Laboratory Phonology
Journal of the Association for Laboratory Phonology

Conklin, J. T. (2023). Examining Recording Quality from Two Methods of Remote Data Collection in a Study of Vowel Reduction. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 14(1), pp. 1–28. DOI: https://doi.org/10.16995/labphon.10544

Open Library of Humanities

Examining Recording Quality from Two Methods of Remote Data Collection in a Study of Vowel Reduction

Jenna T. Conklin, Linguistics Department, Carleton College, Northfield, MN, USA, jconklin@carleton.edu

Remote recording quality of speech data varies significantly by recording devices, formats, and platforms, and past work has asserted that fine-grained sociolinguistic work should not be conducted remotely, while broad questions, such as analyses of the relative position of phonemes in the vowel space, may be amendable to remote data collection. In this study, lossless offline remote recordings taken via smartphone and lossy web-based recording performed over Gorilla are compared to traditional laboratory recordings in order to determine how accurately the remote options replicate a study of English vowel reduction. Four measures of reduction are examined: Relative duration, Euclidean distance, Pillai scores, and normalized formant values of stressed and unstressed vowels. Temporal analyses and Pillai scores were unaffected by recording method, while Euclidean distance and formant values exhibited some statistically significant changes but remained largely in line with laboratory data. These findings indicate that remote offline recording via smartphone or Gorilla may hold promise for studying vowel reduction and other phenomena requiring a similar degree of precision in formant analysis, but researchers should be aware of the specific distortions likely to be incurred with each method, with smartphone recordings having a stronger impact than Gorilla on low and back vowels.

1. Introduction

Remote data collection in speech production research poses multiple challenges to researchers in pursuit of high-quality data comparable to that obtained through in-person, laboratory-based data collection, but the benefits of faster recruitment, access to less commonly studied populations, and fewer infrastructural requirements (such as access to a sound-insulated booth) make remote recording a desirable objective for many researchers. Before attempting remote data collection, researchers should have a good working comprehension of how variation in compressive algorithms, hardware and environment, and sampling rates are likely to impact their data. The degree of disruption these factors introduce to the study is moderated by the particular goals of the study: Sociolinguistic work concerned primarily with small shifts in formant values is more likely to be strongly and adversely impacted by transitioning to remote data collection, while studies focusing on relative positioning of phonemes in the vowel space may be more suitable for remote collection (Freeman & de Decker, 2021b, 2021a). The present study examines the impact of remote data collection on an analysis of vowel reduction, an application of formant analysis requiring an intermediate degree of precision to the two points of focus of past research. By comparing remotely collected data to recordings made in a traditional laboratory setting, it attempts to quantify the impact of remote data collection on acoustic analyses of vowel reduction under two conditions (online lossy and offline lossless recordings) and with two approaches to the formant extraction process (automated formant extraction versus manually supervised formant extraction).

Variation in remote recording can be traced in large part to four sources: Background noise, variation in hardware, variation in software (i.e., the compression codec), and variation in microphone placement. Working together, these factors can exert unequal impacts on the resultant dataset (Sanker, Babinski, Burns, Evans, Johns, Kim, Smith, Weber, & Bowern, 2021). An emerging body of work has evaluated the impact of these factors as they relate to remote data collection, with the majority of its focus placed on variation stemming from hardware and software.

1.1. Variation due to device and algorithm in remote recording

In recent years, multiple studies have compared acoustic measurements of simultaneous recordings taken on an array of devices. Zhang, Jepson, Lohfink, and Arvaniti (2021) compared acoustic measurements of pitch and vowel formants across four recording modalities in a quiet, non-laboratory setting: A portable Zoom H6 Handy Recorder, the Zoom web meeting application using a built-in laptop microphone, and via smartphone using the built-in microphone on each of two different recording apps, Awesome Voice Recorder and Recorder. In a similar study, Sanker et al. (2021) compared several acoustic measures taken simultaneously via five devices and a sixth high-quality, solid-state device, and Ge, Xiong, and Mok (2021) evaluated simultaneous recordings taken across seven devices.

In Zhang et al. (2021), little variation across device was found for f0, but F1 and F2 were significantly affected by device type: The Zoom web application yielded lower F1 and F2 than the other recording methods, with the F2 of front vowels most strongly affected. Notably, no significant discrepancies in F1 or F2 emerged between the laboratory-style recording equipment (the H6 portable recorder) and the two recordings taken via smartphone, indicating that lossless smartphone recording in a quiet space may prove a viable approach to remote data collection. Similarly, Ge et al. (2021) concluded that f0 was more resilient to cross-device differences, and F1 fared better than higher formants. Zoom web recording was again found to introduce various distortions. Along similar lines, Sanker et al. (2021) found no difference from control recordings in F1 for two of the devices, and one device also exhibited no significant difference in F2. However, visible differences were still apparent in plots of vowel by speaker and device – crucially, the lack of a statistically significant effect does not equal a lack of difference. Several common conclusions emerged. Recordings taken via the Zoom web application introduced significant distortions, while some lossless local recording approaches did not differ significantly from control recordings taken on a traditional solid-state device. Pitch was more resilient than formant frequency to distortion by device or compression algorithm, and F1 was more resilient than F2. Because different vowels can be affected unequally and in opposite directions, a statistical effect can be absent while the impact on the analysis is severe and chaotic, a conclusion emphasized by Sanker et al. (2021). Reducing the impact of recording setup to a single dimension applied uniformly across the vowel space is thus ill-advised, and researchers partaking in remote data collection must be informed about the recording setup of their participants and its potential impacts on the resulting files.

While understanding the impact of device type on recordings is important, conflicting results have emerged regarding what type of home recording setup is likely to lead to the least distortion. As already discussed, Zhang et al. (2021) suggested that lossless recordings taken on a smartphone app introduced the least distortion, while Sanker et al. (2021) obtained the best results from a laptop equipped with an external headset microphone, followed closely by an Android phone. (The Apple phone performed notably less well with regard to vowel formants.) Freeman and De Decker (2021a) concluded that laptops (even utilizing the built-in microphone) offered the highest home recording quality. Given this disparity in results, researchers may be better advised to focus on the recording method rather than the device itself, as well as to thoroughly document the type of device used by each participant.

1.2. Impacts of compression on remote data collection

In addition to the device used to capture a recording, the file format used to record can also introduce acoustic distortions. Broadly speaking, audio can be captured in one of two formats. Lossy compression codecs limit file size by strategically deleting information; while these permanent changes typically do not compromise the comprehensibility of recorded speech, there can be staggering implications for acoustic analysis. Some lossy files undergo changes not only to spectral information, but also to temporal information: Sanker et al. (2021) found that the alignment of consonant segments was shifted in compressed files recorded via Zoom compared to simultaneously-captured uncompressed recordings. By contrast, lossless file formats capture the recorded audio in faithful detail, sacrificing file size for full accuracy.

Numerous studies have documented the distortions to the vowel space associated with lossy audio recording. De Decker and Nycz (2011) noted a raising of F1 and expansion of F2 associated with the conversion from a lossless file format to a lossy .mp3 format. Similarly, Freeman and de Decker (2021a, 2021b) found that the 750 – 1500 Hz formant range was particularly susceptible to distortion, most strongly affecting the low back vowels, and Calder, Wheeler, Adams, Amarelo, Arnold-Murray, Bai, Church, Daniels, Gomez, Henry, Jia, Johnson-Morris, Lee, Miller, Powell, Ramsey-Smith, Rayl, Rosenau, and Salvador (2022) also documented a lowering of F1 and raising of F2 in a compressed audio format, although they concluded that Lobanov normalization was able to correct for these distortions. Gender differences may also correspond to the amount of distortion that is introduced, with some studies reporting that the distortion is greater for female speakers (see, e.g., Freeman & de Decker, 2021a) and others reporting disparate correlations between gender and recording method varying by vowel, recording device, and normalization algorithm (Calder et al., 2022).

1.3. Hand measurement

One of the challenges of remote data collection is the potential for increased background noise, microphone noise, and other recording-related artifacts that arise more frequently when participants record themselves outside the lab environment than when the researcher is able to directly control and correct the recording milieu. Thus, remote recording often yields sound files with a higher signal-to-noise ratio than typically found in speech production studies. This added noise can interfere with accurate detection of formant frequencies and obscure results, particularly when formants are measured in an automated fashion (de Decker, 2016). However, it may be possible to correct for this issue by implementing a manual formant-extraction protocol or, as is done in the present study, by using a human-supervised script to measure formant frequencies. One of the objectives of the current study is to determine whether and in what fashion the use of an automated or supervised formant extraction script has a detectable impact on the conclusions of the analysis.

1.4. Research Question

The goal of the current study is to determine whether either of two remote recording methods (lossy recordings taken via Gorilla and lossless recordings taken via smartphone) generate sufficient distortion to formant frequencies to shift the conclusions of an analysis of vowel reduction. To achieve this goal, a replication of an earlier study of vowel reduction completed in person in a laboratory setting (manuscript in preparation) was carried out using two simultaneously-recorded remote data collection procedures. A secondary goal was to determine the extent of improvement to data quality that could be attributed to replacing an automatic formant extraction script with one that required approval by the researcher of each data point based on a visual presentation of the formant tracker and reading. Results were analyzed using four measures of vowel reduction: Comparison of stressed and unstressed F1 and F2 values and visual inspection of a vowel plot, Euclidean distance, Pillai scores, and duration ratio of stressed and unstressed vowels.

2. Methods

2.1. Participants

Twenty monolingual speakers of American English participated in the study; ten completed the task in person in a laboratory setting (2M, 8F; M = 20.9 y.o., SD = 1.96, range = 19 – 26), and ten completed it fully remotely (6M, 4F; M = 35.6 y.o., SD = 10.65, range = 21 – 60). Participants were born and resided in the Midwestern region of the United States at the time of recording; 19 of the 20 reported only limited exposure to other languages. One speaker reported basic proficiency in Tamil, spoken as a heritage language, but described their proficiency as primarily receptive, with spoken abilities limited to domestic topics. Data was collected from in-person and remote participants separately as a result of the COVID-19 pandemic; in-person data also served as a control group for a separate study of vowel reduction in Spanish-English bilinguals (in preparation). The remote group was recruited via Prolific (www.prolific.com) and completed the task through Gorilla Experiment Builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020), while the in-person group was recruited on the campus of a major Midwestern research university and recorded in a sound-insulated booth.

2.2. Stimuli

Participants completed a repetition task in which they heard and repeated English sentences of the general form “the word X means…”. A full list of the sentences is given in Supplemental Materials, Appendix 1: Each contained a single target word with one of five vowels /ɑ, æ, ɛ, ɪ, ʌ/, followed by a predicate defining the target word of approximately the same length, rhythm, and difficulty (i.e., “the word rabbit means a small fuzzy animal”). These vowels were selected to flesh out the periphery of the vowel space as fully as possible using monophthongal vowels that undergo reduction; no high back vowels were included because of the difficulty of finding target words with reduced vowels that could reasonably be assumed to contain an underlying high back vowel. Target words placed the vowel in question in either stressed or unstressed position; all stressed vowels occurred in monosyllables, while unstressed vowels appeared in the second syllable of a disyllable. Monosyllables and disyllables were matched for the segmental content surrounding the target vowel – for example, “bit” ~ “rabbit”. Each vowel was represented by six monosyllables and six disyllables for a total of 60 target words.

Stimuli were recorded by a male native speaker of Midwestern American English (28 y.o. at the time of recording) in a sound-insulated booth using a Shure KSM32 cardioid condenser microphone and a TubeMP preamp connected to a PC, and digitized at 44.1 kHz. The intensity of the recordings was normalized to 70 dB via Praat script (Winn, 2013).

2.3. Procedure and recording equipment

Participants completed a shadowing task in which they heard and repeated target sentences for a microphone. The 60 target sentences were repeated twice and randomized for a total of 120 intended utterances per participant. Participants were instructed to repeat the sentence they heard, speaking as naturally as possible. Visual indicators cued participants when it was time to speak, with a red screen and image of an ear indicating “listen” and a green screen and image of a microphone indicating “speak”. A delay of 0.5 seconds was added after the recorded stimulus before the “speak” cues appeared. Together, the delay before speaking and the longer and variable carrier phrases were intended to encourage speakers not to reproduce the prosody or other subphonemic details of the recorded prompt, in line with work that has shown that when a sentence exceeds working memory capacity, it must be parsed in order to be repeated successfully (Tracy-Ventura, McManus, Norris, & Ortega, 2014). The recording window ended after five seconds, unless the participant advanced to the next trial manually before the full five seconds elapsed. After each block of 30 sentences, participants were offered a short break.

Those in the in-person group completed the study in a sound-insulated booth using Sennheiser HD 380 Pro headphones, with stimulus presentation managed via PsychoPy (Peirce, Gray, Simpson, MacAskill, Höchenberger, Sogo, Kastman, & Lindeløv, 2019). A Shure KSM32 cardioid condenser microphone connected to a TubeMP preamp was used to record their productions in Audacity, and recordings were digitized at 44.1 kHz.

Remote participants completed the task once, but produced two simultaneous recordings obtained via distinct methods to allow for comparison of online and offline remote recording approaches. They were instructed to complete the task in a quiet room and to wear headphones during the study, but natural variation in background noise occurred across participants and over the course of the study. Headphone compliance was ensured via a simple headphone check task utilizing antiphase tones, as described in Woods, Siegel, Traer, and McDermott (2017). Stimulus presentation and online recording was managed via Gorilla Experiment Builder (Anwyl-Irvine et al., 2020) on a laptop or desktop computer. The Gorilla recordings made use of Gorilla’s Audio Recording Zone to capture .weba files, a lossy file format that relies on the OGG Vorbis compression codec (*What Is a WEBA File?*, n.d.). The Gorilla recordings utilized either a built-in or external microphone; models varied across participants. According to self-report, three participants used a microphone built into a headset, two used an external cardioid condenser microphone, two used microphones built into earbuds, one used a built-in laptop microphone, one used an external microphone of unknown type, and one did not report microphone information. Exact models are reported in the Supplemental Materials, Appendix 2. One participant appeared to suffer a microphone error, as all of their Gorilla files were empty; thus, the Gorilla group consisted of only nine participants in the final analysis.

In addition to the online .weba recordings captured via Gorilla, remote participants simultaneously recorded their productions offline in a lossless .wav file recorded using a smartphone. For this recording, participants were directed to download a freely-available audio recording app (Hokusai Audio Editor for Apple users and ASR Voice Recorder for Android owners). They were instructed to adjust the app settings to an acceptable quality (a 128-kbps bitrate for Android users and a 16-bit setting for Apple owners, with a 44.1 kHz sampling rate for both) and to position the smartphone and any freestanding external microphone six to ten inches from their mouth in a stable position. The .wav file was sent to the researcher upon completion of the task.

2.4. Analysis

Target vowels were annotated by hand in both the uploaded smartphone .wav files and Gorilla .weba recordings. To mark boundaries between voiceless consonants and vowels, the onset of periodicity was used as the primary cue to determine the location of the boundary. Where voiced consonants were concerned, increases in intensity and complexity of waveform were used to determine the boundary between consonant and vowel. Where no clear boundary existed, only the steady-state portion of the vowel was annotated, and the item was excluded from analyses of duration. (This group included all utterances of ‘carrot’ and ‘rot’ and many instances of ‘lot’ and ‘pilot’.) Formant frequencies at vowel midpoint and duration of vowels were extracted twice: Once using a Praat script that automatically extracted formant values, and a second time using a Praat script that required manual review of each item (Scarborough, 2005), allowing the researcher to visually confirm that the formants extracted matched the visible formant on the spectrogram. When the formant tracker was visibly distorted, a manual reading was taken by using the FFT-based ‘Get Formants…’ function or by placing the cursor manually in the center of the visible formant. Both scripts relied on Praat’s Burg LPC-based algorithm for formant extraction. Since the study focused on how recording quality impacted analysis and low-quality recordings were an expected aspect of the study design, it was expected that manual review could be critical to ensuring accurate formant readings. Finally, formant frequencies were normalized using log-additive regression normalization (Barreda & Nearey, 2018).

To construct a well-rounded picture of the results, a variety of approaches to quantifying vowel reduction are presented: §3 considers raw formant values, relative position in the vowel space, Euclidean distance, Pillai scores, and duration ratio of reduced and unreduced items. For each measure, a linear mixed-effects model was fit using *lme4* (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2020), including some combination of the factors Extraction Method (Automatic vs. Supervised), Recording Method (In Person, Gorilla, or Smartphone), and Vowel (five levels: /ɑ, æ, ɛ, ɪ, ʌ/), according to which combination yielded the best model fit. Model fit was assessed by comparing nested models using a likelihood ratio test (LRT) and comparing non-nested models using the Akaike Information Criterion (AIC). The models for raw formant values also included a factor Stress (Stressed vs. Unstressed); the others did not include this factor because the difference in stress was worked into the dependent measure for Euclidean distance, Pillai score, and duration ratio.

3. Results

3.1. Vowel position

To fully convey the spectral effects of reduction on vowel position under the various conditions of interest, this section presents the changes enacted to F1, F2, and overall position in the F1xF2 space across recording conditions and extraction methods.

3.1.1. Effect on F1

Figure 1 displays the F1 range of each vowel in stressed and unstressed position across the six conditions of extraction method and recording method. As expected, all vowels except /ɪ/ visibly raised when unstressed (with /ɪ/ an exception due to its initial position resting quite high in the vowel space).

A linear mixed-effects model with a dependent variable of Normalized F1 was fit using *lme4* (Bates et al., 2015). The best-fitting model, determined through comparison of nested models using LRT tests and comparison of non-nested models by AIC, contained four independent terms – Extraction Method (Automatic vs. Supervised), Recording Method (In Person, Gorilla, or Smartphone), Vowel (five levels: /ɑ, æ, ɛ, ɪ, ʌ/), and Stress (Stressed or Unstressed) – as well as the interactions Extraction Method by Recording Method, Recording Method by Vowel by Stress, Recording Method by Vowel, Vowel by Stress, and Recording Method by Stress. It also included random intercepts for Subject and Item and a random slope for Subject in relation to Vowel. This model showed no significant difference between the Automatic and Supervised extraction methods ($\beta = -.001$, $SE = .004$, $t = -.261$, $p = 0.794$), although a difference did emerge for Recording Method when comparing the Smartphone and In Person recordings ($\beta = -.080$, $SE = .017$, $t = -4.752$, $p < .001$). (The Smartphone recordings yielded a lower average F1, corresponding to a more closed vowel on average.) To better understand the complexities of the data, pairwise comparisons (post-hoc t-tests with Bonferroni correction) were run comparing each vowel across recording conditions, holding fixed the levels of Extraction Method and Stress. Results of pairwise comparisons are provided in **Table 1**. (The full output of the omnibus model is given in Supplemental Materials, Appendix 3.) As can be observed in **Table 1**, relatively few deviations from the in-person recordings appeared, with only unstressed /ɪ/ differing significantly between the in-person and Gorilla recordings (higher in Gorilla than in person), and only stressed /ɑ/ and /ʌ/ differing in the smartphone recordings (both lower in the smartphone recordings than the in-person ones). Interestingly, a greater number of deviations appeared between the two remote recording methods, which analyzed data recorded simultaneously from different devices, than between the data obtained through either of the remote methods and the in-person data, which was recorded by different speakers in a laboratory setting.

	Gorilla vs. In Person		**Smartphone vs. In Person**		**Smartphone vs. Gorilla**	
Automatic /ɑ/ Stressed	–0.0142		–0.0822	***	–0.068	***
Automatic /æ/ Stressed	–0.0308		–0.0891		–0.0583	
Automatic /ɛ/ Stressed	0.0235		–0.00293		–0.0264	
Automatic /ɪ/ Stressed	0.0281		–0.0169		–0.045	**
Automatic /ʌ/ Stressed	–0.00863		–0.0764	***	–0.0678	***
Automatic /ɑ/ Unstressed	0.0332		–0.0178		–0.051	
Automatic /æ/ Unstressed	0.0205		–0.0291		–0.0496	
Automatic /ɛ/ Unstressed	0.0662		0.0108		–0.0554	
Automatic /ɪ/ Unstressed	0.0556	***	–0.0141		–0.0697	***
Automatic /ʌ/ Unstressed	0.0423		–0.00558		–0.0479	
Supervised /ɑ/ Stressed	–0.0223		–0.0578	***	–0.0355	
Supervised /æ/ Stressed	–0.0351		–0.0648		–0.0297	
Supervised /ɛ/ Stressed	0.0247		0.00882		–0.0159	
Supervised /ɪ/ Stressed	0.0363		–0.00909		–0.0454	**
Supervised /ʌ/ Stressed	–0.0144		–0.0419	***	–0.0276	
Supervised /ɑ/ Unstressed	0.0348		0.00729		–0.0275	

(Contd.)

	Gorilla vs. In Person		**Smartphone vs. In Person**		**Smartphone vs. Gorilla**	
Supervised /æ/ Unstressed	0.0246		–0.0149		–0.0395	
Supervised /ɛ/ Unstressed	0.0731		0.0174		–0.0558	
Supervised /ɪ/ Unstressed	0.0633	***	–0.00386		–0.0672	***
Supervised /ʌ/ Unstressed	0.0492		0.0279		–0.0214	

Table 1: Pairwise comparisons for F1. The first column indicates the difference in means between the two groups listed. The second column indicates the significance of the p-value of a Bonferroni-corrected t-test between the two groups. Alpha level is set for 0.0025 due to the use of multiple tests – values above this are reported as NS. ** indicates $p < 0.0025$ and *** indicates $p < 0.001$. Blank cells indicate a non-significant result.

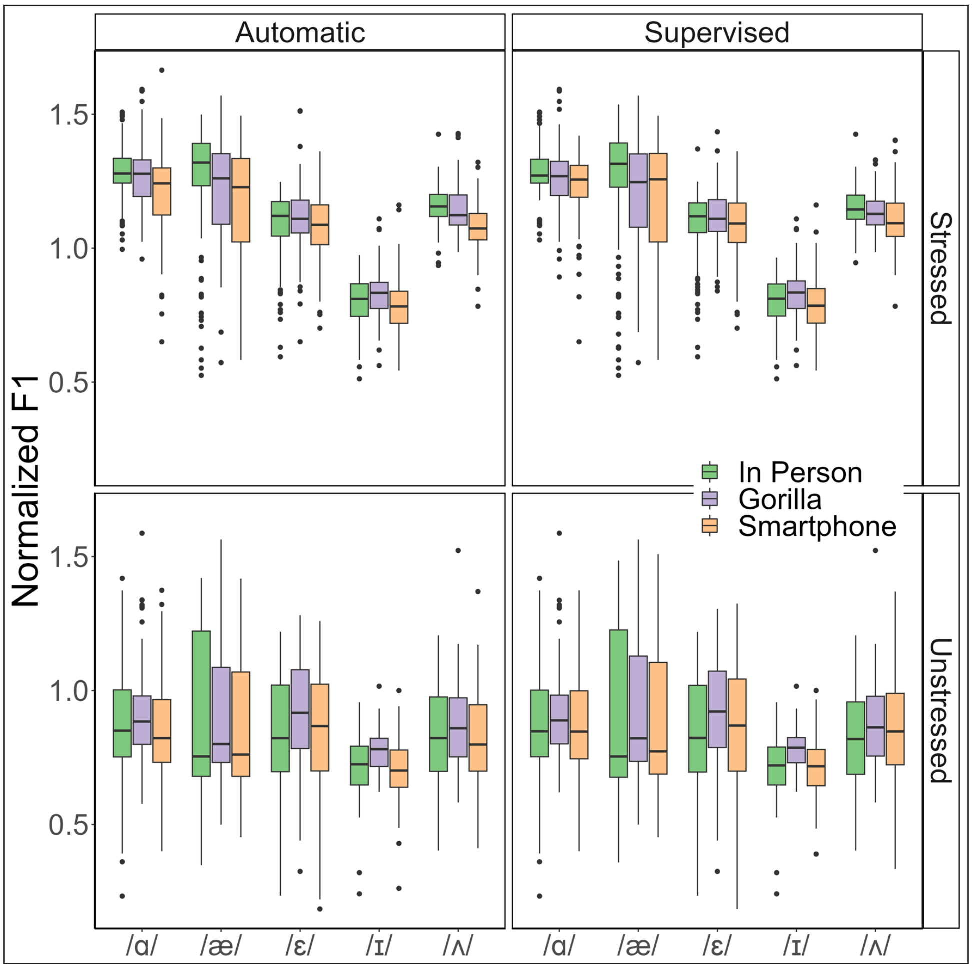

Figure 1: Changes to F1 by Vowel, Stress, Extraction Method, and Recording Method.

Some of this variation – such as the wider range of unstressed /æ/ in the Gorilla and Smartphone recordings – may be due to interspeaker differences, as the in-person recordings were, due to logistical constraints and the desire to run a true replication study remotely rather than a direct comparison, gathered from a separate set of participants. Other variations, such as the elevated F1 in Gorilla for both stressed and unstressed /ɪ/ relative to the smartphone recordings, cannot be ascribed to between-speaker variation and thus must be attributed to differences in hardware and compression algorithm.

3.3.2. Effect on F2

Figure 2 visualizes the F2 range of each vowel in stressed and unstressed position across the six conditions of extraction method and recording method. As was done for F1, to analyze F2, a linear mixed-effects model with a dependent variable of Normalized F2 was fit using *lme4* (Bates et al., 2015). The best-fitting model contained four fully-crossed independent terms: Extraction Method (Automatic vs. Supervised), Recording Method (In Person, Gorilla, or Smartphone), Vowel (five levels: /ɑ, æ, ɛ, ɪ, ʌ/), and Stress (Stressed or Unstressed), and all possible interactions among these four, as well as random intercepts for Subject and Item and a random slope for Subject in relation to Vowel. The model showed no significant difference between the Automatic and Supervised extraction methods (β = .004, SE = .013, t = .285, p = .776), although a difference did emerge for Recording Method when comparing the Gorilla and in-person recordings (β = .057, SE = .025, t = 2.293, p < .05). (The Gorilla recordings yielded a lower average F2, corresponding to a more back vowel quality on average.) The full output of the model is given in Supplemental Materials, Appendix 3.

To better convey the nuances of the differences across recording methods, post-hoc t-tests with Bonferroni correction were used to compare each vowel across recording methods, with the levels of stress and extraction method held constant. The results are displayed in **Table 2**: As can be seen, only stressed /ɑ/ differed from the in-person data in the Gorilla recordings (with a higher F2 obtained via Gorilla), while both stressed /ɑ/ and stressed /æ/ differed from the control group in the smartphone recordings, but only when measured with the supervised script, with both exhibiting a higher F2 in the smartphone recording than the control data. Two differences also emerged between the smartphone and Gorilla recordings, affecting unstressed /ɛ/ and /ɪ/ when measured automatically and manually, respectively.

3.3.3. Overall vowel position

To provide a more visually intuitive view of the changes to formant values due to reduction across conditions, **Figure 3** provides a vowel plot for each of the six conditions. As can be seen, differences between the Automatic and Supervised extraction method plots for the Gorilla and

	Gorilla vs. In Person		**Smartphone vs. In Person**		**Smartphone vs. Gorilla**	
Automatic /ɑ/ Stressed	0.0597	***	0.0255		–0.0341	
Automatic /æ/ Stressed	0.0551		0.0488		–0.00629	
Automatic /ɛ/ Stressed	–0.0266		–0.0126		0.0139	
Automatic /ɪ/ Stressed	–0.0054		–0.0239		–0.0185	
Automatic /ʌ/ Stressed	–0.018		–0.0495		–0.0315	
Automatic /ɑ/ Unstressed	–0.0456		–0.0414		0.00422	
Automatic /æ/ Unstressed	–0.0509		–0.0108		0.0401	
Automatic /ɛ/ Unstressed	–0.0463		0.00907		0.0554	**
Automatic /ɪ/ Unstressed	–0.033		–0.0194		0.0136	
Automatic /ʌ/ Unstressed	–0.0185		–0.0398		–0.0214	
Supervised /ɑ/ Stressed	0.0583	***	0.051	***	–0.00724	
Supervised /æ/ Stressed	0.0597		0.0889	***	0.0293	
Supervised /ɛ/ Stressed	–0.0185		0.00282		0.0213	
Supervised /ɪ/ Stressed	0.01		0.0146		0.0046	
Supervised /ʌ/ Stressed	–0.02		–0.00324		0.0168	
Supervised /ɑ/ Unstressed	–0.0215		0.0131		0.0346	
Supervised /æ/ Unstressed	–0.0103		0.0139		0.0242	
Supervised /ɛ/ Unstressed	–0.0171		0.0164		0.0336	
Supervised /ɪ/ Unstressed	–0.00639		0.0245		0.0309	**
Supervised /ʌ/ Unstressed	–0.0036		0.0141		0.0177	

Table 2: Pairwise comparisons for F2. The first column indicates the difference in means between the two groups listed. The second column indicates the significance of the p-value of a Bonferroni-corrected t-test between the two groups. Alpha level is set for 0.0025 due to the use of multiple tests – values above this are reported as NS. ** indicates $p < 0.0025$ and *** indicates $p < 0.001$. Blank cells indicate a non-significant result.

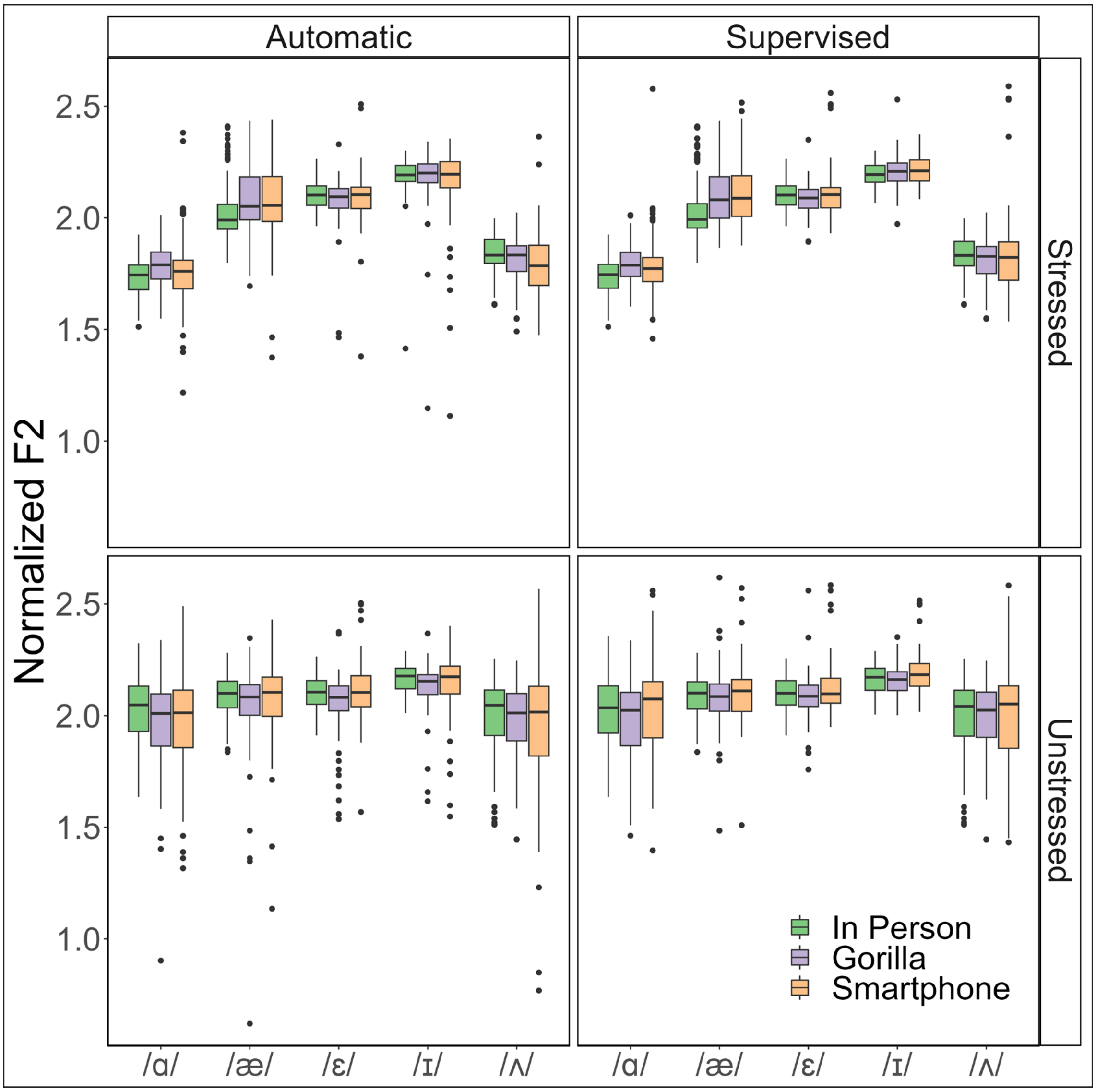

Figure 2: Changes to F2 by Vowel, Stress, Extraction Method, and Recording Method.

In Person data are minimal; in the Smartphone data, some differences are visible, with the Supervised formant script generating less variation in stressed /ɑ/, greater variation in F2 and less variation in F1 for stressed /ʌ/, and less variation in stressed /ɪ/ compared to the Automatic script, as well as a tighter grouping of means for stressed and unstressed /ɪ/ in the Supervised condition. Comparison to the other recording conditions suggests that the data obtained from the Supervised script for the Smartphone condition is most similar to the data found in the Gorilla and In Person recordings from either script, and thus may be more accurate overall.

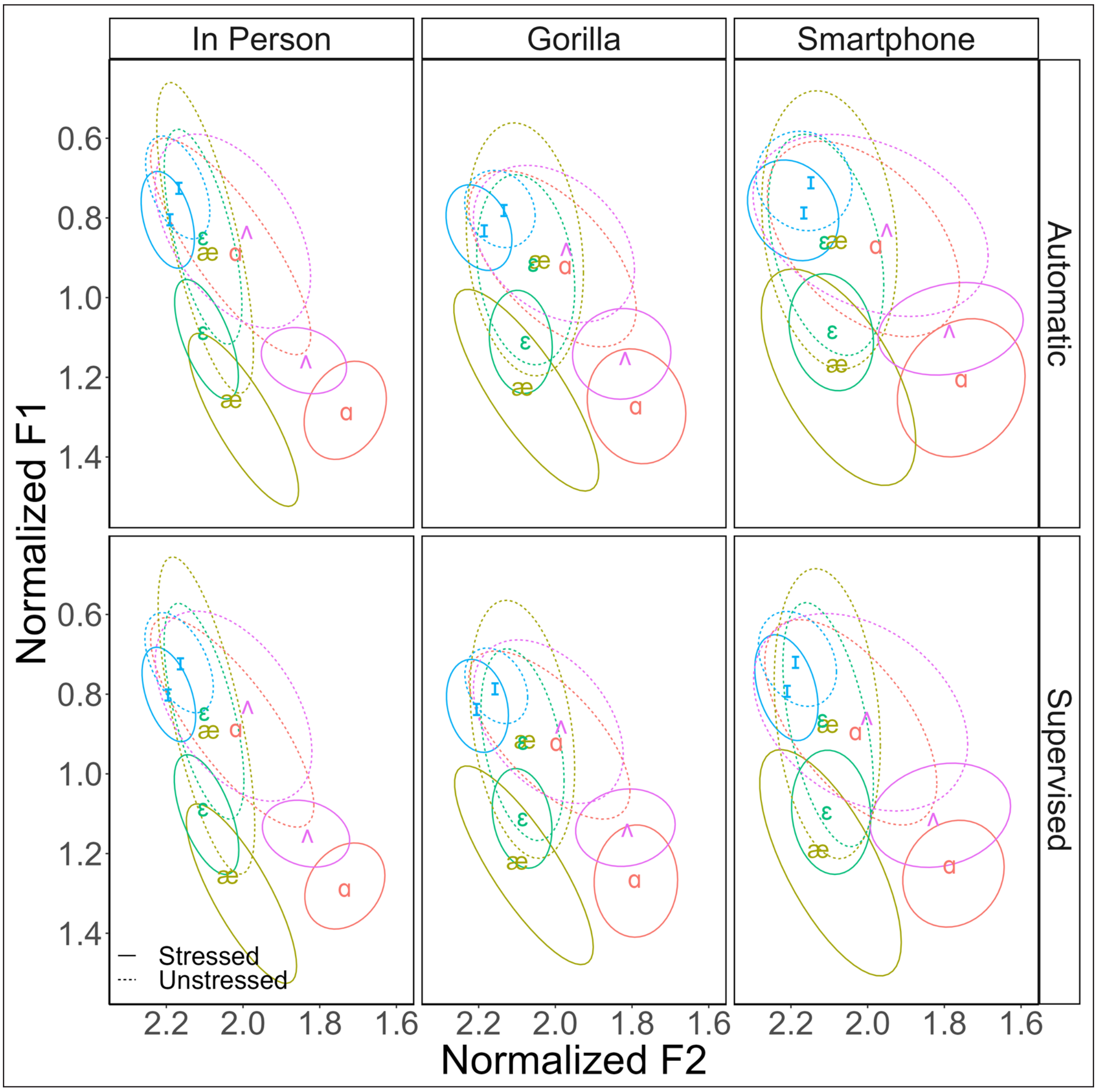

Figure 3: Vowel plot by extraction method, recording method, vowel, and stress. Ellipses show approximately one standard deviation around the mean (67% confidence interval).

3.2. Euclidean distance

To create a singular metric conveying the degree of spectral displacement in the F1xF2 space that could be attributed to vowel reduction, the Euclidean Distance (EuD) between each stressed vowel and its unstressed counterpart was calculated for each speaker and word pair. EuD is a straightforward calculation of the hypotenuse of the triangle created by measuring the distance between two points in the F1 and F2 dimensions. The first stressed and unstressed productions of each word were paired together, and the second stressed and unstressed productions were paired together (i.e., for a given speaker, the first utterance of 'rabbit' and the first utterance of 'bit' contributed to the calcuation of a single EuD measurement). If either member of a pair was missing, no EuD value was computed for that speaker, pair, and repetition. A total of 3342 EuD values were calculated from the 6813 recorded vowels whose formant values were included in the analysis in §3, indicating a rate of about 1.9% data loss due to missing values within pairs.

Figure 4 displays the Euclidean distance of each vowel when recorded in person, using Gorilla, or using a smartphone, as well as when the formant values were extracted using the automatic or supervised scripts. To determine which of these factors had a significant impact on the EuD values, a linear mixed model with a dependent variable EuD, independent variables for Vowel (five levels: /ɑ, æ, ɛ, ɪ, ʌ/), Recording Method (Gorilla, Smartphone, or In Person), and Extraction Method (Automatic or Supervised), an interaction term for Vowel By Recording Method, and random intercepts for Item and Subject was fit using *lme4* (Bates et al., 2015). (Models using other combinations of factors were considered, and this model was selected as the best fit based on comparison of nested models through LRT tests.) The full output of the model is reported in Supplemental Materials, Appendix 3. Importantly, the greatest differences in Euclidean Distance were a result of vowel quality, with the highest vowel /ɪ/ undergoing the least displacement in the F1xF2 space as a result of reduction, and the low vowels /ɑ/ and /æ/ undergoing the greatest. Significant differences in overall EuD across vowels also emerged for the Gorilla recordings in comparison to the In Person data ($\beta = -.088$, SE = .029, $t = -2.80$, $p < 0.01$), as well as between the Automatic and Supervised formant extraction methods ($\beta = -.009$, SE = .005, $t = -2.080$, $p < 0.05$). No difference emerged between the Smartphone and In Person recordings ($\beta = -.047$, SE = .029, $t = -1.607$, $p = 0.121$).

Pairwise comparisons with Bonferroni correction of each vowel across recording methods and extraction methods revealed no significant differences, even though some minor differences in the degree of difference between vowels across recording conditions emerge in the Vowel by Recording Condition interaction captured in Lines 9 – 16 of **Table 3**, Appendix 3. Where significant, these represent minor adjustments to the degree of change in the difference in EuD between the reference vowel /ɑ/ and the vowel in question between the reference recording condition (In Person) and the recording condition cited in that line. Thus, although statistically significant, they do not offer substantive insight into the central question of this study: Do the main conclusions of a vowel reduction analysis vary across extraction methods or recording conditions? With regard to EuD, the most salient takeaway is that automatically extracted formants yielded a greater EuD than formants extracted with the supervised script, while recordings taken via Gorilla yielded a slightly smaller EuD than those taken in person.

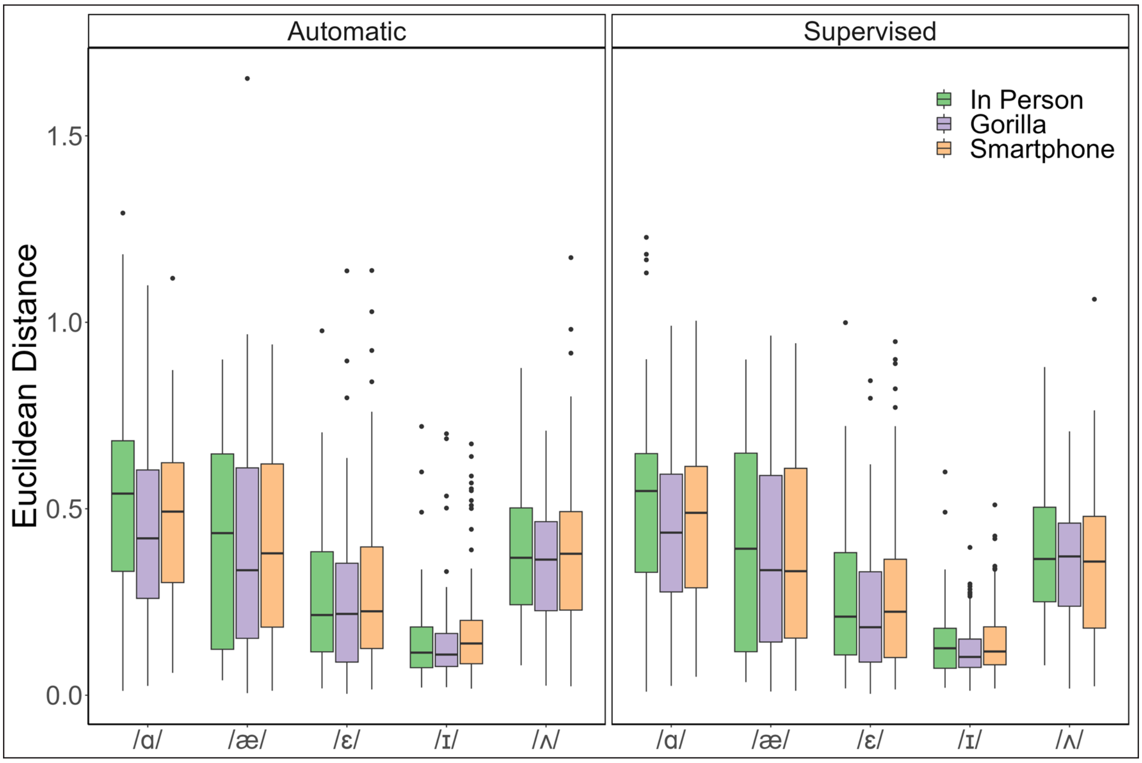

Figure 4: Euclidean distance of stressed and unstressed English vowels by recording condition and extraction method.

3.3. Pillai scores

While Euclidean distance and raw formant values provide important information about spectral reduction, neither delivers a complete picture of all the effects of reduction. Raw formant values are difficult to interpret, and they split the acoustic space into two dimensions that, while well-grounded in the physical production of vowels, do not intuitively capture the centralizing movement of the tongue that occurs in reduction. By contrast, EuD collapses the artificially divided F1 and F2 dimensions into a single vector but fails to represent the direction of the vector or its extent relative to either the stressed or unstressed vowel as a baseline. To overcome these limitations, this section reports the results of an analysis of Pillai scores for each stressed vowel and its unstressed counterpart.

Pillai scores provide a measure of overlap between the distributions of two vowels; the score ranges from 0 to 1, with 0 indicating the highest possible degree of overlap and 1 indicating no overlap (Hay, Warren, & Drager, 2006; Nycz & Hall-Lew, 2013). In this analysis, the degree of overlap between all stressed tokens of a vowel and all unstressed tokens of the same vowel was calculated for each speaker in each of the six conditions (Automatic or Supervised formant extraction and In Person, Gorilla, or Smartphone recordings). The Pillai scores are visualized in **Figure 5** and were analyzed using a linear mixed model. The best-fitting model contained independent factors for Vowel (five levels: /ɑ, æ, ɛ, ɪ, ʌ/) Recording Method (Gorilla, In Person, or Smartphone), and their interaction, as well as a random intercept for Subject and a random slope for Subject in relation to Vowel; the output of this model is given in **Table 3**. As Extraction Method did not improve the fit of the model, we can conclude that the use of the automated formant extraction script compared to the supervised script did not materially affect the degree of overlap between reduced and unreduced vowels. Furthermore, there were clear differences between vowels; /ɑ/ and /ʌ/ showed the least degree of overlap, which makes sense as these vowels underwent the greatest degree of raising, as discussed in §3.1.1. The high vowel /ɪ/ showed the greatest degree of overlap, which was again unsurprising considering its representation in **Figure 3**, while /æ/ and /ɛ/ hovered in the middle of the Pillai score range, indicating some overlap. Vowels with more overlap overall also exhibited greater variation across speakers, as judged by the size of the quartiles in **Figure 5**, suggesting that some speakers maintained a stronger distinction between stressed and unstressed /ɪ/ (for example) than others did. Finally, although Recording Method improved the fit of the model, no significant effect of Recording Method emerged, suggesting that the degree of overlap for the dataset on the whole was stable regardless of recording method. Pairwise comparisons with Bonferroni correction of each vowel across recording methods and extraction methods confirmed this conclusion, revealing no significant differences.

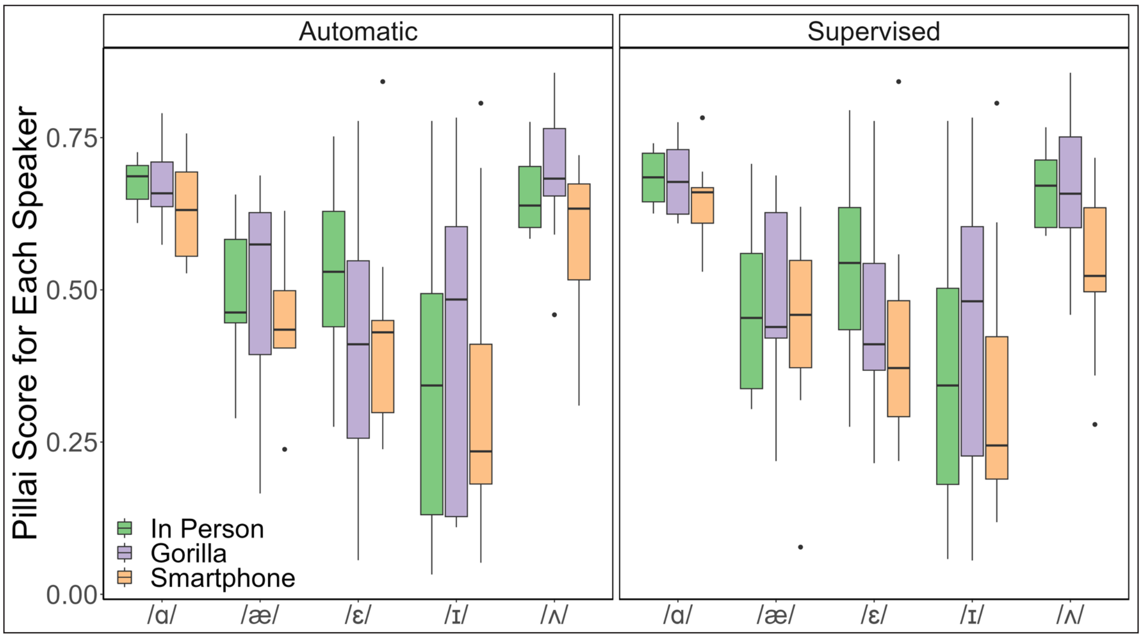

Figure 5: Range of Pillai scores across speakers for extraction and recording method, calculated by comparing each stressed vowel and its unstressed counterpart.

	Term	Estimate	Standard Error	*t* Statistic	*p*
1	(Intercept)	0.680	0.021	32.378	***
2	Vowel /æ/	–0.206	0.047	–4.333	***
3	Vowel /ɛ/	–0.145	0.057	–2.539	*
4	Vowel /ɪ/	–0.325	0.081	–3.990	***
5	Vowel /ʌ/	–0.019	0.034	–0.574	
6	Recording Method Gorilla	–0.009	0.030	–0.309	
7	Recording Method Smartphone	–0.043	0.030	–1.448	
8	Vowel /æ/ * Recording Method Gorilla	0.023	0.068	0.334	
9	Vowel /ɛ/ * Recording Method Gorilla	–0.120	0.081	–1.472	
10	Vowel /ɪ/ * Recording Method Gorilla	0.077	0.115	0.666	
11	Vowel /ʌ/ * Recording Method Gorilla	0.003	0.048	0.067	
12	Vowel /æ/ * Recording Method Smartphone	–0.002	0.067	–0.028	
13	Vowel /ɛ/ * Recording Method Smartphone	–0.071	0.081	–0.872	
14	Vowel /ɪ/ * Recording Method Smartphone	0.018	0.115	0.160	
15	Vowel /ʌ/ * Recording Method Smartphone	–0.059	0.048	–1.240	

Table 3: Output of linear mixed model evaluating Pillai scores calculating overlap between stressed vowels and their unstressed counterparts. The reference category for Vowel is /ɑ/ and for Recording Method is In Person. * indicates p < 0.05, ** p < 0.01, and *** p < 0.001. Blank cells in the rightmost column indicate a non-significant result.

3.4. Duration ratio

To create a single measure incorporating the duration values of paired stressed and unstressed vowels, the Duration Ratio was computed as the quotient of Unstressed Duration/Stressed Duration. This approach yields a measure where a value of 1 indicates no reduction for a pair of utterances, values greater than 1 indicate that the unstressed vowel was longer than the stressed vowel, and values less than 1 denote temporal reduction – the smaller the value, the greater the reduction. **Figure 6** displays the mean duration ratio for each Vowel, Extraction Method, and Recording Method; to analyze the results, a linear mixed model was fit. The best-fitting model contained an independent variable for Vowel (five levels: /ɑ, æ, ɛ, ɪ, ʌ/), Recording Method (Gorilla, Smartphone, or In Person), and their interaction, as well as random intercepts for Subject and Item and a random slope for Subject in relation to Vowel; output is given in **Table 4**. The lack of a significant effect for Vowel, Extraction Method, or Recording Method indicates that major differences in temporal reduction were not present.

This conclusion is not unexpected – if, indeed, changes were found to be due to extraction method, some error in the analytical process would have to be expected, since both scripts extracted duration in an identical fashion from identically annotated, simultaneous recordings. That no differences emerged due to recording method is a reflection of a consistent annotation style if we consider the Gorilla and Smartphone recordings, which annotated the same utterances recorded with different devices and filetypes. While the difference between lossy online .weba recording and lossless offline .wav recording may be expected to produce some spectral distortion, it should not effect change to the temporal dimension. Furthermore, the lack of difference between the In Person recordings and the Gorilla/Smartphone recordings suggests that the sampling techniques and sample size in the study were sufficient to disguise any interspeaker variation or random variation ascribable to individual utterances.

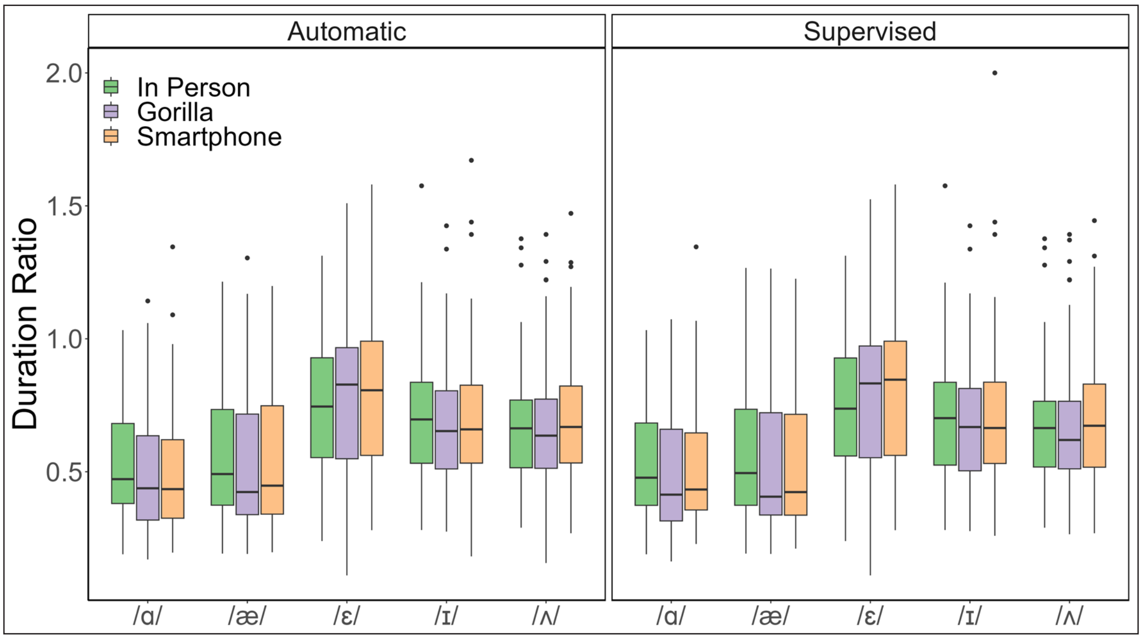

Figure 6: Duration ratio of paired vowels (Unstressed/Stressed) by extraction method and recording method.

	Term	**Estimate**	**Standard Error**	***t* Statistic**	***p***
1	(Intercept)	0.522547	0.104845	4.984	***
2	Vowel /æ/	0.030123	0.13431	0.224	
3	Vowel /ɛ/	0.219348	0.134293	1.633	
4	Vowel /ɪ/	0.179581	0.135675	1.324	
5	Vowel /ʌ/	0.150429	0.135264	1.112	
6	Recording Method Gorilla	–0.026323	0.030269	–0.87	
7	Recording Method Smartphone	–0.014251	0.029937	–0.476	
8	Vowel /æ/ * Recording Method Gorilla	0.000339	0.031002	0.011	
9	Vowel /ɛ/ * Recording Method Gorilla	0.057895	0.030772	1.881	
10	Vowel /ɪ/ * Recording Method Gorilla	–0.001007	0.041192	–0.024	
11	Vowel /ʌ/ * Recording Method Gorilla	0.02857	0.038445	0.743	
12	Vowel /æ/ * Recording Method Smartphone	–0.00633	0.030535	–0.207	
13	Vowel /ɛ/ * Recording Method Smartphone	0.056972	0.030273	1.882	
14	Vowel /ɪ/ * Recording Method Smartphone	0.017694	0.040782	0.434	
15	Vowel /ʌ/ * Recording Method Smartphone	0.039408	0.037948	1.038	

Table 4: Output of linear mixed model evaluating Duration Ratio (Unstressed/Stressed) between stressed vowels and their unstressed counterparts. The reference category for Vowel is /ɑ/, for Recording Method is In Person, and for Extraction Method is Automatic. * indicates p < 0.05, ** p < 0.01, and *** p < 0.001. Blank cells in the rightmost column indicate a non-significant result.

3.5. Summary of results

This study relied on a range of metrics to evaluate differences in vowel reduction across two remote recording methods and two formant extraction methods.

When examining F1 directly, recordings taken via smartphone had a slightly lower average F1 than the in-person recordings. For data taken via Gorilla, F1 was slightly higher for unstressed /ɪ/ than in the control data, while smartphone recordings slightly lowered F1 in low back vowels. For F2, the Gorilla recordings had a slightly (and statistically significantly) higher mean F2. Both remote recordings raised the F2 of /ɑ/ relative to the control, while smartphone recordings also raised the F2 of /æ/. However, the relative position in the vowel space, averaged across speakers, was quite stable across conditions, with the single exception of the smartphone recordings analyzed with the automated formant extraction script. The smartphone recordings analyzed with the supervised script strongly reflected the Gorilla and In Person recordings when considered in the aggregate.

When it came to Euclidean distance (EuD), some statistically significant differences emerged beyond the expected. Vowel quality, naturally, had a strong impact on EuD. Additionally, the data extracted via the Automatic script had a greater EuD than that taken from the Supervised script, and the Gorilla recordings had a smaller EuD overall than the in-person recordings. Here again, the two lowest vowels underwent the greatest change across conditions.

For Pillai scores, only vowel quality predicted differences across categories; there was no significant effect of recording method or extraction method, suggesting that the degree of overlap was fairly stable across methodologies. Visual inspection of the data did suggest a wider spread of Pillai scores across participants in the smartphone recordings for the vowel /v/ under the Automatic extraction method.

Finally, Duration Ratio showed no significant differences due to recording method or extraction method, suggesting that temporal reduction was unaffected by these changes in data collection techniques.

4. General discussion

The core question of this study was whether the use of remote data collection for speech production would impact the conclusions of an acoustic analysis of vowel reduction. Two remote data collection methods, lossy online recording via Gorilla and lossless offline recording via a smartphone app, were compared to data collected in person in a traditional laboratory setting and subjected to a range of statistical analyses. Both remote recording methods diverged from the control data in statistically significant ways, but the differences were not identical across methods. The smartphone recordings aligned best with the control data when looked at through the lens of Euclidean distance, while the Gorilla recordings exhibited fewer changes to normalized F1 and F2 values. Especially notable when considering the F1 and F2 values were the clustering of significant differences in the low and back vowels when comparing the smartphone data to the control group, an area of the vowel space identified in previous work as particularly susceptible to distortion due to recording and audio compression techniques. No differences attributed to recording method emerged when using Pillai scores to parametrize vowel reduction, and apparent vowel position, evaluated visually, was also relatively consistent. Thus, the form of remote recording most appropriate to a given study should be evaluated in line with the vowels of interest and intended analysis, as some approaches to vowel reduction were more resilient than others to differences in recording method, and no single remote recording method yielded results identical to the control data under all analytical approaches.

Testing the impact of formant extraction using an automatic Praat script compared to one visually confirmed by a researcher is an important contribution of the present study. Especially when using recordings likely to be affected by background noise, high SNR, and potentially poor recording quality, it is reasonable to expect that the manually corrected formant extraction procedure may yield higher-quality data. Interestingly, the degree of impact of the two formant extraction procedures differed by analysis and recording method: Overall vowel position was notably impacted by extraction method in the smartphone recordings, with the supervised data falling more closely in line with the control data, while little impact was observed on the Gorilla data from extraction method. Euclidean distance was also affected by extraction method, while Pillai scores were resilient to the slight differences introduced by manual correction of the data. These findings suggest that while the supervised formant extraction script did introduce improvements to data quality, it was primarily in the smartphone recordings that those improvements had a material impact on the conclusions of the analysis.

One finding that emerged repeatedly in earlier work on distortions likely in remote data collection due to hardware and software differences was that the lower back quadrant of the vowel space is especially susceptible to distortion (Freeman & de Decker, 2021a, 2021b). This finding is repeated in the present study, but in an unexpected fashion: The lossless smartphone recordings distorted this portion of the vowel space more strongly than the lossy Gorilla recordings. While the distortion is surprising when considering only the compression codec used in each condition, it is important to remember that file format and microphone type are not the only factors implicit in the choice between a lossy Gorilla recording and a lossless smartphone recording. Modern smartphones typically perform noise cancellation on all incoming sound, both calls and recordings, using sound localization from multiple built-in microphones to detect and correct for background noise (Thorn, 2014).[1] Future research should consider whether the noise

[1] Thanks are due to Ashley Kentner for this astute observation.

cancellation software included in most smartphones could be responsible for some or all of the distortions detected in the lossless recordings.

4.1. Clarifying the intended goal of the current study

Two aspects of this study that may be considered weaknesses from one perspective are (1) the use of different speaker groups for the in-person and remote data and (2) the conflation of recording method and recording environment that occurred due to remote data collection being truly remote, rather than a simulacrum of remote recording techniques conducted in a laboratory as often seen in previous work. If the goal of this study were to evaluate the efficacy of remote recording setups without interference from outside forces (background noise, microphone position, etc.), these would be weaknesses indeed. However, viewed as a remote replication of an original in-person study, whose purpose is to advise researchers of the challenges and degree of variation they may expect to encounter when working with remotely-collected data beyond the already well-documented differences of device type and file format, these apparent weaknesses should more rightly be considered strengths. The differentiation of speaker groups fulfills the purpose it would in any replication study, ensuring that data collected at different times and under different circumstances nonetheless reinforces the generalized conclusions of the original study. Furthermore, the variation due to background noise, differing hardware, microphone placement, etc., reflects the real variation that future fully-remote studies can expect, which is in line with the goal of the study. Well-controlled, laboratory-based studies of the recording equipment and compression codecs likely to be used for remote data collection have been carried out repeatedly, as discussed in §1.1 and §1.2. The original contribution of this study is not to review how well Gorilla and smartphone recording perform under laboratory conditions, but to account for the degree of variation that can be expected under real-world remote recording conditions. Without the additional variation encountered from the sources mentioned earlier, achieving this goal would be impossible.

4.2. Consistency with previous results

Past work on this topic has offered conflicting recommendations for what style of easily-accessible home recording equipment may yield the best results for speech production studies. For instance, Zhang et al. (2021) recommended lossless smartphone recordings as the best alternative to in-person laboratory recording, while Sanker et al. (2021) obtained the best results from a laptop with an external, head-mounted microphone and Freeman and de Decker (2021a) advocated the use of a laptop, regardless of whether the microphone was internal or external. The present results align best with the recommendations from Sanker et al. (2021) and Freeman and de Decker (2021a): Even though the lossy .weba file format used on Gorilla would be expected to cause some distortions, those were fewer than those found on smartphones for the direct formant analysis, despite the smartphones using a lossless format. Why this generalization did not hold true for the Euclidean distance analysis is less clear. It is possible that the inherently relativizing nature of the EuD computation was better able to compensate for the particular distortions present in the smartphone recordings, or perhaps the lack of difference between the control group and smartphone data in EuD was accidental. After all, a certain lack of precision is inherent in EuD as a measure of reduction. While it captures the length of the reduction conceptualized as a vector, both the direction and relative position of that vector remain unanalyzed. In short, while the smartphone recordings did not prove superior in every regard, as might have been expected due to the lossless recording format, it was not unprecedented that the laptop-based Gorilla recordings should meet with a measure of success, based on the results of previous studies.

4.3. Challenges and sources of variation

Of the challenges researchers should expect to encounter when collecting speech production data remotely, acoustic distortion arising from the hardware and software used is among the more predictable elements. Variation due to microphone placement should be expected, but can be mitigated by providing clear instructions and opportunities to test and amend the recording setup before beginning the experimental task. This mitigation was successful only to a degree in the present study, as variation in clarity and SNR was clearly audible throughout the remotely recorded files. Background noise may prove more difficult to manage, as participants are not always aware of the extent of noise in their surroundings or the degree to which it is detectable by the microphone, and creating incentives to manage background noise can be logistically challenging and time-intensive for researchers. In short, researchers should expect that many participants' files will have some degree of background noise, and that some files will likely have loud and disruptive, if intermittent, noise.

Data loss due to unexplained causes should also be expected. In the current study, for instance, one participant submitted only a smartphone file and no Gorilla files (or rather, there was no sound in any of his Gorilla recordings). This was likely due to a failure to recognize that, although Gorilla was "recording", the microphone was not enabled or not connected – in short, to user error on the participant's end. It is worth noting that this error occurred despite the inclusion of a microphone check task at the beginning of the experiment that allowed participants to verify that their microphone was recording clearly.

4.4. Possible impacts of normalization

One as-yet unexamined source of variation – or, more positively, correction – within the current data rests in formant normalization. To account for interspeaker differences and especially gender-based differences, the data were normalized using log-additive regression normalization, a procedure suitable for missing and unbalanced data that aims to preserve sociolinguistic variation while reducing interspeaker variation and contains more stringent protections against overnormalization than, for example, Lobanov normalization (Barreda & Nearey, 2018). Some unique aspects of the current study may have led to unintentional impacts of the normalization procedure on the resulting data. For example, normalization did not account for differences in recording method in any way; all three datasets were normalized together, with only interspeaker differences recognized by the algorithm as overtly-labeled sources of variation. While this was necessary to ensure the resultant values were on the same scale, perhaps the lack of differentiation led to some loss of difference in the vowel categories across recording methods. Similarly, to ensure any differentiation between reduced vowels was maintained after normalization, reduced vowels were coded as members of a single vowel category with their unreduced counterparts (rather than being coded together as members of a /ə/ category). Again, while logistically necessary, it is possible that this decision created a bias of sorts in the normalization procedure, contributing to under-normalization and perhaps skewing the results. In the absence of more work on this form of normalization, or a more extensive analysis of the original unnormalized data, it is difficult to conclude what impacts, if any, the normalization procedure may have had on the conclusions of the current study. As other normalization procedures have been found capable of correcting the distortions introduced by recording in a lossy file format (see, e.g., Calder et al., 2022 on Lobanov normalization), it is not unreasonable to suspect that normalization can introduce a degree of change capable of shifting the outcome of a study. However, unless one undertakes a by-speaker analysis, some normalization is necessary when working with formant data, and log-additive regression normalization is uniquely suited to working with unbalanced data such as that found in this study.

4.5. Conclusions and directions for future research

The goal of the present study was to examine the differences in acoustic output and the conclusions of analyses based on that output across three recording conditions. Where previous work has issued clear recommendations on the suitability of remote data collection for spectral analysis of vowels concerned with minor shifts in vowel quality (not recommended) and broad properties of vowel position (likely suitable for remote data collection) (Freeman & de Decker, 2021a, 2021b), little has been done to differentiate among analyses requiring an intermediate degree of precision. In this study on vowel reduction, it became clear that the degree of divergence from data collected in a laboratory setting depended not only on the phenomenon under inspection, but also on the analytical approach. Pillai scores, a sweeping measure of overlap, were more resilient to differences of recording device and compression codec, than were Euclidean distances, a measure of relative computed distance in the vowel space, which in turn were more resilient than direct analysis of formant values. The results suggest that similar future speech production research based on remotely collected data may be considered reliable under certain circumstances. However, careful consideration of the hardware and software used, as well as the analytical parameters selected, must be made to ensure results are not misleading. Even relative measures, such as vowel reduction, are only resilient to variation up to a point, and distortions large enough to impact the conclusions of an analysis are possible.

Many important questions remain open for future research at the conclusion of this study. The degree to which normalization can correct for or exaggerate distortions introduced by differences of hardware, software, microphone placement, or background noise, for instance, deserves further investigation in the specific context of remote data collection. Further investigation of the acoustic quality of Gorilla's Audio Recording Zone is also warranted, with an additional focus on differences in its .weba and .mp3 output options. Finally, examination of other phonetic speech phenomena beyond vowel reduction is critical to expanding the generalizability of both current and past results.

Additional files

Appendices. A PDF file containing three appendices: The list of target words and carrier phrases used in the experiment, a list of microphones used by participants according to self-report, and additional statistical output not presented in the body of the work. DOI: https://doi.org/10.16995/labphon.10544.s1

Data & Code. A ZIP file containing the data files from the experiment and sample code from the statistical analysis and visualizations. DOI: https://doi.org/10.16995/labphon.10544.s2

Acknowledgements

Many thanks are due to Sophia Chuen for their assistance collecting and analyzing data and to Josh Weirick for recording the stimuli. This work was made possible through financial support from Purdue University and the Headley Fund at Carleton College.

Competing interests

The author has no competing interests to declare.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. DOI: https://doi.org/10.3758/s13428-019-01237-x

Barreda, S., & Nearey, T. M. (2018). A regression approach to vowel normalization for missing and unbalanced data. *Journal of the Acoustical Society of America*, *144*(1), 500–520. DOI: https://doi.org/10.1121/1.5047742

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Calder, J., Wheeler, R., Adams, S., Amarelo, D., Arnold-Murray, K., Bai, J., Church, M., Daniels, J., Gomez, S., Henry, J., Jia, Y., Johnson-Morris, B., Lee, K., Miller, K., Powell, D., Ramsey-Smith, C., Rayl, S., Rosenau, S., & Salvador, N. (2022). Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for vocalic analysis. *Linguistics Vanguard*, 20200148. DOI: https://doi.org/10.1515/lingvan-2020-0148

de Decker, P. (2016). An evaluation of noise on LPC-based vowel formant estimates: Implications for sociolinguistic data collection. *Linguistics Vanguard*, *2*(1). DOI: https://doi.org/10.1515/lingvan-2015-0010

de Decker, P., & Nycz, J. (2011). For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics*, *17*(2). https://repository.upenn.edu/pwpl/vol17/iss2/7

Freeman, V., & de Decker, P. (2021a). Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices. *Language and Linguistics Compass*, *15*. DOI: https://doi.org/10.1111/lnc3.12435

Freeman, V., & de Decker, P. (2021b). Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *The Journal of the Acoustical Society of America*, *149*(2), 1211–1223. DOI: https://doi.org/10.1121/10.0003529

Ge, C., Xiong, Y., & Mok, P. (2021). How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements. *Interspeech 2021*, 3984–3988. DOI: https://doi.org/10.21437/Interspeech.2021-1122

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484. DOI: https://doi.org/10.1016/j.wocn.2005.10.001

Nycz, J., & Hall-Lew, L. (2013). Best practices in measuring vowel merger. *Proceedings of Meetings on Acoustics, 20*(1), 060008. DOI: https://doi.org/10.1121/1.4894063

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. DOI: https://doi.org/10.3758/s13428-018-01193-y

R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., & Bowern, C. (2021). (Don't) try this at home! The effects of recording devices and software on phonetic analysis. *Language*, *97*(4), e360–e382. DOI: https://doi.org/10.1353/lan.2021.0075

Scarborough, R. (2005). *Supervised Formant Reading Script* [Praat Script].

Thorn, T. (2014, February 28). Background noise reduction: One of your smartphone's greatest tools. *TechRadar.* https://www.techradar.com/news/phone-and-communications/mobile-phones/background-noise-reduction-one-of-your-smartphone-s-greatest-tools-1229667 [Accessed October 18, 2023]

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). "Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Proficincy assessment issues in SLA research: Measures and practices* (pp. 143–166). Multilingual Matters. DOI: https://doi.org/10.21832/9781783092291-011

What is a WEBA file? (n.d.). https://docs.fileformat.com/audio/weba/ [Accessed January 6, 2023]

Winn, M. (2013). *Scale intensity* [Praat Script]. http://www.mattwinn.com/praat/Scale_intensity_check_maxima_v3.txt

Woods, K. J. P., Siegel, M., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics*, *79*(7), 2064–2072. DOI: https://doi.org/10.3758/s13414-017-1361-2

Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America*, *149*(6), 3910–3916. DOI: https://doi.org/10.1121/10.0005132