DATA FROM ONLINE PRODUCTION EXPERIMENTS AND CHALLENGES FOR COLLECTING GOOD-QUALITY RECORDINGS FOR PROSODIC ANALYSES

Dorotea Bevivino, Barbara Hemforth, Giuseppina Turco

Université Paris Cité, LLF, CNRS
dorotea.bevivino@u-paris.fr, barbara-edith.hemforth@cnrs.fr, giuseppina.turco@cnrs.fr

ABSTRACT

As an effect of pandemic-related restrictions, experimental research has been challenged to find alternative methods for data collection and analyses. When it comes to laboratory phonology and production studies, few experiments were run online so far, mainly measuring accuracy and latencies. The present study used speech data collected online via a crowd-sourcing platform to perform prosodic analyses on sentence productions. We first ran a pilot study with six participants to test audio quality and feasibility. After preliminary analyses, we conducted a large-scale experiment with remote speech production data collection. The results suggest that we can collect online good-quality recordings for prosodic analyses, particularly duration measurements. However, reduced audio quality and increased accuracy demand for large samples result in increased time and effort to collect and process the desired amount of usable data.

Keywords: online experiments, remote recordings, language production, prosodic analyses, Prolific

1. INTRODUCTION

In recent years, Covid-19 lockdowns and less restricted pandemic-related measures later on have significantly affected the way research is conducted. Experimenters have been challenged to find alternative methods for data collection and analyses not to stop research entirely. Despite recent concerns about low participant validity [1, 2], crowd-sourcing platforms have become increasingly common. Sophisticated experimental features have been implemented on web platforms, and multiple paradigms have been successfully adapted and validated for online data collection in various domains [3, 4, 5]. When it comes to phonetic sciences, remote speech data collection is not new. Some applications had already been developed to investigate production in less-documented languages [6], and remote recordings had been tested for clinical purposes [7]. More recently, new functionalities have been added to common experimental web platforms, and in the last year, some specialized tools have been developed for remote collection of production data [8, 9].

The feasibility of online platforms, web browsers, and everyday devices as alternative equipment for speech studies has been tested and validated in several meta-studies. The intelligibility of the audio stimuli has been confirmed in perception [10]. The time accuracy of the recording devices has been investigated in production [11, 12]. Moreover, several studies have simultaneously recorded and thoroughly compared the sound quality of remote alternatives and high-quality recorders, to test the reliability of acoustic parameters extractions. These extra-controlled multi-tests suggest a distortion effect on all measurements to some extent [13], with F0 being more robust than other formants [14, 15] or other acoustic parameters [16], but still giving pitch visualization problems if HNR levels are poor [17].

Despite this growing body of tools and methodological literature on remote speech data collection, to the best of our knowledge, only a handful of production experiments were indeed run online in the past years. Remarkably, these experiments mainly measured latencies and accuracy [18, 19, 20, 21] without performing more fine-grained acoustic measurements.

It is worth noting at this point that one online production experiment investigating acoustic and prosodic measurements [22] appeared while we were implementing our study. Although both [22] and the present study investigate prosodic measurements from speech data collected online, the two studies differ with respect to some key elements, and we think they independently contribute to the field. Specifically, in [22], participants were recruited from the lab pool (and not a crowd-sourcing platform); the experiment was implemented on a web-platform specifically designed for the lab (and not a widely accessible experimental platform); experimenters collected

multi-word naming productions (and not full sentence productions in a priming paradigm); duration measurements were taken on the entire multi-word utterance and in-between word pauses (and not in specific critical regions of the sentence in the continuous speech stream).

In this paper, we report our experience conducting a speech production experiment online to perform prosodic analyses. We first ran an online pilot study to test audio quality and feasibility. After preliminary analyses and necessary technical adjustments, we ran a large-scale study. Below, we present procedures and challenges we encountered while collecting and processing speech production data online, using a crowd-sourcing platform and web-based experimental software. The data were collected for a large-scale experiment aiming at performing prosodic analyses on specific regions of sentence productions in British English. We believe that both the successful parts of our procedures and the challenges we faced will be helpful for the community to run online production experiments more smoothly.

2. METHODS

Participants' eligibility criteria, experimental materials and design, data collection and data processing procedures were identical for the pilot and the full-scale experiment, unless otherwise specified. The study design and analysis plan were preregistered [23], and all procedures received ethical approval.

2.1. Participants

Young adult native English speakers from the same UK area were recruited online via Prolific, controlled for gender, age, and education level. Six participants took part in the pilot study; 60 in the full-scale experiment. All participants had normal or corrected-to-normal vision, no hearing impairments, and no known neurological, speech, or communication disorders at the time of testing. All participants provided informed consent.

2.2. Study design

Participants were tested on a prosodic priming task. The task consists in repeating a series of sentences auditorily (primes) or visually (targets) presented (see [24] for details on materials and design). The experimental session lasted around 30 minutes, including setup, questionnaires, and practice.

2.3. Technical setup

The task was coded using the Penn Controller for Ibex [25], taking advantage of its MediaRecorder element. The script was implemented and run on the Ibex farm server installed and maintained by CNRS technicians at Université Paris Cité. The server was properly set up to host the recordings. Participants were directly redirected to the experiment website from Prolific. All recordings were made from participants' personal computers, using either the built-in speakers and microphones or external ones.

Device restriction labels (computer, audio, microphone) were included in the study preview on Prolific, to inform participants beforehand and restrict participation. Also, all technical requirements were clearly stated in the study description on the website. Specifically, participants were explicitly advised not to use mobile phones or tablets to join the study, as submissions from these devices would be rejected. We restricted participation to computer-only users primarily because of the nature of the main priming experiment, which required stimuli to be displayed in full-screen size. At the same time, collecting recordings only from computers helped minimize noise due to different devices. To ensure participants only used computers, we checked for screen-width at the beginning of the script and denied access to the experiment if needed.

Similarly, participants were explicitly informed that the task was solely compatible with desktops and laptops using the Chrome browser. Participants were advised against using a different web browser, to avoid the experiment crashing and their submission being rejected. We made the experiment accessible only via Chrome to ensure the most common combination of operating system and web browser [12] and guarantee the stability of the testing platform. Once again, restricting participation to a single web-browser helped minimize potential extra noise in the data.

Along with device and web browser restrictions, participants were explicitly asked to wear a headset and be in a quiet space prior to accepting participation in the study. It was clearly stated that no data analyses would be possible on unclear audios, resulting in a rejected submission. Following [17, 26] suggestions, after being redirected to the experimental page, participants were provided with detailed guidelines to recreate a lab-like environment and ensure good-quality recordings (checking on the audio settings, cancelling excessive ambient noise, avoiding

electronic interference, etc.). No further checks were implemented to ensure that participants actually recorded alone in a silent room and used headsets, due to ethical and practical restrictions on video-recording the experimental sessions.

2.4. Recording procedure

Before starting the practice trials, participants were invited to perform an audio quality check. The recorder interface was displayed on the screen, and participants were asked to record and replay a short precise sentence, while paying attention to the sound quality. Suggestions for improving audio quality were also displayed again. Participants were invited to freely use the interface to take multiple tests until their audio was clear. To encourage engagement with the audio test, the button to proceed to the next trial was displayed only after a short delay.

In the regular experimental trials, the recording interface was no longer displayed on the screen, and participants were instructed to simply repeat sentences using the spacebar to advance. Participants were asked to repeat each sentence, even if not accurate, and to say *Pass* or similar if they could not recall anything. This would enable us to avoid empty recordings and immediately detect potential technical issues or bad-faith participants.

Each recorded sentence was immediately uploaded to the server in the background. At the end of all trials, completion screens remained on the screen until all recordings were uploaded to the server. In the full-scale experiment, participants were encouraged to download the recording zip file locally, to be able to send it to experimenters in case of connection failure or server problems before the upload was complete.

2.5. Data processing and analysis

Each recorded sentence was automatically saved as a separate WebM file and stored in zip files with unique names by PCIbex. WebM files were converted offline to .wav files for analysis. The audio-recorded data were transcribed automatically using Praat [27] scripts. Each file was then manually checked for sound problems, accuracy errors, and to correct any discrepancies between the automatically transcribed expected productions and the speakers' actual productions. The data were hence forced-aligned using the Montreal Forced Aligner [28].

For each repeated sentence, we measured the duration at two specific critical regions, to address the research question of our main experiment. In the full-scale experiment, we also extracted F0 values from a subset of data using ProsodyPro [29], to run exploratory analysis of pitch modulation.

3. PILOT RESULTS

Due to issues with the server and how the original Ibex farm hosts multimedia files, 4 participants experienced difficulties loading the audio stimuli. As a result, their experimental session was slowed down or interrupted. Hence, we listened to and overall evaluated all the collected recordings. More fine-grained analyses were performed only on a subset of data specifically selected to investigate the research question of the main experiment.

After the overall evaluation, none of the 6 participants was excluded for large number of blank recordings or general poor audio quality. In the subset selected for prosodic analyses (N = 158 recordings), no recordings were discarded due to background noise or poor sound quality, few recordings were discarded due to blank (N = 1) or *Pass* (N = 2) productions, several recordings were excluded because the sentence ROI was missing (N = 9) or presented disfluencies (N = 1) (see Fig. 1). No specific problems were encountered when conducting the measurements and running scripts at any stage of the data processing.

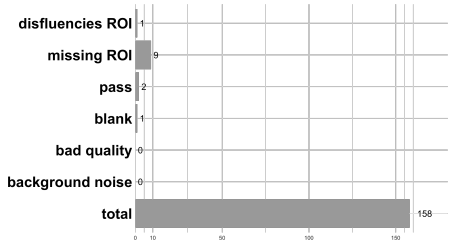

Figure 1: Raw counts of discarded recordings by error type in the subset of pilot data.

These preliminary results attested the general good quality of the recordings to perform prosodic analyses. Thus, online collection of speech production data seemed not only to be feasible and sufficiently reliable for accuracy and latencies but also for prosodic analyses. The results of the pilot study allowed us to proceed safely with the pre-registered experiment on a greater scale and including extended analyses.

4. FULL-SCALE EXPERIMENT RESULTS

Once the resource loading problem was fixed, no participants in the full-scale experiment reported any issues preventing them from completing the task.

Two participants reported being slowed down due to misreading the instructions (they did not press the spacebar and instead let the experiment dictate the pace). With an original target sample size of approximately 80 participants, we began analyzing the data after collecting recordings from the first 40.

After overall preliminary analyses to assess the audio quality of the recordings, we excluded 5 participants due to constant electronic interference and/or severe background noise (e.g., people around, street noise, dog barking, mouse clicking, etc.). During the transcription process, we further excluded 12 participants who produced unusable recordings for more than 20% of their productions. This resulted in a sample of 23 participants out of the 40 whose data were analyzed. Similarly, we did not include in the final analyses bad recordings (e.g., blank or *Pass* productions, noises, fragments, disfluencies at/near the critical regions) from good participants, resulting in a final sample of 1625 productions out of the total 1830 sentences produced by the 23 good participants (see Fig. 2). Due to the large number of excluded participants and recordings, we decided not to proceed further with data processing for the other 20 participants, but rather to replace excluded participants to ensure a balanced sample.

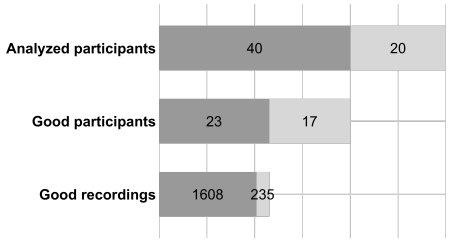

Figure 2: Raw counts of included vs. excluded participants and recordings in the full-scale experiment.

Alike the pilot data, no particular issues arose during file conversion, or measurements, or script execution for speech-to-text alignment or duration extraction. All extracted duration data were successfully and reliably modeled to address the research question of the main experiment [24].

Regarding the exploratory analysis of pitch modulation conducted on a subset of data, we again did not encounter any technical issues while running the scripts to extract normalized F0 curves across speakers. However, an initial examination of the plotted curves revealed a high number of sudden jumps and considerable variability in the curves, rendering pitch visualization unreliable.

5. GENERAL DISCUSSION

Our results suggest that we can conduct production experiments online, via crowd-sourcing platforms and web-based experimental softwares, and still collect good-quality recordings. These recordings not only seem to be reliable for accuracy and latencies, but also for prosodic analyses on sentence productions, particularly duration measurements.

The observations on F0 are in line with previous reports of pitch visualization problems [17]. Also, it is worth noting that for this exploratory analysis, we did not perform any landmarking procedure to normalize timing differences in the individual productions. The nature and structure of our stimuli did not allow for a straightforward isolation of the ROIs. The lack of landmarking may account for some variability and alignment issues in the data. However, it still remains unclear to what extent the variability resulting from different devices, HNR levels, recording distances, and noisy environments might still affect intonational measurements in online data collection.

Notwithstanding the reliability of data for duration measurements, in our full-scale study, we had to discard a large number of recordings (up to roughly half of the analyzed participants) due to technical and behavioral issues that could have been avoided in a lab setting. This is despite the emphasis on the critical importance of audio quality and all reminders provided to participants on the environment setup. While reduced audio quality and increased accuracy demand are physiological for large samples, our results seem rather to denote a general lack of precision and attention during online tasks, in line with previous reports on online participants, even when recruited from lab pools [19]. These findings also point out how our data suffered from the lack of monitoring of participants and their productions during the task, a standard practice in lab data collection. All of this results in increased time and effort required to collect and process the desired amount of usable data.

In conclusion, our study suggests that remote speech data collection is a feasible and reliable option not only for accuracy and latencies but also for prosodic analyses, particularly duration. This approach is a potentially valuable strategy for extending research on less-documented languages or for collecting naturalistic interactional data. However, it appears more suitable for small samples and simple tasks, where noise and attention-related issues can be minimized.

6. REFERENCES

[1] M. A. Webb and J. P. Tangney, "Too good to be true: Bots and bad data from Mechanical Turk," *Perspect. Psychol. Sci.*, 2022.

[2] A. J. Moss, C. Rosenzweig, S. N. Jaffe, R. Gautam, J. Robinson, and L. Litman, "Bots or inattentive humans? Identifying sources of low-quality data in online platforms," *PsyArXiv*, 2021.

[3] M. Vos, S. Minor, and G. C. Ramchand, "Comparing infrared and webcam eye tracking in the visual world paradigm," *Glossa Psycholinguistics*, vol. 1, 2022.

[4] M. S. Slim and R. J. Hartsuiker, "Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js," *Behav. Res. Meth.*, 2022.

[5] A. Chuey, M. Asaba, S. Bridgers, B. Carrillo, G. Dietz, T. Garcia, J. A. Leonard, S. Liu, M. Merrick, S. Radwan, J. Stegall, N. Velez, B. Woo, Y. Wu, X. J. Zhou, M. C. Frank, and H. Gweon, "Moderated online data-collection for developmental research: Methods and replications," *Front. Psychol.*, vol. 12, 2021.

[6] "Lessons learned after development and use of a data collection app for language documentation (lig-aikuma)," in *Proc. 19th ICPhS*, Melbourne, 2019.

[7] S. Jannetts, F. Schaeffler, J. Beck, and S. Cowen, "Assessing voice health using smartphones," *Int. J. Lang. Commun. Disord.*, vol. 54, no. 2, pp. 292–305, 2019.

[8] A. L. Thomas and P. F. Assmann, "SpeechCollectR: An R package for web-based speech data collection," *J. Acoust. Soc. Am.*, vol. 150, no. 4, pp. A356–A357, 2021.

[9] T. T. Schnoor and B. V. Tucker, "SpeakerPool: A remote speech data collection platform," *J. Acoust. Soc. Am.*, vol. 152, no. 4, pp. A60–A60, 2022.

[10] M. Cooke and M. L. Garcia Lecumberri, "How reliable are online speech intelligibility studies with known listener cohorts?" *J. Acoust. Soc. Am.*, vol. 150, no. 2, pp. 1390–1401, 2021.

[11] A. Anwyl-Irvine, E. S. Dalmaijer, N. Hodges, and J. K. Evershed, "Realistic precision and accuracy of online experiment platforms, web browsers, and devices," *Behav. Res. Meth.*, vol. 53, no. 4, pp. 1407–1425, 2021.

[12] D. Bridges, A. Pitiot, M. R. MacAskill, and J. W. Peirce, "The timing mega-study: Comparing a range of experiment generators, both lab-based and online," *PeerJ*, vol. 8, 2020.

[13] C. Sanker, S. Babinski, R. Burns, M. Evans, J. Johns, J. Kim, S. Smith, N. Weber, and C. Bowern, "(Don't) try this at home! The effects of recording devices and software on phonetic analysis," *Language*, vol. 97, no. 4, 2021.

[14] C. Zhang, K. Jepson, G. Lohfink, and A. Arvaniti, "Comparing acoustic analyses of speech data collected remotely," *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp. 3910–3916, 2021.

[15] C. Ge, Y. Xiong, and P. Mok, "How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements," in *Interspeech 2021*, Brno, 2021.

[16] J. Penney, A. Gibson, F. Cox, M. Proctor, and A. Szakay, "A comparison of acoustic correlates of voice quality across different recording devices: A cautionary tale," in *Interspeech 2021*, Brno, 2021.

[17] G. Magistro, "Speech prosody and remote experiments: Atechnical report," *PsyArXiv*, 2021.

[18] M. Kandel, C. R. Wyatt, and C. Phillips, "Agreement attraction error and timing profiles in continuous speech," *Glossa Psycholinguistics*, vol. 1, no. 1, 2022.

[19] A. Fairs and K. Strijkers, "Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors," *PLOS ONE*, vol. 16, no. 10, p. e0258908, 2021.

[20] A. Vogt, R. Hauber, A. K. Kuhlen, and R. A. Rahman, "Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice," *Behav. Res. Meth.*, vol. 54, no. 4, pp. 1954–1975, 2022.

[21] K. Stark, C. van Scherpenberg, H. Obrig, and R. Abdel Rahman, "Web-based language production experiments: Semantic interference assessment is robust for spoken and typed response modalities," *Behav. Res. Meth.*, 2022.

[22] J. He, A. S. Meyer, A. Creemers, and L. Brehm, "Conducting language production research online: A web-based study of semantic context and name agreement effects in multi-word production," *Collabra: Psychology*, vol. 7, p. 29935, 2021.

[23] D. Bevivino, B. Hemforth, and G. Turco, "Prosodic Priming (Replication of Tooley et al. 2018)," 2022. [Online]. Available: osf.io/z24ke

[24] D. Bevivino, G. Turco, and B. Hemforth, "Priming prosodic boundaries across constructions," in *Linguistic Evidence 2022*, Paris, 2022.

[25] J. Zehr and F. Schwarz, "PennController for Internet Based Experiments (IBEX)," 2018.

[26] E. James, "Tools and tips for collecting spoken responses online," 2020.

[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," [Computer program], 2021. Version 6.1.56.

[28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech 2017*, Stockholm, 2017.

[29] L. Xu, "ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis," in *TRASP 2013*, 2013, pp. 7–10.

7. ACKNOWLEDGEMENTS

This work was financed by Labex EFL (ANR-10-LABX-0083). It contributes to the IDEX Université Paris Cité (18-IDEX-0001).