PERCEIVING PROSODIC PROMINENCE VIA UNNATURAL VISUAL INFORMATION IN AVATAR COMMUNICATION

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

1 Introduction

Listeners integrate information from simulated faces in multimodal perception [1], but not always in the same way as real faces [2]. This is increasingly relevant with the dramatic increase in avatar communication in virtual spaces [3]. Prosody is especially relevant, because compared to segmental speech sounds, the visual factors indicating prosodic prominence (e.g. eyebrow raises and hand gestures) frequently bear no biomechanical relation to the production of acoustic features of prominence, but are nonetheless highly reliable [4], and avatar virtual communication systems may convey prosodic information through inappropriate means, e.g., by expressing amplitude via oral aperture (louder sound = larger opening).

Given that people are capable of picking up even small visual differences to aid in speech perception when interacting with other humans [5], oral aperture in an avatar might increase the perceived loudness of the stimuli. Furthermore, when the mouth is moving, differences between louder and quieter words may be attenuated by lack of difference in mouth aperture [1]. On the other hand, listener/viewers may disregard this inappropriate visual cue, leaving perceived loudness unaffected.

2 Methods

Ten native English speakers between the ages of 18 and 30 from the University of British Columbia participated in this study for course credit.

The stimuli consisted of videos of a single Facebook Spaces™ avatar (Figure 1) saying a sentence involving two characters and emphasizing one name (e.g. *Lee emailed BRIE.*). The stimuli followed a 2x2x2 Latin square design: Mouth movement (present or absent), emphasized word (first or second character's name), and which character name was rated for loudness (first or second mentioned). The stimuli were counterbalanced for the vowel quality of first character second character's names ([i:], [u:], or [ɑ:]).

To create naturalistic stimuli, the recording experimenter was asked questions that prompted the stimuli sentences with focus on one or the other of the character's names. (e.g. "Who did Lee email?" - "Lee emailed BRIE.",

In the mouth movement condition, lip shapes were simulated by the Facebook Spaces Beta [6] software by enabling the oculus headset microphone. According to a conference talk at F8 2017 [7] the avatar's lips go through visemes based on the acoustic information. Head and body movement were present in both conditions.

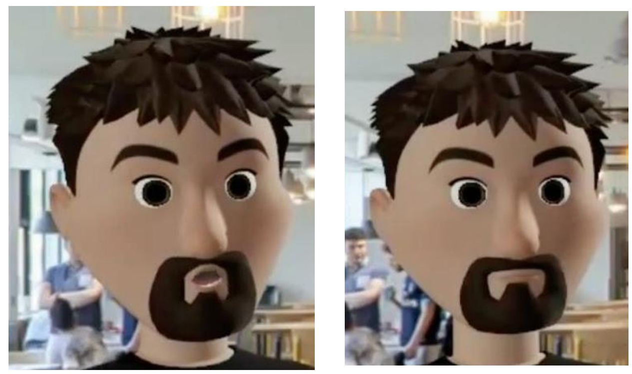


Figure 1: Close-up screen capture of Avatar with mouth opened and closed. Arms and torso were visible in the experiment.

Sound was recorded through a lapel microphone and the software Audacity. Auditory stimuli were normalized for amplitude with a script in Praat [8] and then fine-tuned by a research assistant. The externally recorded audio was synced to the in-app audio using Kdenlive [9] and Final Cut Pro X [10] at 30 fps.

Participants sat in a sound attenuated booth in the Interdisciplinary Speech Research Laboratory at the University of British Columbia. Stimuli were presented using OpenSesame 3.2.4 [11] on an iMac 2017, with AKG K240 headphones. The experiment was presented in two blocks. Both blocks consisted of all 36 tokens in pseudo-randomized order. In the first block participants rated the loudness of the first character's name, and in the second block they rated the loudness of the second character's name, each time using a 5-point Likert scale: *1 Not Loud* to *5 Very Loud.* Once a response was collected for one video, the next would be played automatically.

3 Results

The Likert scale data were aggregated [12], and then analyzed with linear mixed effects models in R using the lme4 [13] and lmerTest [14] packages. Crucially, one of the models included mouth-movement as a term (1), and the other did not (2).

(1) *Response ~ Accented * Character * Mouth_movement + (1|Subject)*
(2) *Response ~ Accented * Character + (1|Subject)*

* taylryan@gmail.com
† dprica@alumni.ubc.ca
‡ esther_wong_yt@hotmail.com
° keoughm@mail.ubc.ca
gick@mail.ubc.ca

Response referred to the Likert scored loudness, *Accented* to which character name was accented, *Character* to whether the loudness of pronunciation of the first-mentioned or second-mentioned character was rated, *Mouth_movement* referred to whether the avatar's mouth was moving, and *(1|Subject)* includes subject as random intercepts.

There was a two-way interaction of *Accented* and *Character* (t = 9.01, p < .0001). The terms *Character* (t = -12.78, p < .0001) and *Accented* (t = -6.71, p < .0001) were also significant. Accented words were perceived as louder than unaccented words, and character names at the beginning of the sentence were perceived as louder than character names at the end of the sentence (Figure 2).

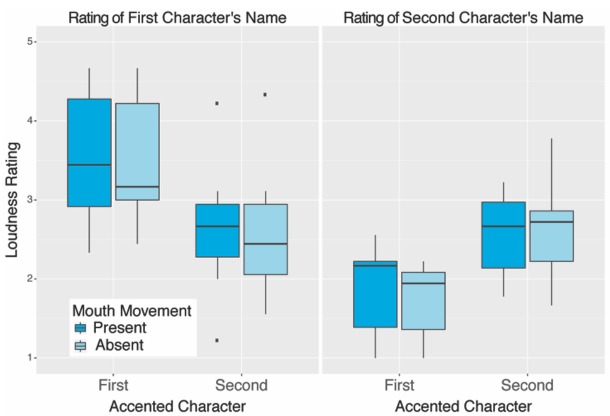


Figure 2: Boxplots demonstrating perceived loudness ratings of first and second characters

Neither *Mouth_movement*, nor any interactions with *Mouth_movement*, were significant (all ts < .53). Adding to the evidence, likelihood ratio test comparing (1) and (2) suggested that (2) is the better model, since there was no significant difference between (1) and (2), and (2) included fewer terms. Taken together, the evidence indicates that *Mouth_movement* played no detectable role in the perception of loudness. An analysis examining *Vowel* also failed to detect any contribution of that term to the model.

4 Discussion and conclusions

This study examined whether the aperture of an avatar's mouth was interpreted as an indicator of loudness. If mouth aperture were taken as an indicator of loudness, then the mouth movement condition would have been perceived as louder than the no mouth movement condition.

The results showed that subjects disregarded mouth aperture when judging the loudness of the words. Not only did they disregard that the mouth was opening more in response to signal amplitude, they even disregarded that for half of the stimuli the mouth didn't open at all.

These results support the finding that speech from avatars is not perceived the same as speech in face-to-face, in-person communication [1]. Mouth movements likely affect speech prosody in other ways than loudness perception, and future work should examine whether mouth aperture has any effect on the general perception of prominence *beyond* loudness, and whether that affects the message level of communication.

Acknowledgments

Research funded by NSERC. The authors wish to thank Gracellia Purnomo for her contributions to the project.

References

[1]~Cohen, M., and Massaro, D. Synthesis of Visible Speech. *Behavior Research Methods Instruments & Computers*, 22(2), 260–263, 1990.

[2]~Keough, M., Taylor, R.C., Derrick, D., Schellenberg, M, and Gick, B. Can. Acoust. 45(3):176-177, 2017

[3]~Bloomberg. Professional services. Computing's next Big Thing? Virtual world may be Reality by 2020. *Bloomberg Professional services.* https://www.bloomberg.com/professional/blog/computings-next-big-thing-virtual-world-may-reality-2020/, 2017. Accessed: May 29, 2018

[4]~Krahmer, E. and Swerts, M., *JML* 57 (3): 396–414, 2007.

[5]~Abel J, Barbosa AV, Black A, Mayer C, and Vatikiotis-Bateson E. The labial viseme reconsidered: Evidence from production and perception. *JASA*, 129(4):2456, 2011.

[6]~Facebook Spaces Beta version 67.0 [Computer Software], Facebook, 2017.

[7]~Facebook. Facebook for developers – F8 2017 Keynote. https://developers.facebook.com/videos/f8-2017/f8-2017-keynote/. Accessed: October 28, 2018.

[8]~Boersma, P. Praat, a system for doing phonetics by computer. *Glot international* 5, 2002.

[9]~Kdenlive. 18.04.1. [Computer software] J. Wood. USA, 2018.

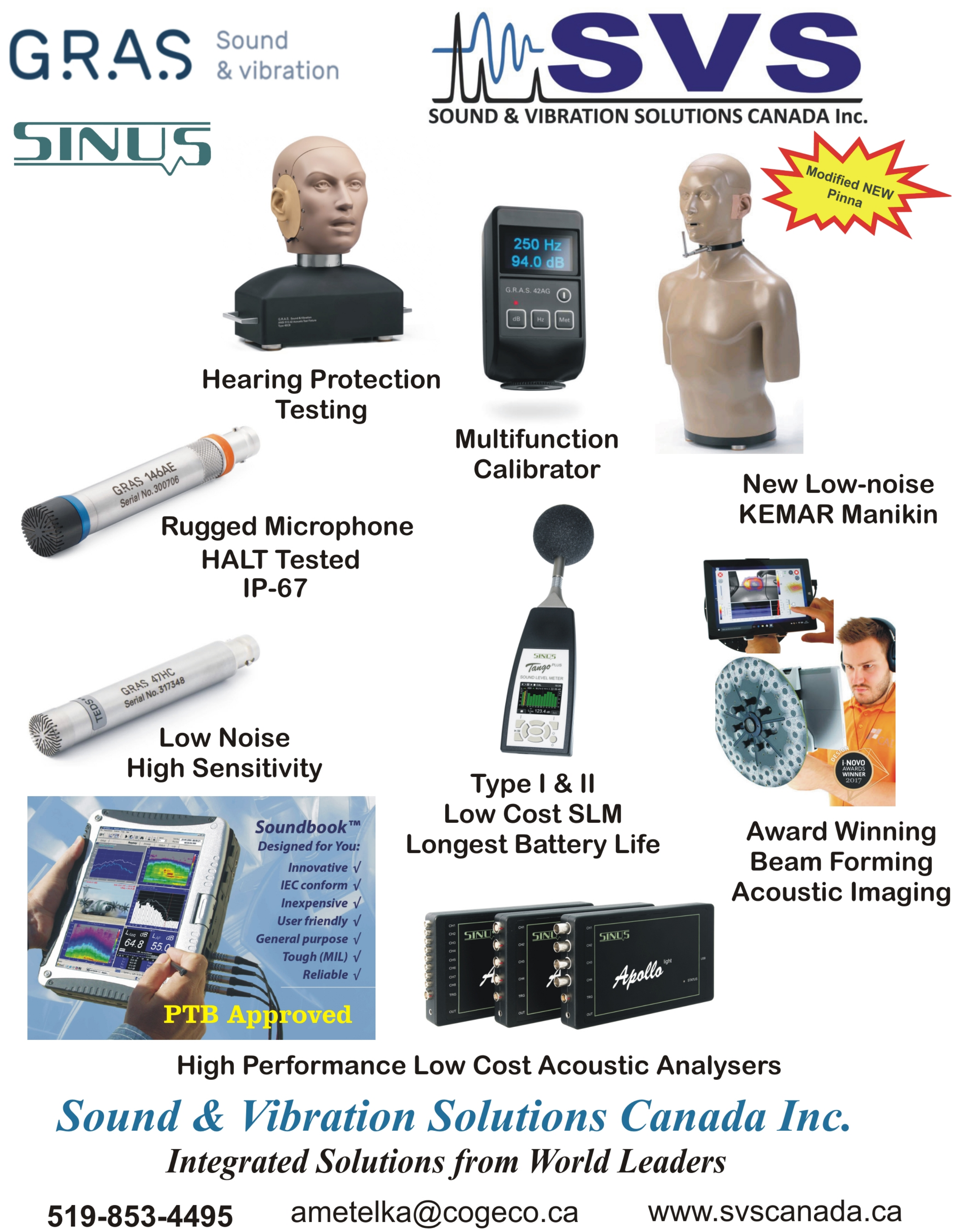
[10]~Final Cut Pro X. 10.1.2. [Computer software] Apple Inc.. USA, 2014.

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

[12]~Harpe. How to Analyze Likert and Other Rating Scale Data. *Curr. in Pharm. Teach. and Learn.* 7, 6:836–50, 2015.

[13]~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.

[14]~Kuznetsova, A., Brockhoff, P.B., Christensen, R.H. lmerTest package: tests in linear mixed effects models. *J. Stat. Software.* 82(13): 1–26, 2017.


GRAS Sound & vibration
SINUS
SVS
SOUND & VIBRATION SOLUTIONS CANADA Inc.
Modified NEW Pinna
Hearing Protection Testing
Multifunction Calibrator
New Low-noise KEMAR Manikin
Rugged Microphone
HALT Tested
IP-67
Low Noise
High Sensitivity
Type I & II
Low Cost SLM
Longest Battery Life
Award Winning
Beam Forming
Acoustic Imaging
Soundbook™
Designed for You:
Innovative √
IEC conform √
Inexpensive √
User friendly √
General purpose √
Tough (MIL) √
Reliable √
PTB Approved
High Performance Low Cost Acoustic Analysers
Sound & Vibration Solutions Canada Inc.
Integrated Solutions from World Leaders
519-853-4495
ametelka@cogeco.ca
www.svscanada.ca