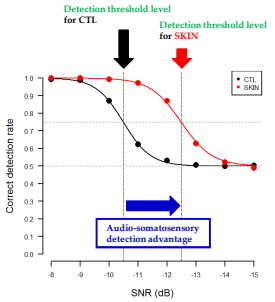
Methods

- We carried out a speech detection test in noise (10 participants) under 2 conditions, (1) with somatosensory information condition (SKIN) and (2) no somatosensory information (audio-only) condition (CTL).
- The subject task was to identify which noise period includes the target utterance.
- Target utterance is /pa/ recorded by a native French speaker.
- 8 SNR levels (from -8 to -15 dB) between target utterance and background noises (white noises) were tested.
- 6 Hz of a sinusoidal pattern was used as the skin stretch stimulation to mimic the speaking motion.



Experimental setup

- 320 trails in total (2 experimental conditions * 8 SNR levels * 20 responses in each SNR level).
- Noise level were set at 80 dB of SPL.
- Data analysis.
  1. Correct detection rate: the percentage of correct detection in each SNR level.
     - The psychometric functions (red and black curves) were estimated to calculate the detection threshold level.
     - Detection threshold level (vertical lines) was defined as a point at 75 % correct detection in the estimated psychometric function.
  2. Reaction time: the duration between the beginning of audio stimulus presentation and key button press.



A representative example of estimated psychometric functions and detection threshold levels.

Summary

- Somatosensory inputs facilitated the correct detection of speech sounds in low SNR environments compared to audio-only condition.
- This result suggests that somatosensory information can intervene in the speech detection process in noisy environments.

Acknowledgments

- This research was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s").
- We wish to thank Gurvan Quiniou for his help in data collection and analysis.