INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Event-related potentials associated with somatosensory effect in audio-visual speech perception

Takayuki Ito[1,2], *Hiroki Ohashi*[2], *Eva Montas*[2], *Vincent L. Gracco*[2,3]

[1]GIPSA lab, CNRS, France
[2]Haskins Laboratories, USA
[3]McGill University, Canada

takayuki.ito@gipsa-lab.fr, ohashi@haskins.yale.edu, montas0392@gmail.com, vincent.gracco@mcgill.ca

Abstract

Speech perception often involves multisensory processing. Although previous studies have demonstrated visual [1, 2] and somatosensory interactions [3, 4] with auditory processing, it is not clear whether somatosensory information can contribute to the processing of audio-visual speech perception. This study explored the neural consequence of somatosensory interactions in audio-visual speech processing. We assessed whether somatosensory orofacial stimulation influenced event-related potentials (ERPs) in response to an audio-visual speech illusion (the McGurk Effect [1]). 64 scalp sites of ERPs were recorded in response to audio-visual speech stimulation and somatosensory stimulation. In the audio-visual condition, an auditory stimulus /ba/ was synchronized with the video of congruent facial motion (the production of /ba/) or incongruent facial motion (the production of the /da/: McGurk condition). These two audio-visual stimulations were randomly presented with and without somatosensory stimulation associated with facial skin deformation. We found ERPs differences associated with the McGurk effect in the presence of the somatosensory conditions. ERPs for the McGurk effect reliably diverge around 280 ms after auditory onset. The results demonstrate a change of cortical potential of audio-visual processing due to somatosensory inputs and suggest that somatosensory information encoding facial motion also influences speech processing.

Index Terms: speech perception, multisensory interaction, skin stretch perturbation, EEG

1. Introduction

While speech perception is often considered an auditory process, other sensory modalities are also involved. Visual information is a predominant source of information to facilitate speech perception when the auditory signal is degraded or ambiguous such as in a noisy environment [2]. One of the properties of the visual signal that appears to influence speech perception is that of motion. That is, the visual signal provides information on certain articulatory properties mostly from the facial skin and oral opening, and these properties are highly correlated with aspects of the acoustic signal (e.g., the speech envelope). As an additional source of information, orofacial somatosensory inputs have been shown to affect speech perception. Air puffs to the cheek that coincide with auditory speech stimuli alter participants' perceptual judgements [3]. Orofacial skin stretch changes the perceptual discrimination of speech as long as the stimulation applied to the facial skin is consistent with the stimulation that normally accompanies speech production [4]. It appears that somatosensory information, specifically from the face, signals motion related properties that have acoustic consequences. These consequences appear to have access to the perceptual process. Whereas the effects of visual and somatosensory inputs in speech perception have been examined separately, it is still unclear how these two inputs interact in the auditory processing of speech.

In the case of audio-visual speech perception, when visual information associated with speaking is incongruent with the auditory information, we perceive the sound differently from what was produced or seen (e.g. McGurk effect [1]). The articulatory movement information from the visual input interacts with the auditory information and shifts the speech perception toward an intermediate sound. Considering that somatosensory influences also produce motion information and have been shown to interact in motion-specific ways to influence speech perception [4], having motion information from two different modalities (vision and somatosensation) may combine to modify auditory processing of speech.

The current study examined the effect of somatosensory inputs during audio-visual speech perception in which the audio-visual information was congruent or incongruent. We recorded changes in event-related potentials in response to the presence or absence of facial skin deformation [5]. Facial skin stretch was used to generate somatosensory information similar to that produced during speech production [6-8]. An effect of somatosensory stimulation on audio-visual stimulation provides information on the potential importance of motion on speech perceptual processing.

2. Methods

We tested 5 native speakers of American English. The participants were all healthy young adults with normal hearing and all reported to be right-handed. All participants were signed in the consent forms approved by Yale HIC.

In the experiment, event-related potentials (ERPs) were recorded in response to several combinations of somatosensory, auditory and visual stimulation. For the audio-visual stimulation, we included McGurk trials [1], that is, the perceptual illusion that occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound. By applying additional somatosensory stimulation during audio-visual perception, we were able to examine evoked potential changes due to somatosensory interaction with audio-visual processing.

http://dx.doi.org/10.21437/Interspeech.2017-139

2.1. Stimulus presentation and task

For the somatosensory stimulation, facial skin deformation was applied using a small robotic device (SenSable Technology, Phantom 1.0). The details of the somatosensory stimulation device have been described in our previous studies [4, 5]. Briefly, two small plastic tabs were attached bilaterally with tape to the skin at the sides of the mouth. The tabs were connected to the robotic device using monofilament. Skin stretch force was applied in a backward direction by taking into account lip opening motion for the production of the stimulus sound /ba/ (see Figure 1). The stretch consisted of a single cycle of a 3-Hz sinusoid with 4 N maximum force. Similar patterns of facial skin stretch has successfully induced somatosensory ERPs in a previous study [5].

For the audio stimulation, the /ba/ syllable was used for all combinations of audio-visual stimulation since audio /ba/ and visual /ga/ combination likely induces a McGurk effect [1]. The stimulus was recorded by a female speaker of American English. We used the single syllable /ba/ for all audio stimulation. The audio /ba/ was slightly ambiguous in order to induce a clear visual effect in audio-visual perception. For the congruent visual condition, facial motion for the production of the /ba/ syllable was synchronized with the audio. For the incongruent visual condition, facial motion for the production of /ga/ was used. The expected perceptual illusion is that of /da/ (or /ga/). Audio stimulation was delivered binaurally through EEG-compatible earphones, which consists of plastic tubes (24 cm) and earpieces (Etymotic Research, ER3A). Visual stimulation was presented on the monitor of a PC laptop.

We tested seven stimulus conditions in total. There were two audio-visual conditions: congruent AV pairs (AV) and incongruent pairs (AVm). Each audio-visual condition was tested with and without somatosensory stimulations (S-AV and S-AVm). In addition to audio-visual condition, three conditions without visual stimulation were recorded: somatosensory alone (S), audio alone (A) and somatosensory-auditory (S-A) conditions. Somatosensory stimulation was delivered 140 ms prior to the auditory onset. This means that the peak amplitude of facial skin stretch correspmds to the timing of the prodiction of /b/. These seven conditions were presented in random order. 80 ERPs were recorded in each condition.

The participant's task was to indicate whether the sound they heard was /ba/ or not. The participants' response was recorded by key press. In the somatosensory alone condition, the participants were instructed to answer not /ba/. Participant judgements constituted the behavioral measures. The participants fixated their gaze on a cross displayed on the video monitor without blinking in order to eliminate artifacts during ERP recording. The cross was removed every 7 trials and the participants were given a short break.

2.2. EEG acquisition and data processing

Event-related potentials were recorded with 64 scalp sites (Biosemi ActiveTwo) and four electrodes for electro oculography (See Figure 1) and sampled at 256 Hz. A hundred responses per condition were recorded.

In pre-signal processing, EEG signals were first filtered with a 1– 30 Hz band-pass filter and then re-referenced to the average across all electrodes. A single epoch was extracted in the range between -500 and 1000 ms relative to the auditory stimulus onset. Bias levels were adjusted using the average amplitude in the pre-stimulus interval (-300 to -200 ms).

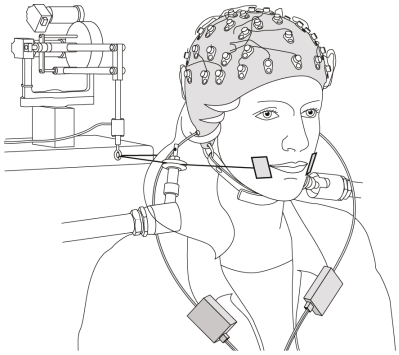

Figure 1: *Experimental setup for EEG recording with facial skin stretch perturbation.*

Trials with blinks and eye movement were rejected offline on the basis of horizontal and vertical electro-oculography (over ±150 μV). More than 90% of trials per condition were included in the analysis. The data was aligned at the auditory onset. In the following analysis, we focused on the potential at Fz specifically since this electrode shows typically the maximum amplitude of auditory ERP. In order to detect a change of ERPs between conditions, we calculated an algebraic summation or subtraction of ERPs.

The first analysis was carried out in terms of detecting an audio-visual interaction due to incongruent stimulation. We focused on the difference in ERP between congruent and incongruent audio-visual pairs. ERP in the congruent audio-visual pairs were subtracted from ERP in the incongruent pairs (Somatosensory-on: S-AVm minus S-AV and somatosensory-off: AVm minus AV). The obtained subtractions in somatosensory-ON and –OFF conditions were further compared using a cluster-based analysis.

Cluster-based permutation analysis [9] was applied to detect what time range shows a reliable difference between two ERPs. In this analysis, we first detected reliable differences in the ERP between two conditions at each sampling point. Paired t-test were applied to obtain the original t-score. Then a permutation was constructed by exchanging all ERP data in two conditions. T-scores were calculated again in this permutated data. The permutation test was repeated 1000 times and generated a distribution of t-score by permutation. We evaluated where the original t-value was in the distribution obtained by the permutation analysis. If the original t-score was in the 5 % range from the edge of the distribution, we determined that the original t-value was reliable. In the end, we extracted the longest length of cluster: the sequence of sampling points that show a reliable difference in permutation analysis, as the significant difference period.

Source localization analysis using sLORETA [10] was applied to the ERP differences in the range from the cluster-based permutation analysis.

In the second analysis, we processed the dataset in terms of somatosensory interaction with audio (-visual) processing. We derived the somatosensory-audio-visual ERPs by summing the ERPs in the somatosensory alone condition with the ERPs in the audio (A) or audio-visual conditions (AV or AVm) based on the assumption that the summed ERP responses should be equivalent to the ERP from the same stimuli presented simultaneously, if neural responses to each of the unisensory

stimuli are independent [11]. We assessed a difference between recorded and reconstructed ERPs through a subtraction analysis. Cluster-based permutation analysis was applied to detect a temporal difference in the subtracted responses.

2.3. Behavioral performance

Behavioral performance was evaluated using judgement probability. 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 since no auditory stimulation was involved. Note that in more than 95% of somatosensory trials participants responded not /ba/ as instructed. Repeated measures ANOVA was used to compare judgement measures across conditions.

We also examined the extent to which the perceptual judgements were correlated with amplitude change in somatosensory-elicited and derived potentials. In the correlation analysis, the behavioral measure was the difference of the participants' judgement probability between somatosensory on and off conditions, and the ERP measure was the somatosensory interaction activity that is the difference between the ERP recorded in response to somatosensory-auditory-visual stimulations and the summed ERP (somatosensory ERP plus auditory-visual ERP).

3. Results

Figure 2 shows event-related potentials in response to audio-visual stimulation. The top panel represents ERPs when somatosensory stimulation was presented together with AV stimulation. The bottom panel represents ERPs in the absent of somatosensory stimulation. The dashed line in each panel represents congruent AV conditions (S-AV for the top and AV for the bottom) and the solid line represents ERP in response to incongruent visual stimulation (S-AVm for the top and AVm for the bottom). The difference related to the incongruent conditions was observed early in the time course and in the periods between 50-300 ms after the audio stimulation in both somatosensory-on and –off conditions.

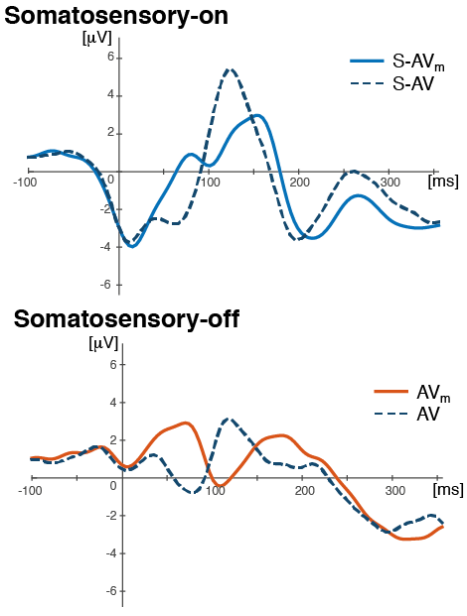

Figure 2: *Audio-visual event-related potentials.*

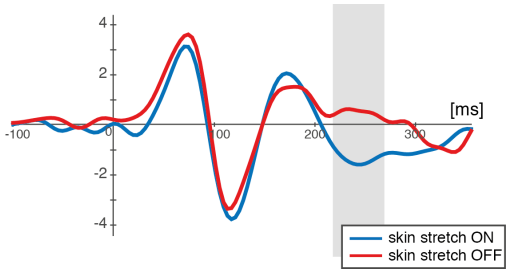

Figure 3: *Contrast of audio-visual ERP between congruent and incongruent conditions*

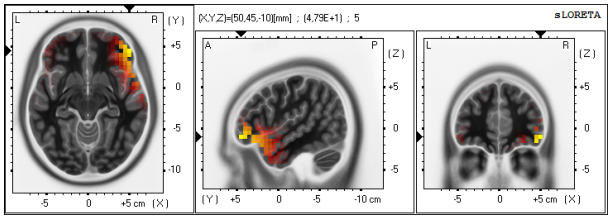

Figure 4: Estimated source location for the change of audio-visual ERP due to somatosensory input

In order to clarify potential changes in the audio-visual interaction to the somatosensory stimulation, we subtracted the congruent ERP from the incongruent ERP (Figure 3). In these subtracted ERPs, the first positive peak appeared around 80 ms after the auditory onset, the negative peak follows around 110 ms and then the second positive peak was around 170 ms. The two subtracted potentials were similar until 200 ms, and then they diverged. Cluster-based permutation analysis indicated that the divergence due to somatosensory stimulation (the gray region in Figure 3) was associated with a significant change in the audio-visual ERP. This indicates the somatosensory effect to AV responses appeared only in the later periods of response (> 200 ms), but not in the earlier periods (< 200 ms).

We applied source localization to the difference in the potentials and found the estimated source to be within the right inferior frontal gyrus (BA47) as shown in Figure 4.

Next, we examined the change in the ERPs of auditory or audio-visual processing due to somatosensory inputs by examining a difference between recorded ERPs and ERPs that was summed ERP in the somatosensory alone condition with the ERPs in the audio (A) or audio-visual conditions (AV or AVm). We found an ERP change due to somatosensory stimulation in all three condition (Figure 5). Like the previous analysis, the change is similar in all three condition up to 220 ms after auditory onset, indicating that somatosensory inputs affected the ERP equally in all three conditions. In the periods between 220-275 ms (gray area in Figure 5), the ERPs in the AV condition was different from the other two conditions. The amplitude of ERPs in this period was used in the correlation analysis mentioned below.

Behavioral analysis showed clearly the effect of the incongruent stimuli to a change in perception [$F(2,23) = 96.62$, $p < 0.001$]. The left panel in Figure 6 shows the judgement probability that the subject identified the presented sound as /ba/. Congruent pairs of audio-visual stimulation (AV and S-AV) showed that the judgement probability was close to 1 indicating that the subject perceived the sound as /ba/. On the contrary, incongruent pairs (AVm and S-AVm) showed that the subject perceived the sound as not /ba/ since the judgement

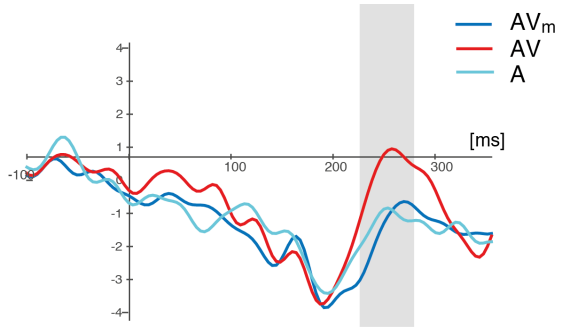

Figure 5: ERPs due to somatosensory interaction in auditory and audio-visual processing.

probability was close to 0. In the auditory alone condition (A and S-A), the subject's judgement was less than 1 with large variance representing a slightly degraded auditory signal that allowed for visual information to facilitate syllable identification.

We also found that the judgement probability was greater when somatosensory stimulation was applied. This change was observed consistently in all three auditory conditions: congruent AV, incongruent AV (AVm) and auditory alone (A) [$F(1,14)= 8.476$, $p < 0.02$]. This indicates that somatosensory inputs did not work as a disturbance of speech perception, rather somatosensory stimulation biased consistently the subject's judgement toward identifying the /ba/ regardless of AV congruency. Note we found no interaction between AV congruency and somatosensory effect.

A correlation analysis was carried out between the behavioral response and the ERP change due to somatosensory stimulation. Here we used the change in judgement probabilities due to somatosensory input (difference between white and gray bar in the left panel of Figure 6) and the ERP change due to somatosensory stimulation (ERP amplitude in gray area of Figure 5). These two variables were modestly correlated ($r= 0.50$; $p = 0.058$) suggesting that the magnitude of change in ERPs is related to the addition of somatosensory stimulation during speech perception.

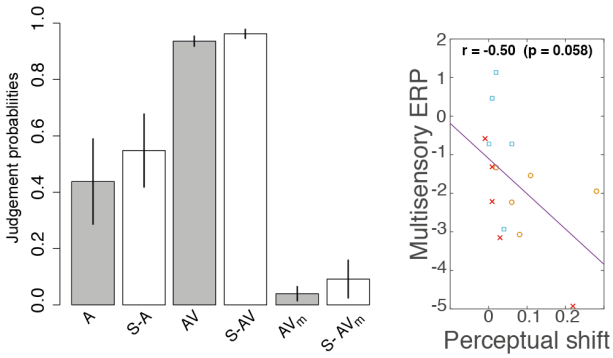

Figure 6: Behavioral response and correlation between cortical potential change and perceptual change

4. Discussion

We found that somatosensory stimulation during audio-visual (AV) speech perception resulted in ERPs changes in the period between 220-275 ms after auditory onsets. Source localization analysis indicated a right inferior frontal gyrus (IFG) region as a possible source of the somatosensory effect. In the behavioral response, we found that somatosensory inputs slightly modified the subject's perceptual judgement when presented with incongruent A-V stimulation. The change in the behavioral measure was modestly correlated with ERP changes due to somatosensory inputs, suggesting clear multisensory convergence in sensory signals for audio-visual speech perception.

In contrast to the somatosensory interaction during audio-visual processing, the influence of each individual sensory input (visual and somatosensory) to speech sound processing was found earlier (< 220 ms after auditory onsets). This suggests that somatosensory-auditory-visual interactions may be processed sequentially with individual somatosensory and visual signals interacted with auditory separately and then integrated after.

One possible mechanism of the McGurk effect may involve the extraction of movement information from visual inputs. The incongruent visual stimulation simply works to conflict or disturb auditory perception due to the perception of a different speech articulatory movement from vision. In contrast, somatosensory inputs provide congruent information concerning speech articulatory motion, facilitating perception and generating a more being congruent with the movement inferred from the acoustic signal. Since multisensory ERP were changed by somatosensory inputs when the sound was perceived as different (the ambiguous or incongruent /ba/) , the ERP changes may be due to a conflict between congruent and incongruent information for speech sound processing.

Our source localization analysis showed that right inferior frontal gyrus (IFG) may be involved in the somatosensory interaction with audio-visual speech perception. A previous fMRI study demonstrated that right IFG was associated with the congruent-incongruent contrast during AV speech processing [12, 13]. In addition, this area has also generated larger activation in an identification task with both vibrotactile and sound stimulations [14] suggesting this area may be involved in multisensory processing of stimulus identity. For the current study somatosensory inputs appear to induce additional activation in the right IFG reflecting multisensory processing.

5. Conclusions

Somatosensory stimulation affected perceptual judgements and also induced a change in ERP for audio-visual speech processing. The effect of somatosensory input on evoked cortical potential change and the associated behavioral response was moderately correlated suggesting a functional coupling among the sensory systems. The results demonstrate a multisensory convergence between somatosensory and audio-visual processing. We here tested only a limited number of the subjects. Further investigation is required to validate the current findings.

6. Acknowledgements

This work was supported by grants, R21DC012502 from National Institute on Deafness and Other Communication Disorders, NIH.

7. References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature,* vol. 264, pp. 746-8, 1976.

[2] W. H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise," *J Acoust Soc Am,* vol. 26, pp. 212-15, 1954.

[3] B. Gick and D. Derrick, "Aero-tactile integration in speech perception," *Nature,* vol. 462, pp. 502-4, Nov 2009.

[4] T. Ito, M. Tiede, and D. J. Ostry, "Somatosensory function in speech perception," *Proc Natl Acad Sci U S A,* vol. 106, pp. 1245-8, Jan 2009.

[5] T. Ito, D. J. Ostry, and V. L. Gracco, "Somatosensory event-related potentials from orofacial skin stretch stimulation," *J Vis Exp,* p. e53621, 2015.

[6] N. P. Connor and J. H. Abbs, "Orofacial proprioception: analyses of cutaneous mechanoreceptor population properties using artificial neural networks," *J Commun Disord,* vol. 31, pp. 535-42; 553, Nov-Dec 1998.

[7] T. Ito and D. J. Ostry, "Somatosensory contribution to motor learning due to facial skin deformation," *J Neurophysiol,* vol. 104, pp. 1230-8, Sep 2010.

[8] R. S. Johansson, M. Trulsson, K. Â. Olsson, and J. H. Abbs, "Mechanoreceptive afferent activity in the infraorbital nerve in man during speech and chewing movements," *Exp Brain Res,* vol. 72, pp. 209-14, 1988.

[9] D. M. Groppe, T. P. Urbach, and M. Kutas, "Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review," *Psychophysiology,* vol. 48, pp. 1711-25, Dec 2011.

[10] R. D. Pascual-Marqui, "Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details," *Methods Find Exp Clin Pharmacol,* vol. 24 Suppl D, pp. 5-12, 2002.

[11] G. A. Calvert, P. C. Hansen, S. D. Iversen, and M. J. Brammer, "Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect," *Neuroimage,* vol. 14, pp. 427-38, Aug 2001.

[12] M. Olivetti Belardinelli, C. Sestieri, R. Di Matteo, F. Delogu, C. Del Gratta, A. Ferretti*, et al.*, "Audio-visual crossmodal interactions in environmental perception: an fMRI investigation," *Cognitive Processing,* vol. 5, pp. 167-174, 2004.

[13] N. M. van Atteveldt, E. Formisano, R. Goebel, and L. Blomert, "Top-down task effects overrule automatic multisensory responses to letter-sound pairs in auditory association cortex," *Neuroimage,* vol. 36, pp. 1345-60, Jul 15 2007.

[14] L. A. Renier, I. Anurova, A. G. De Volder, S. Carlson, J. VanMeter, and J. P. Rauschecker, "Multisensory integration of sounds and vibrotactile stimuli in processing streams for "what" and "where"," *J Neurosci,* vol. 29, pp. 10950-60, Sep 2 2009.