Interspeech 2022
18-22 September 2022, Incheon, Korea

Orofacial somatosensory inputs in speech perceptual training modulate speech production

Monica Ashokumar[1], *Jean-Luc Schwartz*[1], *Takayuki Ito*[1,2]

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

monica.ashokumar@gipsa-lab.grenoble-inp.fr, jean-luc.schwartz@gipsa-lab.grenoble-inp.fr, takyuki.ito@gipsa-lab.grenoble-inp.fr

Abstract

Somatosensory inputs are important to acquire or learn precise control of movement [1]. In the case of speech, receiving somatosensory inputs together with corresponding speech sounds may be a key to formulate or calibrate the speech production system [2]. We here examined whether speech production can be modulated by perceptual training with repetitive exposure to paired auditory-somatosensory stimulation in the absence of actual production of the sound. We carried out a perceptual training using a vowel identification task with /e/-/ø/ continuum. The speech sounds were accompanied with somatosensory stimulation, in which a facial skin-stretch was applied in the backward direction. The vowels /e/ and /ø/ were recorded prior to and following the training and the first three formants were compared. Results showed that the third formant of /e/ was increased following the training, and the rest of formant was not changed. Since the current somatosensory stimulation was related to the articulatory movement for the production of /e/ (lip-spreading), repetitive exposure to somatosensory stimulation in addition to the sound may specifically change the articulatory behavior for the production of /e/. The results suggest that perceptual training with specific pairs of auditory-somatosensory inputs can be important to formulate production mechanisms.

Index Terms: somatosensory system, perception-production link, perceptual training, speech production

1. Introduction

Speech production is a complex mechanism involving the integration of sensory and motor processing for fine tuning of motor control. Based on feedback control mechanisms, receiving a correct auditory - somatosensory pair is important to acquire the target speech output [2]. Previous studies showed that participants adjust their motor behavior in response to an experimentally applied alteration of either the auditory or the somatosensory input [3], [4]. Lametti et al. [5] also showed simultaneous adaptation to both auditory and somatosensory alteration. Interestingly, sensorimotor adaptation to speech motor training also changes speech perception. Indeed, Nasir and Ostry [6] showed that adapting to different external environments during production (with the setting developed in [4]) also changes the vowel category boundary. A study of adaptation to altered auditory feedback similar to [3] also changes category boundary of fricative consonant [7]. Such motor learning effect in perception is likely due to the functional link between speech perception and speech production. In return, this functional link might also induce an effect of perceptual training on speech production.

Interaction between production and perception in speech learning is prominent in language acquisition with repetitive exposure to speech sounds during development [8]. Changes in speech production due to repetitive exposure to speech sound is also seen in adults [9]–[11]. Cooper and Lauritsen [9] showed that repetitive listening of voiceless consonants led to changes in production with a decreased voice onset time for the corresponding consonants. Lametti and colleagues [12] showed that altering the perceptual vowel boundary due to perceptual training affected the amplitude of speech motor learning in response to altered auditory feedback. Listening to the correct pronunciation is also the main strategy in second language acquisition [13]. Recent studies demonstrated that visually received movement information facilitates the acquisition or improvement of production performance [14], [15], indicating that receiving movement-related information in conjunction with corresponding speech sounds, may be important for speech learning.

Somatosensory inputs directly receive articulatory movement information. Externally applying facial skin deformation can be an effective experimental tool to provide movement-related information both to the speech production and perception systems [16], [17]. Ohashi and Ito [18] specifically demonstrated that somatosensory inputs on their own can contribute to the recalibration of perception. Perceptual training with paired auditory-somatosensory stimulation in absence of actual production can also change speech perception [19]. These studies suggest that repetitive exposure to somatosensory inputs in conjunction with speech sounds can contribute to modification or recalibration of the speech representation in production and perception.

The current study examined whether speech perceptual training with orofacial somatosensory stimulation changes the speech production performance. We compared the speech production performance prior to vs. after perceptual training. A vowel identification task using the /e/-/ø/ continuum was applied for speech perceptual training. In a similar task, Trudeau-Fisette et al. [20] have showed that, when orofacial somatosensory stimulation associated with backward skin stretch was applied, the participants' perception was changed to perceive the vowel more as /e/. This can be attributed to the assumption that backward facial skin-stretch can be associated with the articulatory movement of the lip-spreading vowel /e/ and hence causes ambiguous stimuli in the /e/-/ø/ continuum to be more perceived as the vowel /e/. Given this finding, we expected that repetitive exposure of backward somatosensory stimulation paired with auditory stimulation within an /e/-/ø/ continuum might change the production of those vowels as a

10.21437/Interspeech.2022-10993

result of adaptation to paired auditory-somatosensory stimulation. Assessing this expectation is the topic of the present paper.

2. Methods

2.1 Participants

Twenty-eight native French speakers (age range 18-41, 20 female) participated in this study. Participants reported no neurological deficits, hearing or speech disorders. The participants signed the consent form approved by the local ethical committee of the Université Grenoble Alpes [Comité d'Ethique pour la Recherche, Grenoble Alpes (CERGA-Avis-2021-8)]. Half participants (14) were assigned to the target condition with somatosensory stimulation in the training (SOMA group, see 2.5 Experimental Procedure). The other half of the participants were assigned to the control condition (CTRL group, no somatosensory stimulation).

2.2 Speech Task

We focused on two front mid-high vowels, acoustically and articulatory close and differing mainly in lip action, with the lip-spread vowel /e/ and lip-rounded vowel /ø/. Acoustically these two vowels mainly differ in the second formant and to a lesser extent in the third formant, mostly due to the labial contrast with possibly also a slight difference in tongue articulatory anterior-posterior position [21], [22]. We applied the /e/ to /ø/ continuum in our perceptual training based on a study by Trudeau-Fisette et al. [20] that orofacial somatosensory stimulation associated with backward skin stretch reliably changes the participants' perception in this continuum. For the speech production task, we tested the utterance of the French words 'dé' [/de/, "dice" in English] and 'deux' [/dø/, "two" in English].

2.3 Auditory Stimuli

The perceptual training consisted of a vowel identification task using an /e/-/ø/ continuum. Eight stimuli in the continuum were synthesized using the procedure in Menard and Boë [23], in which the second, third and fourth formant frequencies (F2, F3 and F4) were equally shifted in Hz in consecutive stimuli keeping the first and fifth formants (F1 and F5) constant. Details are described in Trudeau-Fisette et al. [20]. These auditory stimuli were presented through a loudspeaker, which was set in front of the participant.

2.4 Somatosensory Stimulation

Somatosensory stimulation associated with the facial skin-stretch was produced using a small robotic device (Phantom 1.0, 3D Systems). Figure 1B displays the experimental setup. Stimulation is based on the assumption that orofacial skin receptors can provide kinesthetic information on speech articulatory movements [17], [24]. Plastic tabs connected to the robotic device through wires were attached to the lateral side of the mouth with double-sided tapes. Based on the previous study [20], we used sinusoidal pattern with 4N of peak force. The onset of stimulation was set a 90 ms lead relative to the onset of auditory stimuli.

2.5 Experimental Procedure

We carried out a perception training paradigm based on the vowel identification task and compared the production performance of corresponding speech sounds before and after the perceptual training (Figure 1A). The experiment began with a baseline speech production task as the pre-test. Ten utterances of the target words were recorded in a picture naming task, in which the pictures depicting 'dé' and 'deux' were presented in a pseudo-random order. This was followed by the perceptual training phase with vowel identification with (target group) or without (control group) somatosensory stimulation. Auditory stimuli were presented in a pseudo-random order. In each trial, the participants were asked to identify if the vowel they heard was /e/ or /ø/. 60 blocks of 8 trials (480 trials in total) were carried out. After this training phase, the speech production task, same as in the pre-test, was carried out again in a post-test phase. In the control condition for the control group, the procedure included the setup of the robot, but no somatosensory stimulation in the training phase.

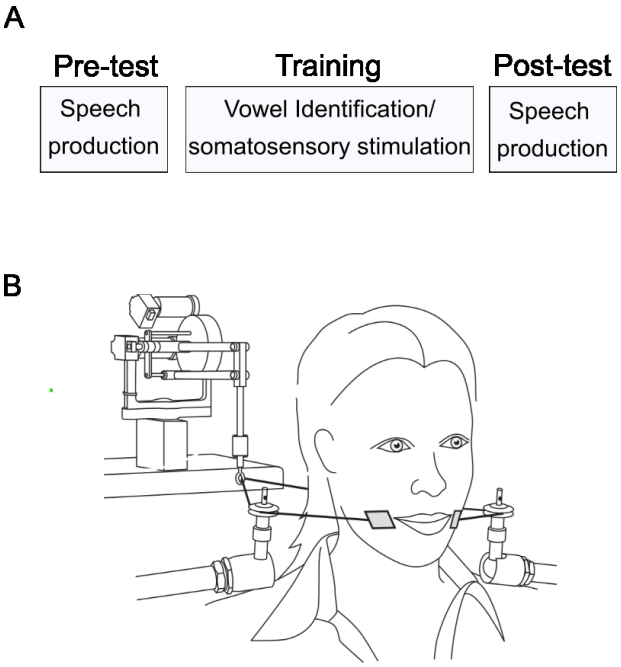

Figure 1: *(A) Experimental protocol (B) Experimental setup for the vowel identification task with somatosensory stimulation. Reproduced with permission from (Ito and Ostry [17]).*

2.6 Statistical Analysis

For each utterance, vowel production periods were manually detected, and then F1, F2 and F3 were extracted using PRAAT [25] by averaging over a 40ms window set around the middle of the vowel production period. To compare the speech production performance in the pre- and the post-test, the difference between the mean values of F1, F2 and F3 for the 10 utterances of each vowel were finally evaluated.

One sample t-test was applied in each formant and group separately to examine whether the difference between pre- and post-tests was reliably different from zero. A one-way ANOVA was also applied to examine whether differences between pre- and post-test were also different between groups (SOMA and CTRL) in each formant.

3. Results

Figure 2 shows representative examples of the spectrogram of the recorded words in pre- and post-test. As shown in this representative example, an increase in the F3 for the vowel /e/ can be observed in the post-test whereas the other formants for both vowels are almost similar in the pre-test and post-test.

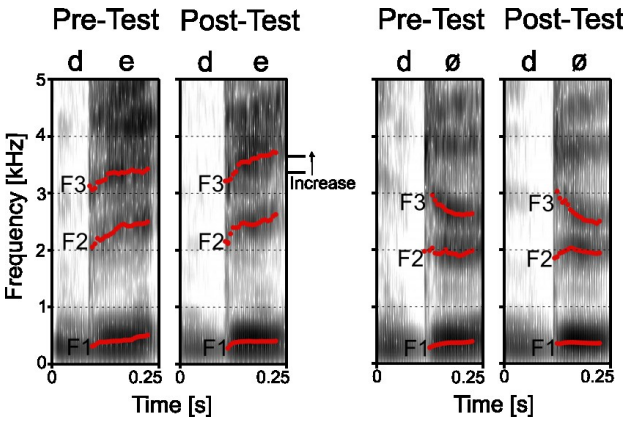

Figure 2: *Representative example of spectrogram of the words 'dé' and 'deux' in pre- and post-test.*

Figure 3 shows the difference between pre- and post-test averaged across subjects. In the production of /e/ (left three panels in Figure 3), we found a reliable change of F3 in the somatosensory group. Separate one-sampled t-tests showed that the change in F3 was reliably different from zero in the experimental group [$t(13) = 2.44$, $p<0.05$], but not in the control group [$t(13) = 0.29$, $p = 0.77$]. One-way ANOVA showed a marginal difference between groups [$F(1,26) = 3.15$, $p = 0.087$]. In contrast, the change in F1 was not significantly different from zero in the experiment group [$t(13) = 0.14$, $p = 0.88$], but marginally different in the control group [$t(13) = 1.81$, $p = 0.09$]. One-way ANOVA showed no significant difference between groups [$F(1,26) = 0.84$, $p =0.36$]. The change in F2 was not significantly different from zero in both the somatosensory group [$t(13) = -1.7$, $p = 0.1$] and control group [$t(13) = -0.57$, $p = 0.57$]. One-way ANOVA showed no significant different between groups [$F(1,26) = 1.51$, $p = 0.22$]. In the production of /ø/ (right three panels in Figure 3), we did not find any reliable change in all formant measures. The change in F1 was not significantly different from zero in both the somatosensory group [$t(13) = -0.62$, $p = 0.54$] and control group [$t(13) = 1.35$, $p = 0.19$]. One-way ANOVA showed no significant difference between groups [$F(1,26) = 1.72$, $p = 0.2$]. The change in F2 was also not significantly different from zero in both the somatosensory group [$t(13) = -1.7$, $p = 0.11$] and the control group [$t(13) = -0.87$, $p = 0.39$]. One-way ANOVA showed no significant difference between groups [$F(1,26) = 0.51$, $p = 0.48$]. There was no significant difference in F3 in both the somatosensory group [$t(13) = 0.72$, $p = 0.47$] and the control group [$t(13) = 0.61$, $p = 0.55$]. One-way ANOVA showed no significant difference for F3 [$F(1,26) = 0.006$, $p = 0.94$].

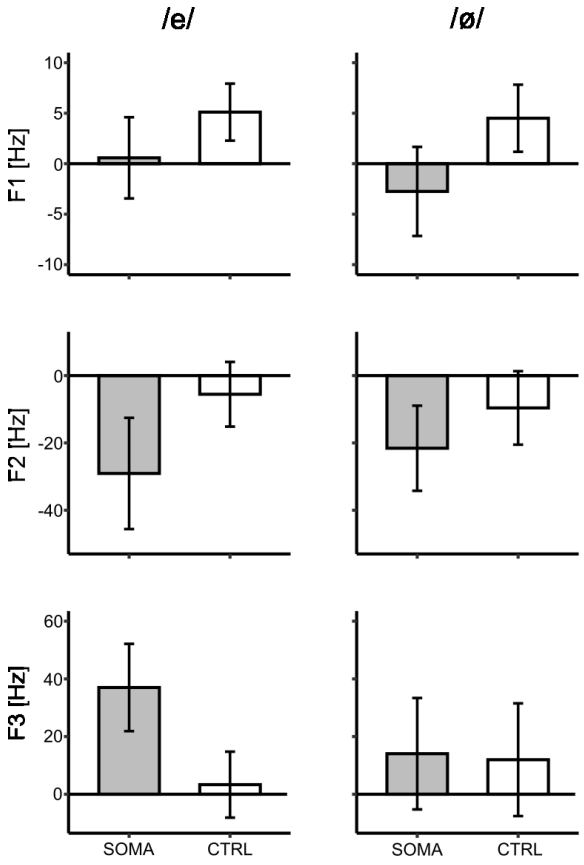

Figure 3: *Difference in F1, F2 and F3 for the vowels /e/ and /ø/ between pre- and post-test. Grey bars represent the somatosensory group (SOMA) and white bars represent the control group (CTRL). Error bars show the standard error across participants.*

4. Discussion and Conclusion

The current study investigated the effect of orofacial somatosensory inputs in speech perceptual training on speech production mechanisms. The speech production performance was evaluated based on the difference in F1, F2 and F3 between the pre- and post-test. We found that F3 of /e/ increased following the perceptual training with somatosensory stimulation. The result suggests that perceptual training using specific pairs of auditory-somatosensory stimulation may change speech production. This is in line with previous studies [9]–[11] that show the effect of perceptual training with repetitive stimuli transferred to the production mechanism through sensory-motor relationships.

Somatosensory inputs can play an important role in the link between production and perception mechanisms. Previous studies showed that somatosensory inputs affect the perception of speech sounds [16], [20]. Specifically, Ohashi and Ito [18] demonstrated that somatosensory inputs in speech motor training contribute to recalibration of speech perception. Ito and Ogane [19] also showed that specific pairs of auditory-somatosensory stimulation in perceptual training also recalibrate the perception of speech sounds. The current finding provides complemented insight on the role of somatosensory information in speech learning.

F3 is known to increase when the front cavity is decreased in size by lip spreading [21], [22]. The current F3 increase for the production of /e/ in the post-test suggests that the participants spread their lips more following the training. Since the backward skin stretch in the current somatosensory stimulation can be associated with the production of /e/, a repetitive presentation of this somatosensory stimulation paired with the presentation of /e/ could be a key to induce the change in F3 following training. Since the somatosensory stimulation evokes the vowel /e/, this could explain why the change concerns, in the post-test, the production of /e/, but not of /ø/.

The findings are consistent with researches showing sensorimotor interaction in both speech perception and speech production [6], [7], [9]–[11]. In summary, repetitive exposure to specific pairs of auditory-somatosensory inputs during speech perceptual training may affect the speech production mechanisms.

5. Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklowdoska-Curie Grant Agreement No 860755 (Comm4CHILD project) and by National Institute on Deafness and Other Communication Disorders Grant R01-DC017439. We would also like to thank Clément Guichet for his help with the data collection.

6. References

[1] R. L. Sainburg, M. F. Ghilardi, H. Poizner, and C. Ghez, 'Control of limb dynamics in normal subjects and patients without proprioception', *J Neurophysiol*, vol. 73, no. 2, pp. 820–35, 1995.

[2] F. H. Guenther, 'Cortical interactions underlying the production of speech sounds', *J Commun Disord*, vol. 39, no. 5, pp. 350–65, Oct. 2006, doi: 10.1016/j.jcomdis.2006.06.013.

[3] J. F. Houde and M. I. Jordan, 'Sensorimotor Adaptation in Speech Production', *Science*, vol. 279, no. 5354, pp. 1213–1216, Feb. 1998, doi: 10.1126/science.279.5354.1213.

[4] S. Tremblay, D. M. Shiller, and D. J. Ostry, 'Somatosensory basis of speech production', *Nature*, vol. 423, no. 6942, pp. 866–869, Jun. 2003, doi: 10.1038/nature01710.

[5] D. R. Lametti, S. M. Nasir, and D. J. Ostry, 'Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback', *J. Neurosci.*, vol. 32, no. 27, pp. 9351–9358, Jul. 2012, doi: 10.1523/JNEUROSCI.0404-12.2012.

[6] S. M. Nasir and D. J. Ostry, 'Auditory plasticity and speech motor learning', *Proc. Natl. Acad. Sci.*, vol. 106, no. 48, pp. 20470–20475, Dec. 2009, doi: 10.1073/pnas.0907032106.

[7] D. M. Shiller, M. Sato, V. L. Gracco, and S. R. Baum, 'Perceptual recalibration of speech sounds following speech motor learning', *J. Acoust. Soc. Am.*, vol. 125, no. 2, pp. 1103–1113, Feb. 2009, doi: 10.1121/1.3058638.

[8] P. K. Kuhl, 'Early language acquisition: cracking the speech code', *Nat. Rev. Neurosci.*, vol. 5, no. 11, pp. 831–843, Nov. 2004, doi: 10.1038/nrn1533.

[9] W. E. Cooper and M. R. Lauritsen, 'Feature processing in the perception and production of speech', *Nature*, vol. 252, no. 5479, Art. no. 5479, Nov. 1974, doi: 10.1038/252121a0.

[10] W. E. Cooper and R. M. Nager, 'Perceptuo-motor adaptation to speech: an analysis of bisyllabic utterances and a neural model', *J. Acoust. Soc. Am.*, vol. 58, no. 1, pp. 256–265, Jul. 1975, doi: 10.1121/1.380655.

[11] D. G. Jamieson and M. F. Cheesman, 'The adaptation of produced voice-onset time', *J. Phon.*, vol. 15, no. 1, pp. 15–27, Jan. 1987, doi: 10.1016/S0095-4470(19)30534-0.

[12] D. R. Lametti, S. A. Krol, D. M. Shiller, and D. J. Ostry, 'Brief Periods of Auditory Perceptual Training Can Determine the Sensory Targets of Speech Motor Learning', *Psychol. Sci.*, vol. 25, no. 7, pp. 1325–1336, Jul. 2014, doi: 10.1177/0956797614529978.

[13] S. E. Lively, D. B. Pisoni, R. A. Yamada, Y. Tohkura, and T. Yamada, 'Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories', *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2076–2087, Oct. 1994, doi: 10.1121/1.410149.

[14] A. Suemitsu, T. Ito, and M. Tiede, 'An EMA-based articulatory feedback approach to facilitate L2 speech production learning', *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3336–3336, May 2013, doi: 10.1121/1.4805613.

[15] C. Haldin *et al.*, 'Speech recovery and language plasticity can be facilitated by Sensori-Motor Fusion training in chronic non-fluent aphasia. A case report study', *Clin. Linguist. Phon.*, vol. 32, no. 7, pp. 595–621, Jul. 2018, doi: 10.1080/02699206.2017.1402090.

[16] T. Ito, M. Tiede, and D. J. Ostry, 'Somatosensory function in speech perception', *Proc. Natl. Acad. Sci.*, vol. 106, no. 4, pp. 1245–1248, Jan. 2009, doi: 10.1073/pnas.0810063106.

[17] T. Ito and D. J. Ostry, 'Somatosensory Contribution to Motor Learning Due to Facial Skin Deformation', *J. Neurophysiol.*, vol. 104, no. 3, pp. 1230–1238, Sep. 2010, doi: 10.1152/jn.00199.2010.

[18] H. Ohashi and T. Ito, 'Recalibration of auditory perception of speech due to orofacial somatosensory inputs during speech motor adaptation', *J. Neurophysiol.*, vol. 122, no. 5, pp. 2076–2084, Nov. 2019, doi: 10.1152/jn.00028.2019.

[19] T. Ito, R. Ogane, 'Repetitive exposure to orofacial somatosensory inputs in speech perceptual training modulates the vowel categorization in speech perception.', Front Psychology, To be published 2022.

[20] P. Trudeau-Fisette, T. Ito, and L. Ménard, 'Auditory and Somatosensory Interaction in Speech Perception in Children and Adults', *Front. Hum. Neurosci.*, vol. 13, p. 344, Oct. 2019, doi: 10.3389/fnhum.2019.00344.

[21] J.-L. Schwartz, D. Beautemps, C. Abry, and P. Escudier, 'Inter-individual and cross-linguistic strategies for the production of the [i] vs. [y] contrast', *J. Phon.*, vol. 21, no. 4, pp. 411–425, 1993.

[22] B. E. Lindblom and J. E. Sundberg, 'Acoustical consequences of lip, tongue, jaw, and larynx movement', *J. Acoust. Soc. Am.*, vol. 50, no. 4, pp. 1166–1179, Oct. 1971, doi: 10.1121/1.1912750.

[23] L. Ménard and L. Boe, 'L'emergence du systeme phonologique chez l'enfant : l'apport de la modelisation articulatoire', 2004, doi: 10.1353/CJL.2005.0003.

[24] T. Ito and H. Gomi, 'Cutaneous mechanoreceptors contribute to the generation of a cortical reflex in speech', *Neuroreport*, vol. 18, no. 9, pp. 907–910, Jun. 2007, doi: 10.1097/WNR.0b013e32810f2dfb.

[25] P. Boersma and D. Weenink, 'PRAAT, a system for doing phonetics by computer', *Glot Int.*, vol. 5, pp. 341–345, Jan. 2001.