IMPACTS OF REMOTE AND TRADITIONAL RECORDING METHODS ON QUALITY OF FORMANT ANALYSIS FOR VOWEL REDUCTION

Jenna T. Conklin

Carleton College
jconklin@carleton.edu

ABSTRACT

Past work on recording quality of remotely-collected speech data has demonstrated that some remote methods are adequate for broad-strokes formant analysis (such as comparing locations of phonemes in the vowel space), but none are suitable for fine-grained formant analysis, such as sociolinguistic work. This study evaluates the adequacy of two remote recording techniques for formant analysis of vowel reduction, which requires an intermediate degree of specificity between these two extremes. It compares lossy recordings captured remotely via Gorilla, lossless smartphone recordings self-administered by remote participants, and traditional recordings created in-person in a sound booth. Results show only minimal variation across recording methods, indicating that remote recording via Gorilla or self-administered smartphone recordings may prove viable options for remote data collection for similar studies, although the convenience of remote recruitment and recording are offset by higher rates of data loss and difficulties in annotation and formant extraction compared to traditional methods.

Keywords: vowel reduction; remote recording; formant analysis

1. INTRODUCTION

Researchers who wish to collect speech data remotely face an array of challenges; variation in compressive algorithms, hardware and environment, and sampling rates conspire to create an unpredictable landscape capable of exerting unequal impacts on the resultant recordings [1]. While it has long been known that variation in recording device and environment can lead to inconsistent acoustic measurements of speech [2], the body of research on the quality of remote recording of speech data has greatly expanded in recent years in response to the COVID-19 pandemic. With regard to the impact of remote data collection on analysis of vowel formants, studies have examined the effect of recording via videoconferencing apps [3]–[5] and varied hardware arrangements [1], [3], [6], [7], largely focusing on the degree of acoustic divergence from traditional recording setups. Many of these studies focused either on large-scale differences, such as the relative arrangement of vowels, or extremely nuanced differences, building a general consensus that analysis of large or relative formant differences can be carried out remotely under the right conditions, while work examining minor shifts will benefit from a traditional recording setting [5], [6]. The current study considers the effect of remote recording on an analysis of vowel reduction, a phenomenon that demands a degree of precision between these two extremes, by comparing data taken in a laboratory setting to that from two remote setups.

1.1. Sources of Variation in Remote Recording

Obstacles to obtaining high-quality recordings remotely are numerous, and only some are within researchers' control. Loss of data due to background noise or microphone placement, for instance, can be offset by providing clear instructions, but even with participants' best efforts, some disruption is likely. One source that can be controlled is the recording format: researchers must choose between lossy audio formats, which reduce file size by selectively deleting data, and lossless formats, which accept larger file sizes in exchange for more faithful recordings. Remote recordings taken online in real time, such as those obtained through Zoom, Skype, Teams, and Gorilla, are lossy, and some distortions to the vowel space relative to traditional recording are to be expected [3]–[5], although [4] was able to correct for these changes through Lobanov normalization. At least one study of remotely-collected speech data quality utilizing lossless recording, wherein participants recorded themselves on a device offline and then uploaded the file, found little difference in formant values from traditional recordings [3], suggesting that offline smartphone recordings might be the most viable alternative to in-person data collection for work considering suitably large effects.

The hardware used for a remote study will also impact the formant values of resultant recordings. [6] examined data recorded using tablets, smartphones, laptops, and a more traditional microphone and found that laptops offered fewer discrepancies than tablets and smartphones, male vowel spaces underwent less distortion than female ones, and the 750-1500 Hz range was especially likely to be distorted, often

affecting the low back vowels. Similarly, [7] compared formant values of Chinese vowels collected simultaneously across seven devices in two different settings (lab and conference room) and found that F1 and F2 values exhibited some systematic differences across devices, with greater variation in F2 compared to F1. These results suggest that remote subjects' personal hardware should be suitable for formant analysis, provided the study is focused on large distinctions or relative arrangements of vowels, but perhaps not for fine-grained sociolinguistic analysis. Researchers should expect some variation to be introduced by the model of the recording device, and should be aware that direct comparison of formant values across recording devices is likely to lead to misleading conclusions.

2. METHODS

2.1. Recording Methods & Subjects

An identical task was administered to two groups of participants, one remote (6M, 4F; M=35.6 yo, SD=10.65), one in-person (2M, 8F; M=20.9 yo, SD=1.96). All subjects were monolingual speakers of Midwestern American English. The in-person group completed the task in a sound-insulated booth wearing Sennheiser HD 380 Pro headphones, the stimuli were presented using PsychoPy [8], and their speech was recorded with a Shure KSM32 cardioid condenser microphone attached to a TubeMP preamp and digitized at 44.1 kHz. Data from the in-person group also served as a control for a separate study on bilingual vowel reduction (currently in preparation).

The remote group was recruited via Prolific [9] and completed the task over Gorilla [10]. One set of recordings was taken using Gorilla's Audio Recording Zone to generate .weba files using whatever microphone participants had for their computer. (.weba is a lossy format created with the OGG Vorbis compression codec [11].) A second, simultaneous recording was created by remote participants on their smartphone; this recording was a lossless .wav file recorded via a free app (Hokusai Audio Editor for Apple users and ASR Voice Recorder for Android owners). (One participant encountered a microphone error of some kind on their computer, and thus contributed only a smartphone recording and no Gorilla recording, leaving nine subjects in the Gorilla group.) Participants adjusted the app settings to at least a 128 kbps bitrate (Apple users selected an even higher-quality 16-bit setting) and a 44.1 kHz sampling rate. They were instructed to place both the smartphone and computer microphone in a stable position 6 – 10 inches from their mouth, or, in the case of a laptop-internal microphone, to sit at a comfortable distance from the screen. When the task was complete, subjects uploaded the smartphone recording to the researcher.

2.2. Task & Procedure

Participants completed a shadowing task in which they heard and repeated a sentence containing a target word with one of five vowels /ɑ, æ, ɛ, ɪ, ʌ/ in stressed or unstressed position. Stimuli were read by a male native speaker of Midwestern American English, recorded in the same manner as the in-person participants described in §2.1. All stressed vowels appeared in monosyllables, and unstressed vowels appeared in disyllables. Each target word had a counterpart in the study with identical structure for the target syllable, but different stress (i.e., bit ~ rabbit, text ~ context). Sixty target words (six for each stressed vowel and six for each unstressed vowel) were included; each subject repeated each target word twice. Target words were embedded in a unique carrier sentence of the form "the word X means…" where the phrase following "means" was relatively consistent in length, rhythm, and complexity.

2.3. Analysis

The beginning and end of the vowel were annotated by hand in Praat [12] and formant values were extracted at vowel midpoint using an LPC-based Praat script that allowed for manual review [13]. If the researcher observed that the automatic formant reading did not match the visual formant, a manual reading was substituted, taken by placing the cursor in the approximate center of the visible formant. Extracted formants were normalized using log-additive regression normalization [14]. The Euclidean distance for each pair of stressed & unstressed target utterances (i.e., each speaker's first repetition of bit ~ rabbit, each speaker's second repetition of bit ~ rabbit , etc.) was calculated to provide a measure of the degree of centralization for unstressed vowels, calculated as shown in (1).

$$(1)\quad EuD = \sqrt{(F1_{V1} - F1_{V2})^2 + (F2_{V1} - F2_{V2})^2}$$

3. RESULTS

3.1. Euclidean Distance

Euclidean distance provides a measure of the distance between two points: in this case, the degree of distance in F1xF2 space between a stressed vowel and its unstressed counterpart (see (1)). A linear mixed-effects model was fit in R [15] to evaluate the degree to which recording method impacted Euclidean

distance using *lme4* [16] and *lmerTest* [17]. The best-fitting model contained a response variable of Euclidean distance, fixed effects for Vowel (containing five levels /ɑ, æ, ɛ, ɪ, ʌ/) and Recording Method (with three levels, Gorilla, Smartphone, and In-Person), as well as an interaction term for Vowel by Recording Method and random intercepts for Subject and Item. Model results showed that the recordings taken via Gorilla differed significantly from those taken in-person (β = -.088, SE = .031, t = -2.865, p < 0.01), while the smartphone recordings did not differ significantly from the in-person data (β = -.047, SE = .030, t = -1.549, p = .131).

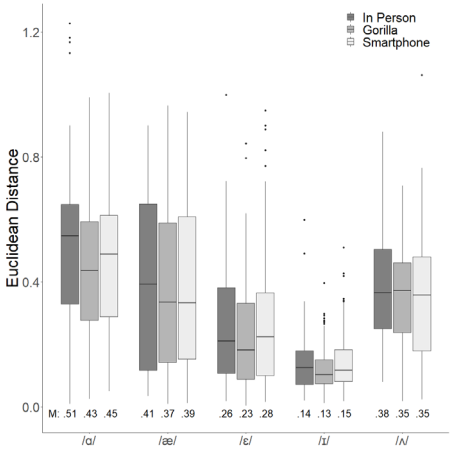

Figure 1: Euclidean distance by Vowel and Recording Method

Figure 1 displays the Euclidean distance of each vowel and recording method. As can be seen, the distribution of Euclidean distance is similar across recording methods for each vowel, but the Gorilla recordings diverge somewhat from the in-person data for vowels /ɑ/ and /ɛ/, while the smartphone recordings track more closely with the in-person data.

3.2. F1

To gain a more complete understanding of the variation in formant values across conditions, normalized formants were also examined. A linear mixed-effects model was fit with normalized F1 as a response variable; the best-fitting model included fixed effects for Vowel, Stress, and Recording Method, interaction terms for Vowel by Recording Method, Stress by Recording Method, and Vowel by Stress, and random intercepts for Subject and Item. Here, the results taken via Gorilla did not differ significantly from the in-person reference data (β = -.019, SE = .019, t = -.978, p = .334), although the uploaded data did (β = -.044, SE = .019, t = -2.306, p < .05). Figure 2 demonstrates that some of the divergence is shared by both the Gorilla and smartphone data (e.g., stressed /æ/) and thus was likely due to between-group differences, while other differences are limited only to the smartphone data and thus must be attributed to distortion introduced by the hardware or user behavior (i.e., microphone placement), as seen in several of the unstressed vowels. The smartphone recordings often exhibit a lower F1 than the Gorilla or in-person data, although this is not consistent across all vowels.

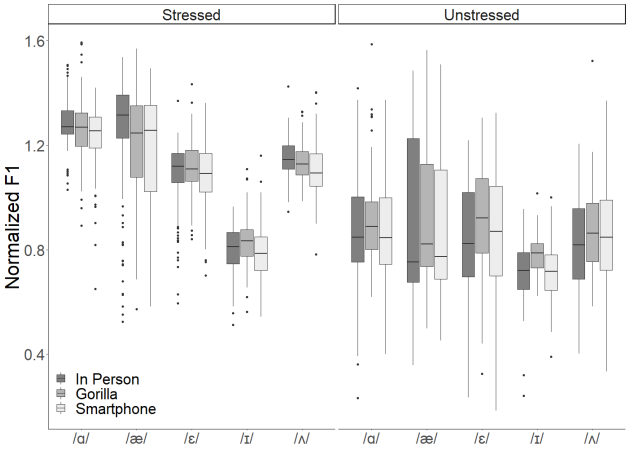

Figure 2: Normalized F1 by Vowel, Stress, and Recording Method

3.3. F2

A third linear mixed-effects model took normalized F2 as a response variable; the best-fitting model also included fixed effects for Vowel, Stress, and Recording Method, all possible interactions among these, and random intercepts for Subject and Item. Both the Gorilla and smartphone recordings differed from the in-person data in F2 (Gorilla: β = .056, SE = .019, t = 3.103, p < .01; smartphone: β = .052, SE = .019, t = 2.873, p < .01). As shown in Figure 3, stressed /ɑ/ and /æ/ had a higher F2 in the remote recordings than the in-person data, while other vowels showed minor shifts in either direction across methods, though F2 was higher in remote recordings for most vowels.

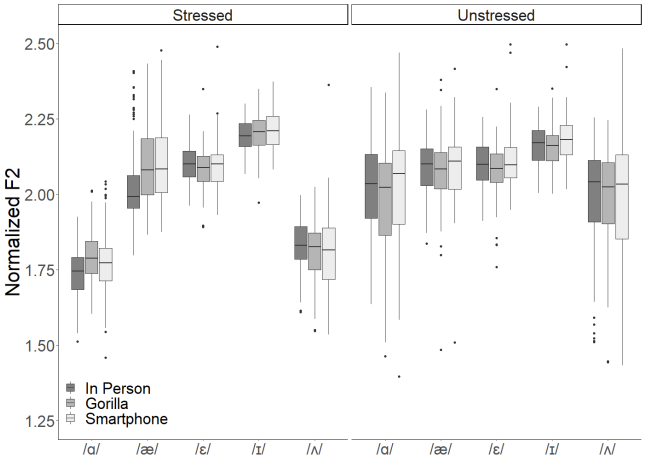

Figure 3: Normalized F2 by Vowel, Stress, and Recording Method

3.4. Pillai scores

Neither Euclidean distance nor direct comparison of formant values provides a measure of the degree of overlap between categories; to understand how distinct a stressed vowel and its unstressed analog are, the Pillai score is informative [18], [19]. Pillai scores closer to 0 indicate a great deal of overlap in the two categories, while scores approaching 1 mark that the two categories have little overlap. Pillai scores comparing each stressed and unstressed vowel for each speaker were calculated and are displayed in Figure 4; additionally, a linear mixed-effects model with a response variable of Pillai score, a random effect of Subject, and fixed effects of Vowel and Recording Method was fit. There was no statistically significant difference between the Pillai scores from the in-person recordings and either the Gorilla (β = -.01, SE = .048, t = -.216, p = .831) or smartphone (β = -.07, SE = .048, t = -1.495, p = .149) recordings. Thus, the degree of overlap between reduced and unreduced instances of the five vowels examined was not meaningfully impacted by switching from a traditional in-person recording setup to either of the two remote recording designs tested.

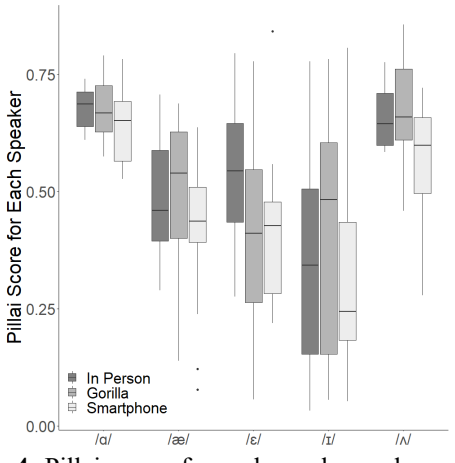

Figure 4: Pillai scores for each speaker and vowel pair

4. DISCUSSION AND CONCLUSIONS

The present study analyzed the reduction of English vowels using three recording methods: one group completed the task in a traditional laboratory setting, and a second group contributed two simultaneous remote recordings, one in a lossy format recorded on a laptop or desktop computer via Gorilla and the second in a lossless format recorded via smartphone. This design intentionally conflated between-speaker differences and variation due to remote participants' individual devices, background noise, and microphone placement to determine whether the results of the study would materially differ based on the researcher's choice to collect data remotely or in person. Analysis of normalized formant values across recording conditions reflected the findings of earlier studies: some variation was detected between the in-person and remote recordings, with the most notable deviance found in low or back vowels. Measures of vowel reduction showed minimal variation across recording methods: no statistically detectable difference emerged when reduction was quantified through Pillai scores, and when Euclidean distance was used, the Gorilla recordings but not the smartphone recordings differed significantly from the traditionally-recorded data. Given that the Gorilla recordings were encoded in lossy .weba files, it is not surprising that they would show greater deviance from the in-person data. The most obvious changes were in low or back vowels, which accords with the findings of [6].

The relative success of the smartphone recording group at replicating the results of an in-person study may be attributed to a handful of causes. First, vowel reduction is both a relative phenomenon, in that any analysis of it must quantify the degree of change of a vowel according to its prosodic prominence, rather than its absolute position in the vowel space. Relative measures are naturally more forgiving than absolute ones when searching for differences across platforms, since direct comparison across recording conditions is not required. Secondly, the scale of difference needed to effect change across groups is relatively large for vowel reduction in comparison to sociolinguistic studies, where the aim of the analysis is to capture the most minute of differences. Finally, it is possible that the normalization procedure served to mask some differences that otherwise would have been salient. (Formants were normalized using a single category for each vowel across recording conditions to ensure the resultant values were on a single scale.) In closing, the present results indicate that remotely-collected lossless smartphone recordings may be a viable alternative for work relying on relative formant values or expecting relatively large shifts in formants across conditions. Recordings taken via Gorilla exhibited greater distortions and should not be used for formant analysis. As others have noted, researchers planning remote data collection should prepare for higher rates of data loss and background noise, and should adjust their recruitment targets accordingly.

5. ACKNOWLEDGEMENTS

Thanks are due to Sophia Chuen for their assistance in collecting and analyzing data and to Josh Weirick and Diego Luna Hernandez for contributing their time and voices to the recording of the audio stimuli. This study was supported by the Headley Fund at Carleton College.

6. REFERENCES

[1] C. Sanker *et al.*, "(Don't) try this at home! The effects of recording devices and software on phonetic analysis," *Language*, vol. 97, no. 4, pp. e360–e382, 2021, doi: 10.1353/lan.2021.0075.

[2] P. De Decker and J. Nycz, "For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?," *Univ. Pa. Work. Pap. Linguist.*, vol. 17, no. 2, Jan. 2011, [Online]. Available: https://repository.upenn.edu/pwpl/vol17/iss2/7

[3] 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, Jun. 2021, doi: 10.1121/10.0005132.

[4] J. Calder *et al.*, "Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for vocalic analysis," *Linguist. Vanguard*, Feb. 2022, doi: 10.1515/lingvan-2020-0148.

[5] V. Freeman and P. De Decker, "Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps," *J. Acoust. Soc. Am.*, vol. 149, no. 2, pp. 1211–1223, Feb. 2021, doi: 10.1121/10.0003529.

[6] V. Freeman and P. De Decker, "Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices," *Lang. Linguist. Compass*, vol. 15, Jul. 2021, doi: 10.1111/lnc3.12435.

[7] 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*, Aug. 2021, pp. 3984–3988. doi: 10.21437/Interspeech.2021-1122.

[8] J. Peirce *et al.*, "PsychoPy2: Experiments in behavior made easy," *Behav. Res. Methods*, vol. 51, no. 1, pp. 195–203, Feb. 2019, doi: 10.3758/s13428-018-01193-y.

[9] "Prolific." London, UK, 2022. [Online]. Available: https://www.prolific.co

[10] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020, doi: 10.3758/s13428-019-01237-x.

[11] "What is a WEBA file?" Accessed: Jan. 06, 2023. [Online]. Available: https://docs.fileformat.com/audio/weba/

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." Nov. 03, 2020. [Online]. Available: http://www.praat.org/

[13] R. Scarborough, "Supervised Formant Reading Script." Jul. 2005.

[14] S. Barreda and T. M. Nearey, "A regression approach to vowel normalization for missing and unbalanced data," *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. 500–520, 2018.

[15] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[17] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest: Tests in linear mixed effects models." 2015. [Online]. Available: https://cran.r-project.org/web/packages/lmerTest/index.html

[18] J. Hay, P. Warren, and K. Drager, "Factors influencing speech perception in the context of a merger-in-progress," *J. Phon.*, vol. 34, no. 4, pp. 458–484, Oct. 2006, doi: 10.1016/j.wocn.2005.10.001.

[19] J. Nycz and L. Hall-Lew, "Best practices in measuring vowel merger," presented at the 167th Meeting of the Acoustical Society of America, Providence, Rhode Island, 2014, p. 060008. doi: 10.1121/1.4894063.