PERCEIVING VISIBLE SPEECH ARTICULATIONS IN VIRTUAL REALITY

Megan Keough [*1, 2], Ryan C. Taylor [†1], Esther Y. T. Wong [‡1], Dimitri Prica [♦1], Murray Schellenberg [♯1] and Bryan Gick [▼1, 2]
[1]Department of Linguistics, University of British Columbia, Vancouver, British Columbia, Canada
[2]Haskins Laboratories, New Haven, Connecticut, USA

1 Introduction

Advances in virtual reality (VR) and avatar technologies have created new platforms for face-to-face communication in which visual speech information is presented through avatars using simulated articulatory movements. These movements are typically generated in real time by algorithmic response to acoustic parameters. While the communicative experience in VR has become increasingly realistic, the visual speech articulations remain intentionally imperfect and focused on synchrony to avoid uncanny valley effects [1]. While considerable previous research has demonstrated that listeners can incorporate visual speech information produced by computer-simulated faces with precise and pre-programmed articulations [2], it is unknown whether perceivers can make use of such underspecified and at times misleading simulated visual cues to speech.

The current study investigates whether reliable segmental information can be extracted from visual speech algorithmically-generated through a popular VR platform. We focused on the platform's most consistent and easily perceived articulator movements: bilabial closure in consonants; and lip rounding, lip spreading, and jaw lowering in vowels (see Figure 1). We report on an experiment using a speech-in-noise task with audiovisual stimuli in two conditions (with articulator movement and without) to ask the following questions: 1) whether the visual information from an avatar improves identification of target words, and 2) whether that visual information improves categorization of the target segment.

1 Methods

19 native English speakers (ages 18-30) were recruited from the University of British Columbia Linguistics subject pool. An additional 10 non-native participants are excluded from analysis. Stimuli consisted of videos of an avatar saying simple sentences ("It's [TARGET]") captured in Facebook Spaces™ with Oculus Rift™ hardware. The stimuli were recorded using the in-app video capture feature which records both the avatar movement and the user's speech. Articulator movement was generated automatically through the app. For stimuli without lip movement, the Facebook Spaces microphone was disabled to prevent audio pick up and articulator movement. For all recordings, a simultaneous audio recording was made [Samson C03U mic] and dubbed into all videos using Kdenlive [3] and Final Cut Pro X [4].Consonant targets followed a 2x2x3 paradigm: articulator movement (with or without) x segment (bilabial or not), x minimal pair (3 pairs). Vowel targets followed a 2x3x2 paradigm: articulator movement (with or without) x segment ([i] [u] [ɑ]) x minimal triplet (2 triplets).

Figure 1: Samples of stimuli videos with targets [u], [i] and [ɑ].

Stimuli were presented on an iMac 2017 computer using OpenSesame 3.2.4 [5] and AKG K240 headphones in a sound-attenuated booth. Stimuli were randomized within a single block. The experiment consisted of 6 blocks (144 tokens in total), with breaks between blocks. A "babble" track was simultaneously presented though Audacity® [6] for the duration of the experiment. Participants were told to imagine that the avatar was giving answers to a crossword puzzle they were solving in a crowded cafe. After each video, participants were asked to type the word they had heard. The signal-to-noise ratio was calibrated empirically during a pilot study to achieve a 40% success rate for two native English speakers.

2 Results

One participant was excluded for not completing the task, leaving 18 subjects for data analysis. Our first question concerned whether visible articulator movement enhanced identification of the target word. Mean accuracy for the Articulator Movement condition was 8% higher than in the No Articulator Movement condition (35.7% vs. 27.6%, respectively) suggesting a small improvement with the addition of articulator information.

To answer our second question, we calculated accuracy rates for target consonant and vowel categorization. For consonants, accurate categorization was defined as responding with a bilabial-initial word when the target word was bilabial (e.g., initial /p/, /b/, or /m/ if the target word was *bit*) and responding with a non-bilabial initial segment when the target was not bilabial-initial (e.g., initial /h/ or /k/ when the target word was *hit*). For vowels, accurate

[*] keoughm@mail.ubc.ca
[†] taylryan@gmail.com
[‡] esther_wong_yt@hotmail.com
[♦] dprica@alumni.ubc.ca
[♯] mhschellenberg@gmail.com
[▼] gick@mail.ubc.ca

categorization was defined as responding with a rounded vowel when the target word contained [u] (e.g., *pool*); a high front vowel (either [i] or [ɪ]) when the target word contained [i] (e.g., *peek*); and a low back vowel when the target word contain [ɑ] (e.g., *Paul*). The data were analyzed using linear mixed-effects models in R [7] with the lme4 [8] and lmerTest packages [9].

For the bilabial target words, participants showed a small *decrease* in categorization accuracy when visible articulator movement was available (-5%). In contrast, participants showed a 20% increase in accuracy for non-bilabial targets in the Articulator Movement condition, suggesting that participants became better at identifying something as not bilabial with the addition of visible articulatory information (see Figure 2). Results from a linear mixed effects model[1] show a significant interaction between Consonant and Articulatory movement (β = -0.25, *SE* = 0.05, *t* = -5.19, $p < 0.001$) such that participants were significantly better at categorizing non-bilabial target words when presented with visible articulator movement.

For vowel target words, participants showed small enhancement effects in the Articulator Movement condition for [u] and [ɑ] target words (4% and 5%, respectively), but a large increase in accuracy for [i] target words. A linear mixed effects model[2] revealed a significant effect of [i] (β = -0.18, *SE* = 0.06, *t* = -3.25, $p < 0.01$) such that participants were worse at categorizing [i] target words overall. In addition, a significant interaction between vowel [i] and Articulator Movement emerged (β= -0.24, *SE* = 0.06, *t* = -3.96, $p < 0.001$) supporting the observation that [i] categorization was enhanced by visible lip spreading.

Discussion and conclusions

The results suggest that even imperfect articulator movement from an avatar improves speech perception to some extent. However, the results also show that the imprecise articulatory movements were not as informative as those from a human source or a pre-programmed synthesized face. In particular, we observed that while a visible articulation of [i] significantly improved segmental categorization, visible articulation of [u] or [ɑ] did not. Perhaps most unintuitively, visible articulation did not improve accuracy of perception of bilabial-initial words, even though the lip movement was readily apparent. In contrast, articulatory movement enhanced categorization of *non*-bilabial sounds.

The avatar's simulated bilabial closure movement was very brief and lacked visible lip compression; the verisimilitude of this articulation was insufficient to aid perception and categorization of bilabials. Further fine-grained perceptual studies are needed to determine the balance of realism and abstraction to optimize perception, and thus successful and naturalistic avatar communication, without increasing signal lag.

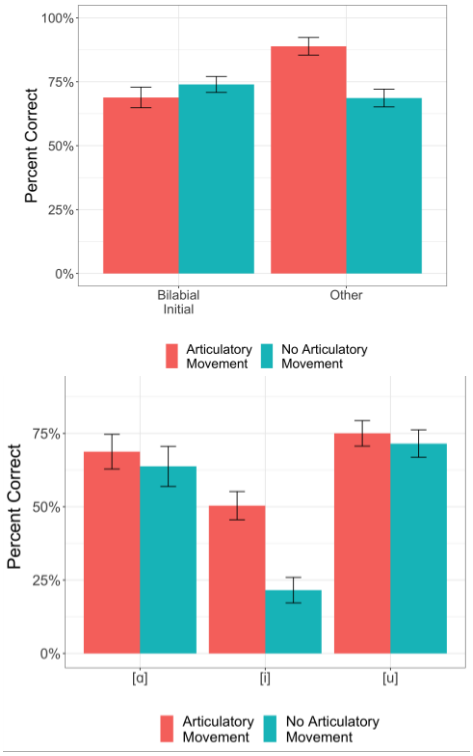

Figure 2: Mean accuracy for consonant (top) and vowel (bottom) target words for both conditions.

Acknowledgments

Research funded by the UBC Hampton Fund and by NIH Grant DC-002717 to Haskins Laboratories.

References

[1] Facebook for Developers - F8 2017 Keynote." Facebook for Developers. https://developers.facebook.com/videos/f8-2017/f8-2017-keynote/. Accessed October 28, 2018.

[2] Cohen, M. & Massaro, D. Synthesis of Visible Speech. *Beh. Res. Meth. Instr. & Comp. 22*(2), 260-263, 1990.

[3] Kdenlive. 18.04.1. May 11, 2018. J. Wood. USA

[4] Final Cut Pro X. 10.1.2. June 27, 2014. Apple Inc.. USA

[5] Mathôt, S., Schreij, D. & Theeuwes, J. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Beh. Res. Meth.*, 44(2), 314-324, 2012.

[6] Audacity Team (2012) Audacity®. Version 2.0.0. Audio editor and recorder. http://audacityteam.org/. Accessed 26/04/2012.

[7] R Core Team. R: A language and environment for statistical computing. R Found. for Stat Comp, 2013. Vienna, Austria. http://www.R-project.org/.

[8] Bates, D., Maechler, M., Bolker, B. & Walker, S. lme4: Linear mixed-effects models using Eigen and S4. R package, version 1(7): 1-23, 2014.

[9] Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. lmerTest package: tests in linear mixed effects models. J. Stat. Soft. 82(13), 2017.

[1] *Accuracy ~ Consonant * Articulatory_movement + (1 + Consonant * Articulatory_Movement|Subject)*

[2] *Accuracy ~ Vowel * Articulatory_movement * + (1 + Vowel * Articulatory_Movement|Subject)*