LEARNING EFFECTS IN MULTIMODAL PERCEPTION WITH REAL AND SIMULATED FACES

Megan Keough[1, 2], Donald Derrick[3, 4], Ryan C. Taylor[1], Bryan Gick[1, 2]

[1]University of British Columbia, [2]Haskins Laboratories, [3]New Zealand Institute of Language, Brain, and Behaviour, [4]Marcs Institute

[1]megan.keough@ubc.ca, [2]donald.derrick@canterbury.ac.nz, [3]rctaylor@mail.ubc.ca, [4]gick@ mail.ubc.ca

ABSTRACT

We have all learned to associate real voices with animated faces since childhood. Researchers use this association, employing virtual faces in audiovisual speech perception tasks. However, we do not know if perceivers treat those virtual faces the same as real faces, or if instead integration of speech cues from new virtual faces must be learned at the time of contact. We test this possibility using speech information that perceivers have never had a chance to associate with simulated faces – aerotactile somatosensation. With human faces, silent bilabial articulations ("ba" and "pa"), accompanied by synchronous cutaneous airflow, shift perceptual bias towards "pa". If visual-tactile integration is unaffected by the visual stimuli's ecological origin, results with virtual faces should be similar. Contra previous reports [8], our results show perceivers do treat computer-generated faces and human faces in a similar fashion - visually aligned cutaneous airflow shifts perceptual bias towards "pa" equally well with virtual and real faces.

Keywords: Speech Perception, Speech Acoustics, Multimodal Phonetics

1. INTRODUCTION

We encounter digitally rendered faces often in our everyday lives through animated films, video games, and even virtual reality. More to the point, we readily associate those faces with human voices. Many audiovisual (AV) speech perception studies make use of this perceptual flexibility and present the task's visual information with a digitally rendered face instead of a real one (e.g., [9]). Animated faces provide the obvious benefit of providing much more fine-grained control when making stimuli.

However, it remains unknown whether perceivers process information from an animated or simulated face as they do from a real, human face. This raises the important question of whether the results obtained in AV tasks with simulated faces truly reflect the same integration processes that occur in real-world interactions. Despite contemporary people's extensive experience of animated faces paired with human voices, it remains possible that the integration found in AV studies reflects an association between auditory speech information and a non-human source learned at the time of the experiment. The current paper tests this possibility with a series of studies using speech information that perceivers have no experience associating with an animated face – aerotactile somatosensation.

1.1. Background

While we may not be consciously aware of the tiny bursts of airflow emitted during the production of some sounds (e.g., aspirated stops), airflow is one of many somatosensory inputs our brain receives while we talk. Air not only flows across our speech articulators and potentially our extremities, but we may also feel the airflow of others when speaking in close proximity. Previous research has demonstrated that the sensation of this speech-related airflow across the skin influences stop consonant perception, pushing the perceiver's percept toward an aspirated token. This aspiration effect occurs both when the airflow is paired with an audio signal [6] and with silent videos of a person producing bilabial-initial syllables (i.e., /pa/ and /ba/) [3]. These results further suggest that integration is automatic enough to occur in the absence of a possible interlocutor.

However, those results [6, 3] may instead show that perceivers extend physical capabilities to a non-present source when the source is human and therefore physically capable of producing the aerotactile information. This raises the question of how perceivers would treat a synthetic source that is not physically capable of producing airflow—a digitally rendered speaker. Unlike a human, a computer's means of producing sound should not be expected to produce a puff of air in the real world. Yet if the behavioral evidence from the AV literature reflects an automatic integration process unaffected by the source's ecological validity, perceivers should show no decrease in the aspiration effect described above. Given the evidence from audio, visual, and aerotactile studies described above, we make two predictions:

(1) Regardless of whether the visual source is from a natural or artificial face, we predict a

slight "ba" bias for tokens presented without air flow, and a slight "pa" bias for tokens presented with airflow.

(2) There will be no difference in integration between participants who view a computer-generated face and participants who view a human face.

2. METHODS

2.1. Visual Stimuli

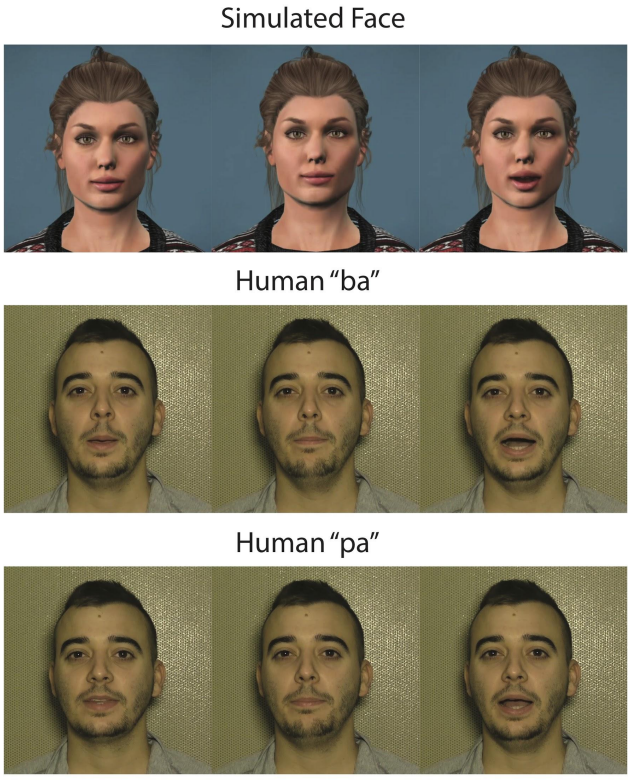
The visual stimuli for the experiment include three sources: 1) Simulated face, 2) human "ba", and 3) human "pa". The *simulated face* stimuli were generated from a two-dimensional female avatar created using the CrazyTalk computer animation software [4]. A single video clip of the face producing a bilabial plosive followed by a low back vowel (e.g., /ba/) was then generated. CrazyTalk's text-to-speech feature was used to generate a sound file that was then synchronized with the avatar's articulation of the syllable by aligning the stop burst in the audio file with the release of the bilabial closure. The resulting clip was then exported to a QuickTime video file.

The *human "pa"* and *human "ba"* stimuli are each a single video of the same human speaker. The "pa" is the speaker producing a voiceless *aspirated* bilabial stop followed by a low back vowel (/pa/ or [p^ha]). The "ba" is the same speaker producing a voiceless *unaspirated* bilabial stop followed by a back vowel (/ba/ or [pa]). For both the human "ba" and the human "pa" the two video clips were obtained from the no-lag condition of Bicevskis et al. [3].

2.2. Aero-tactile Stimuli

For each of the three videos (simulated face, "ba", and "pa") the audio from the video clip was extracted. The resulting sound file was split into a stereo track and a 50 millisecond (ms) 10 kHz sine wave was inserted in the left channel using Audacity [1]. The voltage from the sine wave was used as a trigger to release a gentle puff of air (~ 7 psi) from a California Air Tools 4610 air compressor, itself located outside of the sound-attenuated booth used for the following experiments.

Figure 1: Still shots from the video stimuli. Top = simulated face, middle = human "ba", bottom = human "pa". Left = open mouth prior to plosive production. Middle = compress mouth prior to plosive release. Right = open mouth during vowel production following plosive release.



The sine wave ended 35 ms before the stop burst to account for system latency. This ensured that the puff of air exiting the tube and the release of the bilabial closure would be synchronous.

The left channel of the sound file (with the tone) was then extracted and recombined with the original video clips to create a silent video clip that triggered a simultaneous puff of air. For the no puff condition, the left channel was left empty so as to avoid activating the airflow system. For all conditions, the right channel was not connected to any playback system, so no audio was produced. The airflow was delivered through a ¼ inch vinyl tube running from the compressor by way of a specially designed switch box. The opening of the tube was located 7 cm in front of the participant's suprasternal notch of each participant.

2.3. Procedure

Fifty-six native English speakers, mean age = 21.45 (SD = 5.5), 31 female and 25 male, were recruited from campus and compensated for a fifteen-minute session. Participants provided informed consent, and the experiments were approved by the

University of British Columbia's human research ethics committee. All participants completed a language background questionnaire and reported no speech or hearing difficulties. Before the task, participants were informed that they may feel air during some trials but were given no other instructions regarding the air. All participants were tested in a sound-attenuated booth and instructed to keep their head and back against a high-backed chair.

Participants were assigned to one of four groups: 1) 11 to simulated face, 2) 17 to human "ba", 3) 16 human "pa", and 4) 12 simulated face without puff. The experiments were run on an iMac computer using PsychoPy [10]. Participants viewed repetitions of the relevant single silent video of a bilabial-initial syllable. Following Bicevksis et al.'s standard [3], multitalker babble played through Direct Sound Ex-29 headphones from a second computer located outside the sound booth. For groups 1–3, half of the videos were presented with a synchronized puff of air on the participant's neck, and the other half were not. Participants in the simulated face/no puff group never felt airflow. For each trial, participants were asked to indicate on a keyboard which syllable ("pa" or "ba") they felt the talker in the video had said. Trial order was randomized and the response keys were counterbalanced.

2.4. Statistical analysis

Interaction effects were visualized in R [11] using ggplot2 [12]. In addition, Generalized linear mixed-effects models (GLMM) were run on the interactions between trial order, condition (visual vs. visual + tactile), and visual stimuli type (simulated face, human "ba" and human "pa") for both the response, and the response times. Model fitting was applied in a stepwise backwards iterative fashion, and models were back-fit using the Akaike information criterion (AIC) in order to measure the quality of fit. This technique identifies the best fit for the data, allowing elimination of interactions in a statistically appropriate manner. The two final models can be seen in Formulas (1) and (2).

(1) response ~ condition
+ (1 + condition | participant)

Where *response* is a numerical value, 1 for "pa" and 0 for "ba", and *condition* is one of audio, or audio + tactile.

(2) log(response time) ~ trial order
+ (1 + trial order | participant)

Where log(response time) is the log of the response time, and trial order is the order in which the individual 2AFC question was asked (1 through 70).

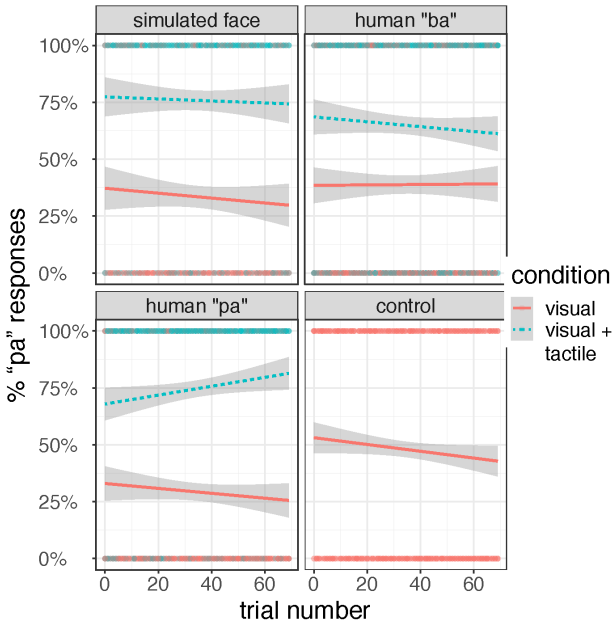
Note the simplicity of these models: other interactions or main effects terms were either not significant, failed to converge, or failed to show significance during back-fit analysis.

3. RESULTS

Participant responses to visual only (no puff) and visual + tactile (puff) stimuli, by trial number, are presented in Figure 2. They show that for the simulated face and the two human faces, participants have a response bias shift towards "pa" for the visual + tactile condition as compared to the visual only condition. However, while there is some variation across the groups, it is apparent in Figure 2 that there was no significant interaction between trial order and visual stimulus type. In other words, participants did not significantly shift their response behavior over the course of the experiment for any of the visual stimuli types.

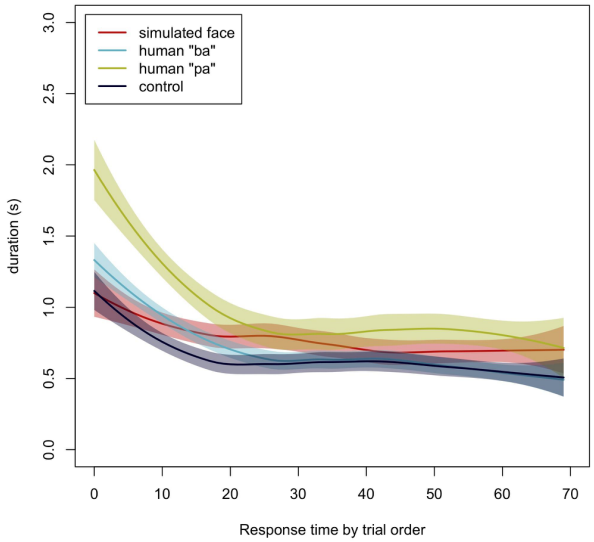
The best-fit (back-fit) generalized linear mixed-effects model comparing responses shows that the only significant variable was condition (visual only vs. visual + tactile, t-value = 6.527, $p < 0.001$). Trial order and visual stimuli type were not significant, and did not have a significant impact on the maximum-likelihood model fit.

Figure 2: Percent "pa" responses by visual stimuli and trial order. Note that for the control experiment, the only condition was "visual"; participants did not feel airflow at any point during the experiment.



As seen in Figure 3, response times sped up throughout the experiment. The statistical analysis revealed that the trial order effect is significant (t-value = -6.829, $p < 0.001$). However, while participants in the human "pa" group were slower to respond at the start of the experiment, the interaction between trial order and visual stimuli type did not emerge as significant during back-fit analysis.

Figure 3: Response times by visual stimuli type and trial order. Thickness of shaded lines based on confidence intervals from locally weighted polynomial regression estimations (LOESS).



4. DISCUSSION

Our results clearly indicate no difference between the simulated face, human "pa", human "ba", and control groups, suggesting that perceivers treat real and simulated faces equivalently for the purposes of integration. While we had previously reported [8] significantly different patterns of integration for simulated faces, the current analysis contradicts those results. The lack of interaction between the trial order and the visual stimulus type shows that the participants were not learning to associate the aerotactile speech information and the simulated face over the course of the experiment; instead, the cross-modal information was integrated from the start.

Our results are in line with several cross-modal studies showing that perceivers may not be looking for an ecologically valid, localized source in their immediate environment when integrating cross-modal cues. For example, the source of the visual information in an audiovisual task need not appear to coincide in space with the source of the auditory information [7, 2, 5]. Such findings are remarkable given that stimuli coming from opposite directions are unlikely to originate from the same natural source.

Our results add to these findings showing that perceivers are not overly concerned with whether speech information originates from an ecologically valid source. Instead, they appear to integrate signal-relevant and synchronous speech information regardless of whether the apparent source is real — no real-time learning is required to accomplish this task.

Acknowledgements
We would like to thank Esther YT Wong, Sharon Kwan, and Terrina Chan for their invaluable assistance running subjects. Funding was provided by NIH Grant DC-002717 to Haskins Laboratories.

5. REFERENCES

[1] Audacity Team. 2018. *Audacity (Version 2.3)*. <https://www.audacityteam.org/>
[2] Bertelson, P., Vroomen, J., Wiegeraad, G., De Gelder, B. 1994. Exploring the relation between McGurk interference and ventriloquism, In *Third International Conference on Spoken Language Processing,* 559–562.
[3] Bicevskis, K., Derrick, D., Gick, B. 2016. Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives. *Journal of the Acoustical Society of America,* 140(5), 3531–3539.
[4] CrazyTalk Core team. 2018. CrazyTalk 8. <http://cdn.reallusion.com/help/eng/ct/CrazyTalk_8_User_Manual.pdf>
[5] Fisher, B.D., Pylyshyn, Z. W. 1994. The cognitive architecture of bimodal event perception: a commentary and addendum to Radeau. *Current Psychology of Cognition*, 92–96.
[6] Gick, B., Derrick, D. 2009. Aero-tactile integration in speech perception. *Nature*, 462(7272), 502–504.
[7] Jones, J. A., & Munhall, K. G. 1996. Spatial and temporal influences on audiovisual speech perception. In *International Journal of Psychology*, 31(3–4), 4734-4734.
[8] Keough, M., Taylor, R. C., Derrick, D., Schellenberg, M. Gick, B. 2017. Sensory Integration from an Impossible Source: Perceiving Simulated Faces, *Canadian Acoustics*, 45, 176–177.
[9] Massaro, D. W., & Cohen, M. M. 1990. Perception of synthesized audible and visible speech. *Psychological Science*, 1(1), 55–63.
[10] Peirce, J. W. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformation*, 2(10), 1–8.
[11] R Core Team. R (2016). *A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, Vienna, Austria <https://www.R-project.org/>.
[12] Wickham, H. 2016. ggplot2: *Elegant Graphics for Data Analysis*. Springer-Verlag, New York.