PARTICIPANT ERROR IN ONLINE PRODUCTION DATA COLLECTION IN GORILLA

REFLECTION ABOUT ONLINE PRODUCTION DATA COLLECTION BASED ON A BAG-RAISING PRODUCTION STUDY

Lisa Sullivan, University of Toronto
Challenges for Change Workshop, Labphon 18, June 22, 2022

1. STUDY BACKGROUND/METHOD

This BAG-raising (/æ/ raises before /g/) production study used Gorilla (Anwyl-Irvine et al., 2020) to collect audio recordings of BAG-raising from 50 English speakers in Ontario and Colorado (plus 39 exclusions, 28 of which were for failing to meet inclusion criteria). Participants completed a word list reading task with 60 stimulus items including 22 target pre-velar words and 22 target vowel space words. On each trial, participants were presented with the stimulus item and asked to read it once. When they were done, they could click 'done' and then 'next' to proceed to the next trial. The word list was repeated 3 times for a total of 180 words per participant.

The task can be viewed here: https://app.gorilla.sc/openmaterials/417553

2. MOTIVATION FOR EXPERIMENT DESIGN

The trial design of this experiment as based on feedback received from the speaker who recorded stimuli for a previous perception experiment on Korean names (Kang & Sullivan, 2020). The stimuli for this experiment were collected using Gorilla to test the platform. The speaker completed two recording sessions, both of which produced audio data of sufficient quality for manipulation in Praat (Boersma & Weenink, 2021).

In the first session, the experiment was designed such that the speaker had 10 seconds to make their recording for each trial. Once these 10 seconds were up, the experiment proceeded automatically to the next trial. The speaker recorded that he felt this was a long time to wait to proceed to the next trial and that it would be beneficial to be able to stop the procedure.

Based on this feedback, the experiment was modified for the second session by allowing the participant to click 'done' when they were finished speaking to end the recording, and then 'next' to proceed to the next trial. This sped up the process and allowed the participant to take a break if necessary. The speaker preferred this version of the experiment, so this procedure was used for the BAG-raising study.

The two versions of the Korean name recording task can be viewed here:
https://app.gorilla.sc/openmaterials/74290

3. REFLECTION

Online production data collection was one of the major challenges of the BAG-raising study. Initial concerns about this related to the quality of the data collected given the lack of control over the types and quality of recording devices, as well as the possibility of background noise if participants were recording in noisy environments. One step taken to mitigate this concern was to limit participation in the experiment to those who were using personal computers, rather than cell phones or tablets, thereby reducing the variety of devices that could be used and number of possible recording locations. There is some evidence that computers may provide better recording quality than tablets or phones (Freeman and De Decker, 2021b, 2021a)

While it is not clear if these limitations had an effect, recording quality was not a major concern in this study, with only 2 participants being excluded due to technical issues being due to recording quality issues, as well as an additional 2 being excluded due to corrupt data files, which could be related to difficulty saving recordings back to the computer. Of the 61 participants who met the study inclusion criteria, only 4 (6.6%) were excluded for reasons related to the quality of recording devices or background noise.

What emerged as more concerning was the fact that many participants cut off recordings before they were complete, leading to many, potentially high quality, recordings being excluded. The online audio recording system in Gorilla required that each word be recorded and saved as a separate sound file. Figure 1 shows the waveforms of two recordings of *lag*, *bake* and *vague* spoken by the same speaker. The waveforms in the top row show tokens that had to be excluded due to being cut off. The waveforms in the bottom row were usable as the full vowel was recorded.

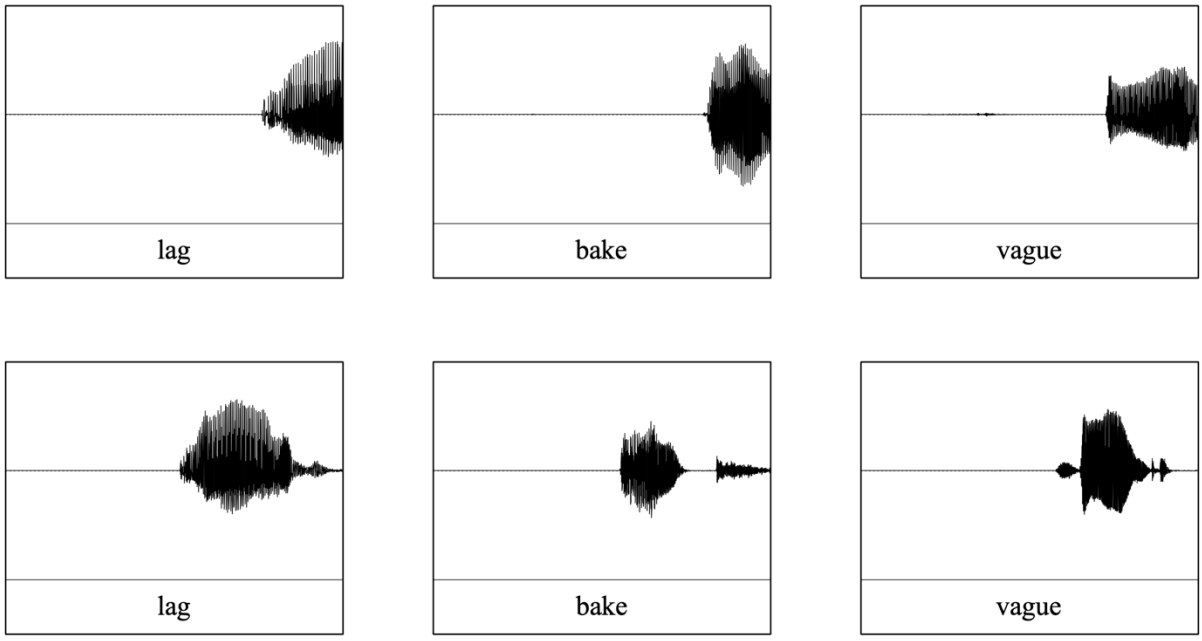

Figure 1 Waveforms of *lag*, *bake* and *vague* spoken by the same speaker. The waveforms in the top row represents recordings that had to be excluded from analysis.

In the experiment design, in order to reduce the length of time the experiment took, as well as to reduce participant boredom, participants were able to click 'next' to end the recording and proceed to the next word anytime from the onset of the recording. They were instructed to do so after they finished speaking; however, many participants clicked 'next' before they had finished speaking, which cut off the recording. This resulted in the inclusion criteria for participants being adjusted to include any participant who had at least one token of each vowel space word.

The vast majority the 620 tokens (3.4% of the total number of utterances by all included speakers) excluded for the participants included in the analysis were related to this problem. While this was 10 or less tokens (under 6%) for most participants (35/50, 70%), there were 10 participants (20%) who had 20 or more exclusions (11% to 47% of their utterances data). Additionally, the remaining 7 participants who were excluded from the analysis (7%) were excluded due to having too many cut off tokens. While some of these exclusions were due to issues not related to participants cutting of their recordings, the vast majority were, and these exclusions could have been avoided by preventing the participant from cutting off their recording before they finished speaking.

Moving forward based on this experience, I would like to continue collecting production data online. The vast majority of data appears to be of sufficiently high quality for acoustic analysis, and the vast majority of exclusions were due to participant/experimenter error rather than data quality. In order to address this concern, I would like to prevent participants from being able to click 'next' and end the recording for 1-2 seconds after each trial loads. This would, hopefully, reduce the number of exclusions due to cut off recordings without dramatically increasing the time to completion for participants.

REFERENCES

Anwyl-Irvine, A. L., J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods* 52(1), 388–407.

Boersma, P. and D. Weenink (2021). *Praat: doing phonetics by computer* [computer program]. version 6.1.38.

Freeman, V. and P. De Decker (2021a). Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices. *Language and Linguistics Compass* 15 (7), e12435.

Freeman, V. and P. De Decker (2021b). Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *The Journal of the Acoustical Society of America* 149(2), 1211–1223.

Kang, Y. & Sullivan, L. (2021.) Phonology of gender in Korean personal names. *Korean Linguistics in Crosslinguistic Context*. Cornell University, Online.