SENSORY INTEGRATION FROM AN IMPOSSIBLE SOURCE: PERCEIVING SIMULATED FACES

Megan Keough [*1], Ryan C. Taylor [†1], Donald Derrick [‡2], Murray Schellenberg [♯1] and Bryan Gick [◊1,3]
[1]University of British Columbia, Vancouver, British Columbia, Canada.
[2]University of Canterbury, Christchurch, Canterbury, New Zealand.
[3]Haskins Laboratories, New Haven, Connecticut, USA.

1 Introduction

Recent research has shown that aero-tactile cues influence speech perception even without the presence of an acoustic signal [1]); when participants viewed a silent bilabial articulation that co-occurred with a puff of air felt on the skin, they were significantly more likely to perceive it as aspirated. These results and others [2] suggest that this integration is relatively automatic, enough so that it occurs in the absence of an interlocutor who could be the airflow source. However, it may be that perceivers are willing to extend physical capabilities to these non-present sources because they are human and therefore possible sources of the aero-tactile cue.

The results from [1] established that participants integrate aero-tactile information that is presented alongside videos of real people. The current study examines whether aero-tactile information presented synchronously with visual speech information from an impossible source – a computer-animated face on a computer monitor – can affect perception of consonants. Unlike a human, a computer's means of producing sound should not be expected to produce a puff of air in the real world. However, we predict that artificial air flow will be perceived as speech aspiration when paired with a computer-generated avatar. Based on the findings in [1], we predict that participants will demonstrate a baseline /ba/ bias in trials without airflow. Evidence of integration from an impossible source would support the idea that visual-tactile integration is an automatic process that occurs even in the absence of an interlocutor capable of producing the stimuli.

2 Method

11 Native English speakers were recruited from the University of British Columbia Linguistics Department subject pool and were compensated $5 for a thirty-minute session. All participants reported no speech or hearing disorders. Participants were seated in a sound booth in a high-backed chair and were instructed to keep their back against the chair as much as possible during the study. They were shown an animated video of a computer-animated head producing a bilabial plosive while listening to multi-talker babble noise through headphones. Some of the presentations were accompanied by a light, synchronous puff of air on the neck. Participants were asked to indicate what syllable they thought the avatar had produced (i.e., pa or ba) using the keyboard.Response keys were counterbalanced across participants.

Figure 1: Computer generated avatar used in the study thought the avatar had produced (i.e., pa or ba) using the keyboard. Response keys were counterbalanced across participants.

A two-dimensional female avatar (see Figure 1) was created using computer software (CrazyTalk 8). A single video clip was then made of the avatar producing a bilabial plosive. An accompanying sound file was created using the software's TTS feature and the avatar's stop closure and release were synchronized with the closure and burst in the waveform. The clip was then exported to a QuickTime file. This initial clip was used for the trials that did not have accompanying airflow. To create the video clip for the puff condition, the audio was extracted from the video clip and a 50 ms 10 kHz sine wave was inserted in the left channel. To account for system latency, the tone was placed 35 ms earlier than the stop burst. This ensured that the visible release of the bilabial closure and the release of the airflow from the tube would be synchronous. The left channel of the sound file was then extracted and recombined with the QuickTime file to create a silent video clip that would trigger a synchronous air puff.

The puff was created using a California Air Tools 4610 air compressor connected to a switchbox via a ¼ inch diameter vinyl tube all located outside the sound booth. A second ¼ inch vinyl tube passed from the switch box through the access port of the sound booth and was attached to a flexible boom arm fitted to a microphone stand. The open end of the tube was positioned ~7 cm in front of the participant's suprasternal notch. Using Direct Sound EX-29

[*] megan_keough@alumni.brown.edu
[†] taylryan@gmail.com
[‡] donald.derrick@gmail.com
[♯] mhschellenberg@gmail.com
[◊] bryan.gick@mail.ubc.ca

headphones, participants listened to a babble track played from a separate computer located outside the sound booth. The experiment was run using PsychoPy [3] on an iMac computer directly in front of the participant.

3 Results

Statistical analysis employed a Generalized Additive Mixed- Effects Model in R [4], following the formula:

Response ~ Condition + s(Trial Number, Participant, bs = "fs", m = 1)

Where *Response* is coded 0 for /ba/ and 1 for /pa/, *Trial Number* is the ordered position of the individual token presentation in time, and *Condition* is "puff" or "no puff". The first term *Condition* is the fixed term. The second term *s(Trial Number, Participant, bs = "fs", m = 1)* is the random effect of trial order by participant. The fixed term (*Condition*) is significant (z-value 2.01, $p = 0.044$. The random effect of trial order is highly significant (Chi Sq = 141.8, $p < 0.001$). In the puff condition, participants responded /pa/ in 76% (SE = 0.06) of the trials. In contrast, when there was no puff, participants reported seeing /pa/ 34% (SE = 0.08) of the time. The results have an adjusted R-squared of 0.974, accounting for 95.6% of the deviance. Almost all variance is accounted for by the effect of trial order, as seen in Figure 2. The results show that participants mostly answered that they perceived /ba/ at the beginning of the experiment, yet during the experiment they eventually all reported perceiving /pa/. In the initial state, the bias toward /b/ is below -4. There is a second phase, where the propensity to answer /p/ rapidly increases the change of state. In the final state, the bias remains above 4. This sequence of events occurred for all participants. The start of the state change varied considerably between participants although the slope of change was very similar.

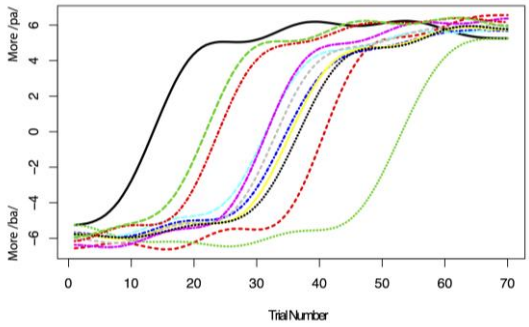

Figure 2: Effects of trial order on participant answer.

4 Discussion and conclusions

The current study tested the hypothesis that perceivers can integrate aero-tactile speech information from an impossible source. It was predicted that participants would provide more /pa/ responses when they are presented with both visual and aero-tactile stimuli than when presented with only the visual stimulus. Condition emerged as a significant factor such that participants reported significantly more /pa/ responses during trials in which they felt synchronous airflow. However, Condition effect is overshadowed by the large effect of Trial Number. As evidenced in Figure 2, participants began the task with a /ba/ bias unrelated to Condition. By the second half of the experiment, however, they exhibit a /pa/ bias. This result is markedly different from the response biases found in [1]. There, the authors reported a /ba/ bias in the Visual-only condition and found no effect of trial order on response. This suggests that while participants in the current study were indeed integrating the aero-tactile speech information from a computer-generated source, they did so differently from participants in [1], who observed real-life productions of the syllables.

One possible explanation is that the effect of trial order is related to the nature of the visual stimulus. Participants were presented with a single visual token, a fact they appear to have noticed. They initially experienced this articulation as a /ba/, reflecting the expected bias [1]. Then, as they experienced more trials that included airflow (which felt more /pa/-like), they may have begun assuming that this single articulation was /pa/. Thus, the aero-tactile stimulus appears to have caused them to associate the video of the avatar with a /p/ rather than a /b/, so that they effectively learned the articulation was a /p/.

While /b/ and /p/ have traditionally been considered a single viseme [5] and thus visually indistinguishable, our findings support those of [6] and [1], whose work suggest a visual distinction between the two sounds. If /b/ and /p/ were in fact visually identical articulations, we would predict a replication of the findings in [1]. Instead, our findings support the idea that there are subtle visual differences between the two articulations that perceivers are sensitive enough to detect. This study showed that aero-tactile stimuli alone was sufficient to cause participants to associate the simulated face video with /p/ within seventy trials. Experience alone is sufficient to train listeners to distinguish these "sounds".

Acknowledgments

This work was supported by an NSERC Discovery Grant (RGPIN-2015-05099) awarded to the fifth author.

References

[1] K. Bicevskis, D. Derrick, and B. Gick.. Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives. *JASA*, 140:5, 2016.

[2] B. Gick, D. Derrick. Aero-tactile integration in speech perception. Nature. 462(7272):502, 2009.

[3] J.W. Peirce. Generating stimuli for neuroscience using PsychoPy. *Front*. *Neuroinformat*. 2:10, 2009.

[4] R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria <https://www.R-project.org/>, 2016.

[5] C. G. Fisher. Confusions among visually perceived consonants. JSHR, 11 :4, 1968.

[6] J. Abel, A. V. Barbosa, A. Black, C. Mayer, and E. Vatikiotis-Bateson. The labial viseme reconsidered : Evidence from production and perception. JASA, 129:4, 2011.