Somatosensory contribution to audio-visual speech processing

Takayuki Ito, Hiroki Ohashi, Vincent L Gracco

Title:
Somatosensory contribution to audio-visual speech processing

Authors: Takayuki Ito [1,2*], Hiroki Ohashi [2], Vincent L. Gracco [2,3]

Author's affiliations:

1: Univ. Grenoble-Alpes, CNRS, Grenoble-INP, GIPSA-Lab, 11 rue des Mathématiques, Grenoble Campus BP46, F-38402, Saint Martin D'heres Cedex, France

2: Haskins Laboratories, 300 George Street, New Haven, CT, 06511, USA

3: McGill University, 845 Sherbrooke Street West, Montréal, QC, H3A 3R1, Canada

*: **Corresponding author**: Takayuki Ito

Address: GIPSA-Lab, CNRS, Univ. Grenoble-Alpes, Grenoble-INP,

11 rue des Mathématiques, Grenoble Campus BP46,

F-38402, Saint Martin D'heres Cedex, France

Email: takayuki.ito@gipsa-lab.grenoble-inp.fr

Tel: +33 (0)4 76 82 70 46

Fax: +33 (0)4 76 57 47 10

Abstract

Recent studies have demonstrated that the auditory speech perception of a listener can be modulated by somatosensory input applied to the facial skin suggesting that perception is an embodied process. However, speech perception is a multisensory process involving both the auditory and visual modalities. It is unknown whether and to what extent somatosensory stimulation to the facial skin modulates audio-visual speech perception. If speech perception is an embodied process, then somatosensory stimulation applied to the perceiver should influence audio-visual speech processing. Using the McGurk effect (the perceptual illusion that occurs when a sound is paired with the visual representation of a different sound, resulting in the perception of a third sound) we tested the prediction using a simple behavioral paradigm and at the neural level using event-related potentials (ERPs) and their cortical sources. We recorded ERPs from 64 scalp sites in response to congruent and incongruent audio-visual speech randomly presented with and without somatosensory stimulation associated with facial skin deformation. Subjects judged whether the production was /ba/ or not under all stimulus conditions. In the congruent audio-visual condition subjects identifying the sound as /ba/, but not in the incongruent condition consistent with the McGurk effect. Concurrent somatosensory stimulation improved the ability of participants to more correctly identify the production as /ba/ relative to the non-somatosensory condition in both congruent and incongruent conditions. ERP in response to the somatosensory stimulation for the incongruent condition reliably diverged 220 ms after stimulation onset. Cortical sources were estimated around the left anterior temporal gyrus, the right middle temporal gyrus, the right posterior superior temporal lobe and the right occipital region. The results

demonstrate a clear multisensory convergence of somatosensory and audio-visual processing in both behavioral and neural processing consistent with the perspective that speech perception is a self-referenced, sensorimotor process.

Keywords: Electroencephalography, Event-related potentials, Orofacial somatosensory processing, Audio-visual speech perception, Multisensory integration.

1. Introduction

Our perceptions are based on what we know and what we know is dependent on what we experience. For speech, the actions of the vocal tract associated with spoken language production provide a common currency (or parity) in that the sounds of the language that we hear are directly related to the movements and configurations of the vocal tract that we produce. While speech perception is often considered to be based on the acoustic properties or auditory objects of the signal (Bizley and Cohen 2013; Diehl and Kluender 1989), alternative findings suggest that perception can be modulated by external (sensory) input that codes motor (or sensorimotor) information to the listener (Gick and Derrick 2009; Ito et al. 2009; Ogane et al. 2020; Sams et al. 2005; Sato et al. 2013). For example, air puffs to the cheek of a perceiver that coincide with auditory speech stimuli alter participants' perceptual judgements (Gick and Derrick 2009), while orofacial skin stimulation changes the auditory perceptual discrimination of speech (Ito et al. 2009; Ogane et al. 2020; Trudeau-Fisette et al. 2019). These observations are consistent with the Motor Theory of Speech Perception (Liberman and Mattingly 1985) and the Direct Realist account suggests that perceiving speech is perceiving the vocal tract gestures (cf. (Galantucci et al. 2006) for overview). Consistent with these views is that listening to speech activates cortical areas related to speech production in the motor and premotor cortex (Fadiga et al. 2002; Grabski et al. 2013; Pulvermuller et al. 2006; Tremblay and Small 2011; Watkins et al. 2003; Wilson et al. 2004). Moreover, ventral motor and somatosensory areas reflect phonological information during speech perception (Schomers and Pulvermüller 2016).

For speech perception, visual information is a source of information that modifies speech perception when the auditory signal is degraded, ambiguous or incongruent (Girin et al. 2001; McGurk and MacDonald 1976; Sumby and Pollack 1954). The visual signal provides information on certain articulatory properties mostly from the facial skin and oral opening. If movement-related somatosensory input applied the listener interacts with an internal model used for perceptual evaluation, then orofacial stimulation as in our previous study, should have an effect when speech perception is based on audiovisual signals. Here we used incongruent auditory and visual speech [e.g. the McGurk effect (McGurk and MacDonald 1976)] to test the hypothesis that speech perception reflect a sensorimotor process based on prediction from an internal model which is action-based. We used movement related stimulation to the listener through orofacial skin stretch to evaluate the neural response to somatosensory stimulation of the facial skin through an analysis of the change in electroencephalographic (EEG) activity. Orofacial somatosensory input associated with facial skin deformation provides motion information for speech production (Connor and Abbs 1998; Ito and Gomi 2007; Ito and Ostry 2010; Johansson et al. 1988), and has been shown to interact in motion-specific ways to influence speech perception (Ito et al. 2009; Ogane et al. 2020; Trudeau-Fisette et al. 2019). The stimulation associated with facial skin deformation also changes cortical potentials for auditory speech perception (Ito et al. 2013, 2014), but the stimulation of lip tapping does not (Möttönen et al. 2005). This interaction is also specific with somatosensory inputs arising from orofacial area, but not from other body parts (Ito and Ostry 2012; Ogane et al. 2020). It appears that somatosensory input from the listener interacts with an internal model of speech production during speech perception as part of

an embodied process. We expect that information from orofacial somatosensory inputs associated with facial skin deformation would similarly modulate the audiovisual processing of speech providing supporting evidence for speech perception as an embodied process.

In addition to the behavioral effects, neuroimaging methods have identified the brain regions related to audiovisual speech. The superior temporal cortex contains a multisensory region that is modulated by temporal asynchrony in audiovisual processing (Macaluso et al. 2004; Miller and D'Esposito 2005; Stevenson et al. 2010; Wright et al. 2003) and is engaged by perceptual fusion (Bushara et al. 2003). Specifically the superior temporal sulcus (STS) can be considered an important site for the McGurk illusion (Beauchamp et al. 2010; Marques et al. 2014; Nath and Beauchamp 2012; Sekiyama et al. 2003). In addition, the right anterior temporal sulcus is activated for voice specific responses that are not related to the acoustic features of voices (von Kriegstein et al. 2003). Other areas that have been shown to be involved with audiovisual processing include the middle intra-parietal sulcus, along with motor speech regions of the brain involved in resolving and fusing incongruent audio-visual speech (Miller and D'Esposito 2005; Nath and Beauchamp 2012). An Event-Related Potentials (ERP) study of the McGurk effect (Bernstein et al. 2008) reported early (<100 msec) and simultaneous activations in areas of the supramarginal and angular gyrus, intra-parietal sulcus, the inferior frontal gyrus (Broca's area), and the dorsolateral prefrontal cortex. Early (<200 msec) processing of audiovisual stimuli is prominent in the left hemisphere, except for right hemisphere prominence in superior parietal cortex and secondary visual cortex.

Using orofacial stimulation and the McGurk stimuli we recorded EEG activity and examined the evoked responses and cortical sources. We predicted that the behavioral responses and ERPs would differ between the congruent and incongruent audiovisual conditions due to McGurk effect and this difference would be modulated by somatosensory inputs. Further and consistent with our assumption of internal modeling of the motor action, we expected the cortical sources to be located in brain areas known for action-perception coding.

2. Methods

2.1 Participants

Twelve native speakers of American English participated in the experiment. Sample size was determined based on our previous studies (Ito et al. 2013, 2014, 2020). The participants were all healthy young adults with normal hearing and right-handed. All participants signed informed consent forms approved by the Yale University Human Investigation Committee. The data of three participants were excluded from the analysis because they did not show an effect for the incongruent visual stimulation.

No part of the study procedures and study analyses was pre-registered prior to the research being conducted. The conditions of the ethics approval do not permit public archiving of anonymized study data. Readers seeking access to the data should contact the corresponding author. Access will be granted subject to completion of a formal data sharing agreement. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to

data analysis, all manipulations, and all measures in the study. Since the digital material for audio-visual stimulation includes identifiable information (full-face view producing the speech), they have not been placed in a public repository and can never be shared intact (even on request) because the stimuli cannot be de-identified. The code for stimulus presentation is publicly available at https://osf.io/vdkwa/. We applied known analytic methods for data analysis with the details provided in the following sections; there were no custom-made procedures.

2.2 Audio-visual stimulation

For all congruent audio-visual stimulation, we used the /ba/ syllable and for the incongruent condition we used the facial motion for the production of /ga/. The stimulus was recorded by a female speaker of American English. The expected perceptual illusion in the incongruent condition is that of /da/. Visual stimulation was presented on the monitor of a PC laptop. Audio stimulation was delivered binaurally through EEG-compatible earphones (Etymotic Research, ER3A).

2.3 Somatosensory stimulation

The details of the somatosensory stimulation procedure are described in our previous studies (Ito et al. 2015). Briefly, a small robotic device (Phantom 1.0, SensAble Technologies) applied skin stretch to two small plastic tabs that were attached bilaterally with tape to the skin at the sides of the mouth. A single cycle of a 3-Hz sinusoidal pattern with 4 N maximum force pulled the skin in a backward direction to evoke the ERP (Ito et al. 2014, 2020).

2.4 Experimental procedure

In total seven stimulation conditions were used; auditory only (A), somatosensory only (Soma), somatosensory-auditory (SomaA), auditory-visual (AV), somatosensory-auditory-visual (SomaAV), incongruent auditory-visual (AVm), incongruent somatosensory-auditory visual (SomaAVm). The conditions were presented in pseudo-random order with the constraint that all seven conditions were tested every seven trials. The interval from the subject's response to the onset of following stimulus was varied between 1000 and 1500 ms. Figure 1 represents the temporal sequence of three stimulations for one trial. As shown here, visual and somatosensory stimuli occurred earlier than the auditory stimulus. The onset of somatosensory stimulation was set 140 ms earlier than the onset of auditory stimulation so the peak of somatosensory stimulation would occur around the acoustic burst of the /b/ in "ba". This adjustment is based on previous findings that an interaction during speech processing is induced when the somatosensory stimulation leads the auditory stimulation (Ito et al. 2014; Ogane et al. 2020).

The participant's task was to indicate by key press whether the sound they heard was "ba" or not. Each response ("ba" and not "ba") were assigned into two keys in keyboard, respectively. The subject pressed these keys by using two fingers (mostly index and middle fingers) and were asked to respond during the trial interval. In the somatosensory alone condition (Soma) there was no auditory stimulation and the participants were instructed to press the key associated with not "ba". Participant

judgments constituted the behavioral measure. During the task, participants fixated their gaze on a cross in the middle of the computer screen and were instructed to maintain their gaze to eliminate eye blink artifacts during the ERP recording.

2.5 Behavioral performance

The probability that the participant classified the syllable as /ba/ was calculated for each condition. The somatosensory alone condition was not included in the analysis. Repeated measures analysis-of-variance (ANOVA) was used to compare judgement measures across the conditions.

2.6 EEG acquisition and pre-processing

EEG was recorded using a 64-electrode Biosemi ActiveTwo system (256 Hz sampling rate) along with eye motion components from electro-oculography. One hundred ERPs for each stimulus condition (seven hundred ERPs in total) were recorded. For the pre-processing, EEG signals were filtered using a 1-30 Hz band-pass filter and re-referenced to the average across all electrodes. EEG signals were then segmented into epochs between –300 and 500 ms relative to the onset of auditory stimulation. The bias level of each epoch was adjusted to the average amplitude in the pre-stimulus interval (–300 to –200 ms). By applying independent components analysis (Onton et al. 2006), the extracted components corresponding to large signal noise and artifacts including eye-blink and motion were excluded by manual inspection. Finally, the processed ERPs were averaged across trials in each condition on a per-participant basis.

2.7 ERP waveform analysis

Here we focused on Cz because the largest amplitude of the auditory ERPs is usually represented in the mid sagittal plane. To extract the specific interaction related-components, we examined the resulting waveforms generated through a summation or subtraction identified in the Results section below. To deal with the multiple comparison problem we employed a cluster-based analysis based on a permutation test generating a distribution of t-scores (Groppe et al. 2011). The procedure was repeated 1000 times identifying the sequence of sampling points that showed a reliable difference.

2.8 Source Localization

To estimate the cortical sources associated with the specific comparisons we used the sLORETA/eLORETA software package (Pascual-Marqui 2002) using a realistic head model (Fuchs et al. 2002) from the Montreal Neurological Institute's MNI152 template (Mazziotta et al. 2001) with the standard electrode positions on the MNI152 scalp corresponding to the Biosemi EEG system. sLoreta images were calculated for each participant and for each extracted response. Using these images, statistical parametrical mapping was conducted by applying non-parametrical tests across participants (Nichols and Holmes 2002), based on estimating the empirical probability distribution for the max-statistic (e.g. the maximum of a t-stat) under the null hypothesis via randomization. This methodology corrects for multiple testing for all electrodes and voxels. We applied a log F-value in comparing responses, and a log t-value for the amplitude of single responses that differed from zero.

3. Results

3.1 Behavioral results

Figure 2 represents the probability that the subjects correctly identified the stimuli as congruent (/ba/) and incongruent (not /ba/). The judgment probabilities were high (close to 1) for the congruent condition, reduced under the A and Soma-A conditions (around .5) and further reduced (close to 0) for the incongruent AV and Soma-AV conditions. The results also show that our auditory stimulation was relatively ambiguous because the probability of auditory alone correctly identified was around chance. Importantly, we found that somatosensory stimulation increased the probability in all three conditions including the McGurk condition [$F(1,24) = 8.727$, $p < 0.01$] with no interaction [$F(2,24) = 2.330$, $p > 0.1$].

3.2 Recorded ERP responses

The three panels in Figure 3A represent averaged ERPs at Cz for all seven conditions. Auditory stimulation clearly induced a typical P1-N1-P2 ERP sequence (red line in the left panel of Fig. 3A) and the somatosensory ERP (black line in the left panel of Fig. 3A) was also consistent with the previous observations of a negative-positive sequence (Ito et al. 2014, 2015) corresponding to somatosensory the N1-P2 components (Inui et al. 2003). The somatosensory ERP (Soma) was relatively larger than the auditory ERP (A) while the Soma and SomaA patterns are similar but with a reduction in amplitude in the negative and positive-going peaks for the SomA condition. Between these two conditions (Soma and SomaA), the first negative peak was not different [F(1,8)

= 2.429, $p > 0.15$] and the second positive peak was reliably different [$F(1,8) = 6.733$, $p < 0.05$].

For audiovisual processing under the congruent condition without somatosensory stimulation (AV), the evoked response (red line in the right panel in Fig. 3A) differs from the ERP of auditory alone (A). Under the incongruent condition (AVm), the N1 peak changes polarity (positive-going) and is followed by a large negative peak with both conditions converging at around 300 msec.

For the congruent AV plus somatosensory stimulation (SomaAV), there is a large negative going peak early (around 50 msec) and a later positive going peak at around 110 msec. For the incongruent condition (SomaAVm), the early peak changed in a positive direction with the later peak changed in the negative direction relative to the congruent condition.

3.3 ERPs for incongruent audio-visual stimulation

The incongruent visual stimulation resulted in a difference in the ERPs between the two processing conditions (the right two panels of Fig. 3A) as mentioned above. The two ERPs responses for congruent (red line) and incongruent (blue line) conditions diverged in the period between 40 ms (early) and 130ms (late) after the auditory onset. The shaded areas represent reliable differences obtained by the cluster-based analysis [42 - 86 ms and 105 - 125 ms for SomaAVm and SomaAV, and 39 - 86 ms and 105 – 129 ms for AVm and AV]. These periods correspond to the P1-N1 sequence in the auditory ERP. ANOVA also showed significant differences in each time period (the early period: with somatosensory $F(1,8) = 29.48$, $p < 0.001$, and without somatosensory $F(1,8) = 25.24$, $p <$

0.005, and the late period: with somatosensory $F(1,8) = 21.69$, $p < 0.005$, and without somatosensory $F(1,8) = 26.03$, $p < 0.001$). The results indicate that the incongruent visual stimulation was differently processed in the relatively early periods (40-130 ms after auditory stimulation) related to the McGurk perturbation.

Figure 3B is the difference in the AV and AV+Soma conditions after subtracting the congruent conditions from the incongruent conditions. Source localization was applied to each peak of the positive-negative sequence separately using a 30 ms time window. Due to similarity of the early and late differences, we applied source localization to the data averaged across the two responses to increase the SNR ratio. The estimated sources were located in right occipital lobe ([20, –90, 30]; MNI coordinates, $p < 0.01$, the left panels of Fig. 3D) around the peak of the early positive response (58-85 ms) and in right inferior occipital lobe ([25, –90, –20]; MNI coordinates, $p < 0.01$, the right panels of Fig. 3D) around the peak of the following negative response (105-132 ms). In addition, as highlighted in Figure 3B, cluster based-analysis showed a reliable difference in the period between 226 and 266 ms (shaded area in Figure 3B). This is consistent with the result of ANOVA [$F(1,8) = 9.671$, $p < 0.05$]. The source localization reflected a significant difference in the right posterior superior temporal lobe ([65, –35, 20]; MNI coordinates, $p < 0.05$) as shown in Fig. 3C.

3.4 Localization of the somatosensory modulation

In order to examine the effect of the orofacial stimulation on the different AV processing conditions we compared the recorded and summed somatosensory-audio-visual ERPs. Figure 4A represents three pairs of recorded (blue-solid line) and summed

ERPs (black-dashed line). In all pairs, the recoded ERPs were reduced from the summed ERPs illustrating the effects of the somatosensory stimulation. The cluster-based analysis showed that the differences were seen in the periods 125 – 246 ms for SomaAVm, 98-234 ms for SomaAV, and 152-238 ms for SomaA (shaded areas in the figures) with the actual ERP reduced from a simple addition of the respective conditions. The difference from the subtraction between recorded and summed responses are shown in Figure 4B. The responses in the conditions with visual stimulation were initiated approximately 100 msec earlier while the differences were similar between 150 ms to 220 ms with a later divergence specifically in the audio-visual conditions (SomaAVm and SomaAV). We applied source localization to the time period where the differences were similar (172-210 ms). ANOVAs for the peaks in this time periods showed reliable differences (Congruent: $F(1,8) = 26.02$, $p < 0.001$ and Incongruent: $F(1,8) = 13.26$, $p < 0.001$). For source localization, we took an average between the congruent and incongruent conditions to increase the SNR ratio. The estimated source was located around the left anterior middle temporal gyrus ([-55, 5, -25]; MNI coordinates, $p < 0.01$) (Figure 4C).

3.5 Somatosensory-visual interaction

We were also interested in whether the somatosensory and visual inputs interact separately from auditory processing. To examine this, we extracted the visual-related component in audio-visual processing and evaluated whether the visual related-components were modified due to somatosensory inputs. The visual-related components were obtained by subtracting the somatosensory-auditory ERP from somatosensory-auditory-visual ERP (SomaAVm - SomaA and SomaAV - SomaA) and by subtracting the

auditory ERP from audio-visual ERP (AVm-A and AV-A), shown in Figure 5A. The left panel represents the ERPs with incongruent visual stimulations and the right panel represents the ERPs with congruent visual stimulations. By subtracting the ERPs without somatosensory stimulation from the ones with somatosensory stimulation in each panel (Figure 5B), we found that the visual related-components were changed with the somatosensory stimulation in the period around 100-150 ms. Cluster-based analysis showed a significant difference for the congruent visual related component (101-133 ms shaded area in the right panel of Figure 5A) but did not show any significant change for the incongruent visual related component. This statistical difference between congruent and incongruent visual stimuli was consistent with the cluster-based analysis showing reliable differences between recorded and summed responses over the period between 101-133ms in the congruent condition. This is also consistent with the result of ANOVA [Congruent $F(1,8)= 11.35$, $p < 0.01$ and Incongruent: $F(1,8) = 4.38$, $p = 0.07$]. Finally, we applied source localization on the visual-related component taking an average of the two responses shown in Figure 5B. The estimated site was the right middle temporal gyrus ([55, 10, –25] in MNI coordinates, $p< 0.01$, Fig. 5C).

4. Discussion

The main focus of this study was to determine whether somatosensory stimulation applied to the face of a perceiver modulated their audiovisual speech perception. Previous work has demonstrated that stimulation of the facial skin of a perceiver can shift their perceptual category if the stimulation is consistent with the production of the particular phonetic segment begin perceived. In the present study, using audio-visual speech

processing of congruent and incongruent A-V stimuli, we found that somatosensory stimulation of the orofacial skin positively affected speech processing by improving judgment probability. Although there was only minimal improvement for the congruent AV condition, this is partially explained by the almost ceiling level performance by the participants. Judgement probability was improved with orofacial stimulation for the auditory only and the incongruent audiovisual processing conditions. The change in evoked potentials were observed for all conditions and a number of possible sources of the somatosensory modulation of audiovisual processing were estimated. The main regions associated with orofacial stimulation were located in the right posterior part of superior temporal gyrus, a region associated with action-perception coupling and the left anterior superior temporal region, an area associated with perceiving human voices. The implications of the current findings are discussed below.

4.1 Cortical sources of somatosensory interactions

The current source modeling estimated two potential sources for the integration of the orofacial somatosensory stimulation during audiovisual speech processing. The left anterior middle temporal gyrus was associated with processing both congruent and incongruent conditions. This region, part of the ventral auditory stream, is associated with phoneme and word recognition (DeWitt and Rauschecker 2012) and phonotactic processing (Obrig et al. 2016). The right posterior part of superior temporal gyrus (STG) was also identified as a source of somatosensory interaction for the McGurk effect. This area is involved in visual speech perception, in which auditory processing is not involved together with the other visual processing areas such as the fusiform gyrus, and middle

temporal gyrus (Calvert et al. 1997; Campbell et al. 2001) [see review in (Bernstein and Liebenthal 2014)]. Moreover, the STG responds strongly to a moving face (Pitcher et al. 2011). Considering that our facial skin deformation provides kinesthetic information associated with speech movement (Ito and Ostry 2010), the processing of motion-related information appears consistent with the STG as an additional region associated with somatosensory integration for the McGurk stimuli. Moreover, the right hemisphere focus is consistent with right hemisphere lateralization for face processing (Kanwisher et al. 1997; Pitcher et al. 2011) and audio visual processing (Davis et al. 2008) and audio-visual speech perception (Möttönen et al. 2004). Together the left anterior STG and the right STS/G appear to be sites for the integration of the somatosensory stimulation for the different audiovisual speech conditions.

4.2 Timing considerations

Previous studies using magnetencephalography have detailed the time course of audio-visual speech processing for the McGurk effect (Hertrich et al. 2007; Möttönen et al. 2004). The early component (< 200 ms) is processed in the sensory-specific areas and the latter component (> 250 ms) is processed in multisensory regions of the human temporal cortex (Möttönen et al. 2004). A similar time course can be seen in our results. We found clear differences between congruent and incongruent audio-visual stimulation up to 220 ms after the auditory onset, and this component was modified by the somatosensory input after 220 ms. In these two components, the estimated sites are sensory-specific regions (around visual cortex) for the early component and the STG site in the later component. Since somatosensory inputs did not modify the response between

congruent and incongruent visual stimulation, it appears that somatosensory input accompanying the speech perceptual stimuli is processed relatively late for the McGurk effect.

4.3 Speech perception as a sensorimotor process

Speech perception is clearly an audiovisual process. However, the current and previous observations that somatosensory inputs can influence speech perception suggests that speech perception is not coded in terms of sensory input only. Rather, it appears that the process involves a transformation of the sensory signals integrating the different sources of information to onto some common space. The common factor across the visual and auditory modalities is that they all are providing input to the perceiver on the motion of different parts of the vocal tract; visual input on the visible articulation, auditory input in the vocal tract configuration. The somatosensory input from the orofacial stimulation provided additional movement-related information consistent with an integration of the external input into this common space. That is, while the stimulation in the current study provided additional sensory input to the perceiver, the input was not related to the talker's vocal tract but rather to the perceiver's.

Viewed within a broader theoretical framework, somatosensory influences on speech perception suggest that speech perception is an embodied process relying on a body-centric representation for sensory identification (Goldman 2012, 2013). At a fundamental level of cognitive functioning, embodied processes are reflected in observations like wearing a heavy backpack making hills look steeper (Bhalla and Proffitt 1999) or

grasping a baton making reachable objects look closer (Witt et al. 2005). The assumption is that perception is modified by our ability to act on what we are perceiving. For speech perception, the Motor Theory of Speech Perception suggests that speech decoding is based on the recovery of the motor cause of speech stimuli, and that articulatory/ motor representations provide the basis of speech communication (Liberman et al. 1967; Liberman and Mattingly 1985). Listening to speech activates cortical areas related to speech production in the motor and premotor cortex (Fadiga et al. 2002; Grabski et al. 2013; Pulvermuller et al. 2006; Tremblay and Small 2011; Watkins et al. 2003; Wilson et al. 2004) and behavioral studies have shown that articulatory movements preceding or accompanying the presentation of auditory stimuli modify speech perception, by e.g. motor stimulation (Sato et al. 2011). As such, the speech perception process, while engaged by and relying on auditory and visual input, also appears to be integrating movement-related information into a common reference frame (the vocal tract). The somatosensory input to the perceiver, while sensory in nature, provides input that reflects a change in the configuration of the perceiver's vocal tract. It is suggested that speech perception is a sensorimotor process that is affected by both the speaker who generates the signals and the listener who incorporates them into an embodied representation.

5. Conclusion

Our results indicate that somatosensory inputs associated with facial skin deformation interact with audio-visual speech processing. The estimated cortical sites associated with somatosensory interaction were located in regions associated with action-perception coupling and with perceiving human voices. The results demonstrate a multisensory

convergence of somatosensory and audio-visual processing during speech perception and support the involvement of an action-perception mechanism in the perceptual process.

Funding

This work was supported by the National Institute on Deafness and Other Communication Disorders Grants R21DC012502, R21DC013915 and R01DC017439.

Acknowledgements

References

Beauchamp MS, **Nath AR**, **Pasalar S**. fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *J Neurosci* 30: 2414–2417, 2010.

Bernstein LE, **Auer ET**, **Wagner M**, **Ponton CW**. Spatiotemporal dynamics of audiovisual speech processing. *Neuroimage* 39: 423–435, 2008.

Bernstein LE, **Liebenthal E**. Neural pathways for visual speech perception. *Front Neurosci* 8, 2014.

Bhalla M, **Proffitt DR**. Visual-motor recalibration in geographical slant perception. *J Exp Psychol Hum Percept Perform* 25: 1076–1096, 1999.

Bizley JK, **Cohen YE**. The what, where and how of auditory-object perception. *Nat Rev Neurosci* 14: 693–707, 2013.

Bushara KO, **Hanakawa T**, **Immisch I**, **Toma K**, **Kansaku K**, **Hallett M**. Neural correlates of cross-modal binding. *Nat Neurosci* 6: 190–195, 2003.

Calvert GA, **Bullmore ET**, **Brammer MJ**, **Campbell R**, **Williams SC**, **McGuire PK**, **Woodruff PW**, **Iversen SD**, **David AS**. Activation of auditory cortex during silent lipreading. *Science* 276: 593–596, 1997.

Campbell R, **MacSweeney M**, **Surguladze S**, **Calvert G**, **McGuire P**, **Suckling J**, **Brammer MJ**, **David AS**. Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Brain Res Cogn Brain Res* 12: 233–243, 2001.

Connor NP, **Abbs JH**. Orofacial proprioception: analyses of cutaneous mechanoreceptor population properties using artificial neural networks. *Journal of communication disorders* 31: 535–42; 553, 1998.

Davis C, **Kislyuk D**, **Kim J**, **Sams M**. The effect of viewing speech on auditory speech processing is different in the left and right hemispheres. *Brain Res* 1242: 151–161, 2008.

DeWitt I, **Rauschecker JP**. Phoneme and word recognition in the auditory ventral stream. *PNAS* 109: E505–E514, 2012.

Diehl RL, **Kluender KR**. On the objects of speech perception. *Ecol Psychol* 1: 121–44, 1989.

Fadiga L, **Craighero L**, **Buccino G**, **Rizzolatti G**. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur J Neurosci* 15: 399–402, 2002.

Fuchs M, **Kastner J**, **Wagner M**, **Hawes S**, **Ebersole JS**. A standardized boundary element method volume conductor model. *Clin Neurophysiol* 113: 702–712, 2002.

Galantucci B, **Fowler CA**, **Turvey MT**. The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review* 13: 361–377, 2006.

Gick B, **Derrick D**. Aero-tactile integration in speech perception. *Nature* 462: 502–4, 2009.

Girin L, **Schwartz J-L**, **Feng G**. Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America* 109: 3007–3020, 2001.

Goldman AI. A Moderate Approach to Embodied Cognitive Science. *RevPhilPsych* 3: 71–88, 2012.

Goldman AI. Joint Ventures: Mindreading, Mirroring, and Embodied Cognition. 2013.

Grabski K, **Schwartz J-L**, **Lamalle L**, **Vilain C**, **Vallée N**, **Baciu M**, **Le Bas J-F**, **Sato M**. Shared and distinct neural correlates of vowel perception and production. *Journal of Neurolinguistics* 26: 384–408, 2013.

Groppe DM, **Urbach TP**, **Kutas M**. Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology* 48: 1711–25, 2011.

Hertrich I, **Mathiak K**, **Lutzenberger W**, **Menning H**, **Ackermann H**. Sequential audiovisual interactions during speech perception: a whole-head MEG study. *Neuropsychologia* 45: 1342–1354, 2007.

Inui K, **Tran TD**, **Qiu Y**, **Wang X**, **Hoshiyama M**, **Kakigi R**. A comparative magnetoencephalographic study of cortical activations evoked by noxious and innocuous somatosensory stimulations. *Neuroscience* 120: 235–248, 2003.

Ito T, **Gomi H**. Cutaneous mechanoreceptors contribute to the generation of a cortical reflex in speech. *Neuroreport* 18: 907–10, 2007.

Ito T, **Gracco VL**, **Ostry DJ**. Temporal factors affecting somatosensory-auditory interactions in speech processing. *Frontiers in psychology* 5: 1198, 2014.

Ito T, **Johns AR**, **Ostry DJ**. Left lateralized enhancement of orofacial somatosensory processing due to speech sounds. *J Speech Lang Hear Res* 56: S1875-81, 2013.

Ito T, **Ohashi H**, **Gracco VL**. Changes of orofacial somatosensory attenuation during speech production. *Neurosci Lett* 730: 135045, 2020.

Ito T, **Ostry DJ**. Somatosensory contribution to motor learning due to facial skin deformation. *J Neurophysiol* 104: 1230–8, 2010.

Ito T, **Ostry DJ**. Speech sounds alter facial skin sensation. *J Neurophysiol* 107: 442–7, 2012.

Ito T, **Ostry DJ**, **Gracco VL**. Somatosensory event-related potentials from orofacial skin stretch stimulation. *Journal of visualized experiments : JoVE* e53621, 2015.

Ito T, **Tiede M**, **Ostry DJ**. Somatosensory function in speech perception. *Proc Natl Acad Sci U S A* 106: 1245–8, 2009.

Johansson RS, **Trulsson M**, **Olsson KÂ**, **Abbs JH**. Mechanoreceptive afferent activity in the infraorbital nerve in man during speech and chewing movements. *Exp Brain Res* 72: 209–14, 1988.

Kanwisher N, **McDermott J**, **Chun MM**. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci* 17: 4302–4311, 1997.

von Kriegstein K, **Eger E**, **Kleinschmidt A**, **Giraud AL**. Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research* 17: 48–55, 2003.

Liberman AM, **Cooper FS**, **Shankweiler DP**, **Studdert-Kennedy M**. Perception of the speech code. *Psychol Rev* 74: 431–61, 1967.

Liberman AM, **Mattingly IG**. The motor theory of speech perception revised. *Cognition* 21: 1–36, 1985.

Macaluso E, **George N**, **Dolan R**, **Spence C**, **Driver J**. Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage* 21: 725–732, 2004.

Marques LM, **Lapenta OM**, **Merabet LB**, **Bolognini N**, **Boggio PS**. Tuning and disrupting the brain—modulating the McGurk illusion with electrical stimulation. *Front Hum Neurosci* 8, 2014.

Mazziotta J, **Toga A**, **Evans A**, **Fox P**, **Lancaster J**, **Zilles K**, **Woods R**, **Paus T**, **Simpson G**, **Pike B**, **Holmes C**, **Collins L**, **Thompson P**, **MacDonald D**, **Iacoboni M**, **Schormann T**, **Amunts K**, **Palomero-Gallagher N**, **Geyer S**, **Parsons L**, **Narr K**, **Kabani N**, **Le Goualher G**, **Boomsma D**, **Cannon T**, **Kawashima R**, **Mazoyer B**. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos Trans R Soc Lond B Biol Sci* 356: 1293–1322, 2001.

McGurk H, **MacDonald J**. Hearing lips and seeing voices. *Nature* 264: 746–8, 1976.

Miller LM, **D'Esposito M**. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J Neurosci* 25: 5884–5893, 2005.

Möttönen R, **Järveläinen J**, **Sams M**, **Hari R**. Viewing speech modulates activity in the left SI mouth cortex. *Neuroimage* 24: 731–7, 2005.

Möttönen R, **Schürmann M**, **Sams M**. Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neurosci Lett* 363: 112–115, 2004.

Nath AR, **Beauchamp MS**. A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59: 781–787, 2012.

Nichols TE, **Holmes AP**. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* 15: 1–25, 2002.

Obrig H, **Mentzel J**, **Rossi S**. Universal and language-specific sublexical cues in speech perception: a novel electroencephalography-lesion approach. *Brain* 139: 1800–1816, 2016.

Ogane R, **Schwartz J-L**, **Ito T**. Orofacial somatosensory inputs modulate word segmentation in lexical decision. *Cognition* 197: 104163, 2020.

Onton J, **Westerfield M**, **Townsend J**, **Makeig S**. Imaging human EEG dynamics using independent component analysis. *Neurosci Biobehav Rev* 30: 808–22, 2006.

Pascual-Marqui RD. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods and findings in experimental and clinical pharmacology* 24 Suppl D: 5–12, 2002.

Pitcher D, **Dilks DD**, **Saxe RR**, **Triantafyllou C**, **Kanwisher N**. Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage* 56: 2356–2363, 2011.

Pulvermuller F, **Huss M**, **Kherif F**, **Moscoso del Prado Martin F**, **Hauk O**, **Shtyrov Y**. Motor cortex maps articulatory features of speech sounds. *Proc Natl Acad Sci U S A* 103: 7865–70, 2006.

Sams M, **Möttönen R**, **Sihvonen T**. Seeing and hearing others and oneself talk. *Brain Res Cogn Brain Res* 23: 429–35, 2005.

Sato M, **Grabski K**, **Glenberg AM**, **Brisebois A**, **Basirat A**, **Menard L**, **Cattaneo L**. Articulatory bias in speech categorization: Evidence from use-induced motor plasticity. *Cortex; a journal devoted to the study of the nervous system and behavior* 47: 1001–3, 2011.

Sato M, **Troille E**, **Ménard L**, **Cathiard M-A**, **Gracco V**. Silent articulation modulates auditory and audiovisual speech perception. *Exp Brain Res* 227: 275–288, 2013.

Schomers MR, **Pulvermüller F**. Is the Sensorimotor Cortex Relevant for Speech Perception and Understanding? An Integrative Review. *Front Hum Neurosci* 10, 2016.

Sekiyama K, **Kanno I**, **Miura S**, **Sugita Y**. Auditory-visual speech perception examined by fMRI and PET. *Neurosci Res* 47: 277–287, 2003.

Stevenson RA, **Altieri NA**, **Kim S**, **Pisoni DB**, **James TW**. Neural processing of asynchronous audiovisual speech perception. *Neuroimage* 49: 3308, 2010.

Sumby WH, **Pollack I**. Visual Contribution to Speech Intelligibility in Noise. *J Acoust Soc Am* 26: 212–15, 1954.

Tremblay P, **Small SL**. On the context-dependent nature of the contribution of the ventral premotor cortex to speech perception. *Neuroimage* 57: 1561–1571, 2011.

Trudeau-Fisette P, **Ito T**, **Ménard L**. Auditory and Somatosensory Interaction in Speech Perception in Children and Adults. *Front Hum Neurosci* 13, 2019.

Watkins KE, **Strafella AP**, **Paus T**. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41: 989–94, 2003.

Wilson SM, **Saygin AP**, **Sereno MI**, **Iacoboni M**. Listening to speech activates motor areas involved in speech production. *Nat Neurosci* 7: 701–2, 2004.

Witt JK, **Proffitt DR**, **Epstein W**. Tool use affects perceived distance, but only when you intend to use it. *J Exp Psychol Hum Percept Perform* 31: 880–888, 2005.

Wright TM, **Pelphrey KA**, **Allison T**, **McKeown MJ**, **McCarthy G**. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb Cortex* 13: 1034–1043, 2003.

Figure Legends

Figure 1: The temporal relationship of audio(A)-visual(V)-somatosensory(S) stimulations. Zero represents the onset of auditory stimulation, which was used for the alignment in ERP data analysis. The A is the audio signal for /ba/, the V is the visual image used for the congruent visual stimuli, and the S is the trajectory of the stretching force the for somatosensory stimulation.

Figure 2: The bars represent the average probability of the participants identifying whether the audio signal was "ba" for each condition. Error bars represent the standard error across the participants.

Figure 3: **A**: The averaged event-related potentials for all seven conditions recorded at Cz. For the left panel, the red arrows represent auditory P1-N1-P2 peaks and the black arrows represent somatosensory N1-P2 peaks, respectively. **B**: The subtracted responses for SomaAV and AV reflecting the difference between the incongruent and congruent conditions. **C** and **D**: The estimated cortical sources in the shaded period in **B** are shown in panel **C** and for the difference in the early and late peaks are shown in panel **D**. Color bars represent the range of sLORETA values.

Figure 4: **A**: A comparison between the recorded and summed responses to estimate the somatosensory interaction for the audio-visual processing. The responses were obtained by summing somatosensory-alone with the audio-visual responses. **B**: The results of the subtracted responses for all the conditions. **C**: The cortical source localization estimates

from the epoch around the peak of somatosensory-auditory interaction (arrow in **B**). Color bar represents the range of sLORETA values.

Figure 5: **A**: The estimate of the component associated with the visual interaction with auditory-somatosensory processing. Each component was obtained by subtracting the ERPs without visual stimulation from the ERPs with visual stimulation. **B**: The ERP differences for the congruent and incongruent condition with and without visual stimulation. **C**: The estimated cortical source for the epoch around the peak of the ERP differences in panel **B.** Color bar represents the range of sLORETA values.