Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for vocalic analysis

https://doi.org/10.1515/lingvan-2020-0148
Received December 29, 2020; accepted August 25, 2021; published online February 28, 2022

Abstract: In this study, we explore whether Zoom is a viable method for collecting data for sociophonetic research, focusing on vocalic analysis. We investigate whether recordings collected through Zoom yield different acoustic measurements than recordings collected through in-person recording equipment, for the exact same speech. We analyze vowel formant data from 18 speakers who recorded Zoom conversations at the same time as they recorded themselves with portable recording equipment. We find that, overall, Zoom recordings yield lower raw F1 values and higher F2 values than recording equipment. We also tested whether normalization affects discrepancies between recording methods and found that while discrepancies still appear after normalizing with the Watt and Fabricius modified method, Lobanov normalization largely minimizes discrepancies between recording methods. Discrepancies are also mitigated with a Zoom recording setup that involves the speaker wearing headphones and recording with an external microphone.

Keywords: COVID-19; methods; sociophonetics; sound change; Zoom

1 Introduction

Fieldwork has been a part of the sociolinguistic enterprise from the field’s inception. While some sociophonetic research has analyzed data collected in labs, sociophoneticians and variationist sociolinguists have long relied on in-person, conversational data collected in the field (Labov 1984; Schilling 2013: Ch. 5). However, the effect of the COVID-19 pandemic on data collection cannot be overstated. For a subfield which has relied on the ability to collect data in person through the traditional sociolinguistic interview (Labov 1972, 1984; Tagliamonte 2006: Ch. 3), with a traditional recording setup involving in-person equipment (DeDecker and Nycz 2013; Hall-Lew and Plichta 2013; Meyerhoff et al. 2015; Schilling 2013: Ch. 6), the advent of a pandemic which restricts our ability to be within six feet of other people also restricts our ability to collect the in-person data we are accustomed to. With this unexpected shift, sociophoneticians and variationist sociolinguists have been spurred to explore other data collection methods, primarily ones which do not necessitate in-person interaction. Zoom, an online videoconferencing software, has become the primary means of social interaction in a post-COVID world for many of us. Thus, some researchers have started to consider Zoom as a data
collection source for sound change research (e.g., Freeman and DeDecker 2021). Of course, as sociolinguists, we want to be sure that the patterns being gleaned from the sounds we are analyzing are a result of social factors – e.g., the age, gender, region, or ethnicity of a speaker – or language-internal factors – e.g., the neighboring phonological environment or duration of the sound – rather than artifacts that arise based on technical differences between data collection methods. And while recent important work has made great strides in probing the effects different data collection methods have on the sounds variationists analyze (e.g., DeDecker and Nycz 2011; Hall-Lew and Boyd 2017, 2020; Leemann et al. 2016, 2020), and one study has explored Skype as a possible sociophonetic data source (Bulgin et al. 2010), the ubiquity of Zoom is a relatively recent phenomenon, and the effects of Zoom as a data collection tool remain ripe for exploration. Thus far, one important study has explored differences between sound recordings collected via Zoom and via portable recording equipment, among two speakers recorded in a sound booth (Freeman and DeDecker 2021). However, it is still unknown whether differences between recording methods are exhibited across a larger speaker sample, and when Zoom data is collected outside of a sound booth, in a conversational format more akin to a sociolinguistic interview.

In this study, we explore whether Zoom is a viable data source for sociophonetic research by aiming to address the primary research question: Do recordings collected through Zoom yield different acoustic measurements than recordings collected through a more traditional in-person approach, for the exact same speech, when other factors are controlled for? Much sociophonetic work has centered on explorations of the social effects that condition relative vowel positions and arrangements, including sound change research (e.g., the California Vowel Shift, Hagiwara 1997; Podesva et al. 2015; the Northern Cities Shift, Eckert 1989; Labov et al. 2006; D’Onofrio and Benheim 2020; the Canadian Vowel Shift, Clarke et al. 1995; Boberg 2005; the Southern Vowel Shift, Thomas 2004; Labov et al. 2006). Here, we explore differences between Zoom and in-person recordings, with respect to vowel formant measurements correlating to height (formant 1) and frontness (formant 2) of articulation, to probe whether Zoom recordings are comparable to in-person recordings for sociophonetic research on vowels. We expand on Freeman and DeDecker’s (2021) study by exploring the effects of Zoom recordings on a larger speaker sample (18 speakers), using conversational sociolinguistic interview data. We analyze Zoom and in-person recordings for the same speech for each speaker and explore whether different recording setups and normalization methods can be used to mitigate discrepancies between recordings.

2 Methods

The data come from conversations in English between 18 graduate students enrolled in a sociophonetic analysis seminar. The participants include six males, 10 females, and one nonbinary participant from a range of geographic backgrounds around the world, all aged in their 20s and 30s. As we are investigating the intra-speaker variable of the recording method, rather than inter-speaker phonetic variation, the demographic and geographic makeup of the sample is incidental. Each conversation (nine total) included two participants conversing on Zoom. The conversations mimicked the structure of a sociolinguistic interview in the field, including a conversational question and answer portion and a word-list reading task (Labov 1972); however we focus on the conversational portion.

Each participant recorded their conversation through Zoom at the same time as they recorded their speech with portable audio recorders, yielding two recordings of the exact same speech for each speaker. We were not strict on controlling for the hardware used to record via Zoom, as it is unlikely that a sociolinguist in the field would be able to control for the specific hardware their subjects have available to them during Zoom interviews. Instead, we code for different Zoom setups in the statistical analysis, including factors such as whether the participant used headphones during the Zoom conversation, whether the participant’s Zoom recording was collected from the built-in computer microphone or from an external microphone, and whether the speaker was the host of the Zoom conversation (i.e., whether the speaker’s Zoom audio was recorded through their own computer equipment as the host, or second-hand via the Zoom video on the other speaker’s
Zoom recordings were automatically recorded in M4A format at 32.0 kHz (the software does not currently offer the ability to record at 44.1 kHz at the time of writing), but these files were later converted to WAV in Audacity (2020) at a 16-bit bit rate to enable compatibility with software used in measuring and processing the data. One limitation to acknowledge is that Zoom conversations were recorded using default settings, and the software includes automatic noise-reduction capabilities (though at the time of writing, the software includes settings that allow the user to turn off automatic adjustment of microphone volume, and to adjust how much background noise is filtered with multiple levels). However, each speaker recorded their conversations in a quiet environment in their own home, and no speaker reported any significant noise disturbances during the conversation.

The in-person recordings on the portable audio recorders were recorded at a sampling rate of 44.1 kHz and a bit depth of 16-bit, as is standard for sociophonetic analysis (DeDecker and Nycz 2013). Lapel microphones were clipped on each speaker’s chest about 6 inches from their mouth. We chose the SLINT omnidirectional condenser lavalier microphone due to the ubiquity of lavalier condenser microphones in sociophonetic analysis (DeDecker and Nycz 2013) and due to its affordability. Fifteen of the 18 participants used Olympus 822/823/852/853 model portable audio recorders. We selected this series of recorders because of their affordability and their capability of recording at a 44.1 kHz sampling rate and 16-bit bit depth. Recorders of this type have been used in previous studies of sociophonetic variation (e.g., Calder 2019a, 2019b; Van Hofwegen 2016, 2017). However, three speakers used different recorders, as the Olympus recorders were not easily available in their geographic location at the time of the recordings: the TASCAM DR-100MKII, the Voicetracer DVT 2050, and the Philips VTR8060. All of these recorders also had the ability to record at a 44.1 kHz sampling rate and 16-bit bit depth. Recorders capable of recording in WAV format were recorded at a sampling frequency of 128 Kbps and 44.1 kHz sampling rate; these files were then converted to WAV format in Audacity at a sampling rate of 44.1 kHz and bit depth of 16-bit. We initially coded for the type of recorder used to control for recording differences and possible differences between recordings originally collected in WAV format and those converted from MP3, but this never emerged as a statistically significant factor, so we did not include this information in our final statistical models. One possible explanation for the lack of significant differences between portable recorders is that each speaker used exactly the same microphone, regardless of the recorder they used. In addition, previous work also found no significant difference between WAV recordings and those converted from MP3 for vocalic analysis (Bulgin et al. 2010). Finally, all stereo recordings were converted to mono to enable compatibility with Praat scripts used to collect measurements.

Given that the Zoom and in-person recordings were of the same speech, we transcribed 30 min of each speaker’s in-person recording in ELAN (2020), duplicated this transcription file, and adjusted the boundaries of the intervals in the duplicated transcription so that the intervals aligned with the audio file for the Zoom recording. We generated forced-aligned Praat (Boersma and Weenink 2021) TextGrids for each transcribed audio file using FAVE (Rosenfelder et al. 2015).

Praat scripts were used to extract vowel tokens (with primary stress) from the following 10 Wells lexical sets representing vowel phonemes in English: FLEECE, KIT, DRESS, TRAP, LOT, THOUGHT, STRUT, GOAT, GOOSE, and POOL. These vowel classes were chosen to include a representation of vowels at different locations in the vowel space, and to enable Lobanov normalization. We excluded tokens from analysis that were under 60 ms long and that were adjacent to other vowels, glides, rhotics, or liquids (excepting the POOL class, which was necessarily pre-liquid). All extracted tokens were manually inspected in Praat to ensure the tokens were of the correct sound, and TextGrid interval boundaries were hand-corrected from their automatic alignments based on visual spectrogram cues. We excluded tokens if there was any audible overlap with another speaker or if the automatic alignment and extraction yielded the incorrect sound. We analyze exactly the same vowel tokens across recording methods, accepting up to 25 tokens per vowel class, per recording method, per speaker (though many vowel classes yielded far fewer tokens). For each token, we then used Praat scripts to measure formants 1 and 2 at the vowel midpoint, in order to test whether recording method affects

---

1 The Olympus WS 822/823, TASCAM DR-100MKII, Voicetracer DVT 2050, and Philips VTR8060 recorded in WAV format, while the Olympus WS 852/853 recorded in MP3 format.
measurements that correlate to vowel height (F1) and frontness (F2). Scripts were set to extract five formants for each vowel, with a hertz ceiling of 5,000 Hz for men and 5,500 Hz for women,² following previous studies (e.g., D’Onofrio et al. 2016; Podesva et al. 2015). The scripts also provided duration measurements for each hand-corrected token, and all durations were log-transformed. Finally, since sociophonetic studies of vowels often rely on normalized values to make cross-speaker comparisons, all formants were normalized using two normalization methods that are relatively common in variationist study: the Lobanov (1971) method, and the Watt and Fabricius modified method (henceforth WFM; see Fabricius et al. 2009). These two methods have been shown to perform comparatively better than other methods at reducing individual anatomical differences while maintaining differences in articulation resulting from social factors (see Adank et al. 2004; Fabricius et al. 2009). Both methods are vowel-extrinsic and speaker-intrinsic; however, while Lobanov normalizes using the grand mean of the vowel space, WFM normalizes based on points that represent the corners of the vowel space. Vowels were normalized using the NORM web-based interface (Thomas and Kendall 2007), and Zoom data was normalized separately from in-person data.

Statistical analysis was performed with mixed effects linear regressions in RStudio (2020), using the lme4 package (Bates et al. 2015). First, the entire data set was analyzed in two models: one for F1 and one for F2, with vowel class as a fixed effect, following Hall-Lew and Boyd (2020). Other fixed effects we included were speaker gender³ (as gender has been shown to pattern with formant measurements in predictable ways; e.g., Hillenbrand et al. 1995) and importantly, whether the data came from the Zoom recording or from the in-person recording. Random effects included speaker, word, preceding environment, following environment, and the geographic region the speaker grew up in. Also, since previous studies have shown that recordings collected via Zoom (Freeman and DeDecker 2021), Skype (Bulgin et al. 2010), and other online methods (DeDecker and Nycc 2011) affected measurements for male and female speakers to different degrees, interactions between gender and recording method were also explored in our statistical models.

Finally, the F1 and F2 data were analyzed in separate models for each vowel class, with the above effects tested in each model. In addition, we tested a number of additional fixed effects. Logged duration was included in the models to test whether vowel formant data patterned in predictable ways, that is, whether longer tokens were less centralized in the vowel space. We also included factors related to the speaker’s Zoom setup. One factor was whether the speaker was the host of the Zoom conversation: that is, whether the speaker’s Zoom data was recorded on their own computer as the host of the conversation or whether the speaker’s data was recorded through another person’s computer. We also included recording setup as a factor, with speakers binned into three groups: whether the speaker used headphones and an external microphone (e.g., earbuds or a headset) while conversing on Zoom (number of speakers = 13); whether the speaker used headphones and the computer-internal microphone on Zoom (n = 3); or whether the speaker recorded with a computer-internal microphone without headphones on Zoom (n = 2). We did not have any speakers record using an external microphone without headphones. We only included these effects in the separate vowel-class models, as they were shown to affect the data differently depending on vowel class. We also tested interactions between effects, but only included these in final models where significant. All of the above models were first run with F1 and F2 in hertz as dependent variables – in order to test whether differences in measurements arise between Zoom and in-person equipment at the recording stage – and then re-run with Lobanov-normalized and WFM-normalized F1 and F2 as dependent variables – in order to test whether discrepancies remain after normalization.

3 Results

The initial models for raw F1 and F2 over the entire data set reveal significant differences in formant measurements between the Zoom and in-person recordings, such that raw F1 is overall lower in the Zoom recording

² The nonbinary speaker’s measurements were taken with a hertz ceiling of 5,500 Hz. For more on the methodological limitations of collecting vowel measurements for nonbinary speakers, see Miles-Hercules and Zimman (2019).
³ For this study, the gender factor was coded in three categories: male, female, and nonbinary.
(estimate = −37.029***), and raw F2 is overall higher in the Zoom recording (estimate = 19.115***), as shown in Figure 1. In other words, the gap between F1 and F2 is wider for Zoom recordings. Gender is a significant predictor as well, with men overall exhibiting lower F1 (estimate = −98.482**) and F2 (estimate = −237.774***)
values than other speakers in the sample. An interaction between gender and recording method also emerges for F2, such that the difference between Zoom and in-person recordings is stronger for men than for women and nonbinary speakers in general. This pattern can be seen in Figure 3 and seems to manifest primarily in vowels higher in the vowel space. Finally, vowels as fixed effects pattern as expected, with higher vowels exhibiting lower F1 than lower vowels, and fronter vowels exhibiting higher F2 than backer vowels.5

Results for Lobanov-normalized F1 and F2 show that recording method no longer significantly predicts formant measurements when vowels are normalized, which is reflected in the left plot in Figure 2. In addition,

4 Significance stars signify the following: *p < 0.05; **p < 0.01; ***p < 0.001. A dot (.) signifies p < 0.10, suggesting the factor approaches significance but not does reach the significance threshold of p < 0.05.
5 See Tables 1 and 2 in Appendix for details on all fixed effects and interactions for the linear mixed effects models with raw F1 and raw F2 as dependent variables.
gender only emerges as significant for normalized F1, such that men, in general, exhibit higher values than other speakers (estimate = 0.1233*). One possible explanation is that men exhibit lower articulations of vowels than other speakers in the sample, when differences in vocal tract size are accounted for. Another explanation is that men exhibit smaller vowel spaces than women in many English dialects, even after normalization (see Heffernan 2010 for a discussion). This gender effect holds regardless of recording method. Vowels as fixed effects again pattern as expected based on place of articulation.

On the other hand, WFM-normalized F1 and F2 exhibit significant differences between Zoom and in-person recordings, such that Zoom exhibits higher values than in-person recordings for both F1 (estimate = 0.03168***)

---

6 See Tables 3 and 4 in Appendix for details on all fixed effects and interactions for the linear mixed effects models with normalized F1 and F2 (Lobanov method) as dependent variables.
and F2 (estimate = 0.1941***). This suggests that, while Zoom maximizes the difference between F1 and F2 for non-normalized measurements, Zoom instead shifts both F1 and F2 up in spectral space for WFM-normalized values. In addition, a significant interaction between gender and recording method emerges, such that differences between recording methods appear larger for women and the nonbinary speaker than for men. This effect seems to be more pronounced lower in the vowel space, as shown in the right-hand plots in Figure 4. Vowels as fixed effects retain patterns that suggest place of articulation effects.7

Overall, considering the entire data set together, the recording method appears to significantly affect raw hertz measurements, and some discrepancies between recording methods remain for WFM-normalized values, but Lobanov normalization largely seems to mitigate discrepancies.

We now consider the models for each individual vowel class, with significant effects for raw formants presented in Table 1, Lobanov-normalized values in Table 2, and WFM-normalized values in Table 3.8 Overall, the patterns within the individual vowel classes exhibit the larger patterns of the entire data set. As shown in Table 1, recording method is a significant predictor of raw F1 for most vowel classes, with F1 being lower in the Zoom recording than the in-person recording. There is a suggestive relationship between recording method and F1 for POOL and DRESS, but this does not reach statistical significance. The only vowel for which there is neither a significant or suggestive relationship between recording method and F1 is TRAP, which is the lowest vowel in the vowel space, that is, the vowel with the highest F1 value. In addition, F2 is significantly higher in the Zoom recording than in the in-person recording for DRESS, GOOSE, and LOT, and for many other vowel classes, F2 is significantly higher in the Zoom recording for speakers with a sub-optimal recording setup (i.e., especially for speakers recording without headphones and speakers recording with computer-internal microphones). Observing the patterns in Table 1 and the Figure 1, we see that F1 differences are greater for vowels higher in the vowel space, and F2 differences are often greater for vowels farther front in the vowel

---

7 See Tables 5 and 6 in Appendix for details on all fixed effects and interactions for the linear mixed effects models with normalized F1 and F2 (Watt and Fabricius modified method) as dependent variables.
8 For individual regression tables, see Tables 7–66 in Appendix.
Table 1: Significant model coefficients for linear mixed effects models for each vowel class, with raw formant frequency (in hertz) as dependent variable (for all tables, the conditions “Headphones and int mic” and “No headphones and int mic” diverge from the reference level “Headphones and external mic”).

<table>
<thead>
<tr>
<th>Vowel class</th>
<th>Formant</th>
<th>Recording (Zoom)</th>
<th>Gender (M)</th>
<th>Logged duration</th>
<th>Headphones and int mic × Recording (Zoom)</th>
<th>No headphones and int mic × Recording (Zoom)</th>
<th>Host (Y) × Recording (Zoom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRESS (n = 582)</td>
<td>F1</td>
<td>-16.61 (.)</td>
<td>-149.8 (***)</td>
<td>61.54 (***)</td>
<td>-28.48 (*)</td>
<td>-30.7 (**)</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>20.79 (*)</td>
<td>-199.8 (**)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLEECE (n = 520)</td>
<td>F1</td>
<td>-47.96 (***)</td>
<td>-71.67 (.)</td>
<td>109.9 (.)</td>
<td>129.81 (***)</td>
<td>-52.18 (***)</td>
<td>-36.96 (*)</td>
</tr>
<tr>
<td>F2</td>
<td>-52.26 (***</td>
<td>-31.46 (.)</td>
<td></td>
<td></td>
<td></td>
<td>538.79 (***</td>
<td></td>
</tr>
<tr>
<td>GOAT (n = 540)</td>
<td>F1</td>
<td>-38.59 (***)</td>
<td>-103.36 (**)</td>
<td>53.3 (.)</td>
<td>93.97 (***)</td>
<td>76.75 (.)</td>
<td>-32.71 (**)</td>
</tr>
<tr>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-57.26 (.)</td>
<td></td>
</tr>
<tr>
<td>GOOSE (n = 246)</td>
<td>F1</td>
<td>-29.17 (***)</td>
<td>-49.8 (.)</td>
<td>43.91 (.)</td>
<td></td>
<td>-75.07 (***)</td>
<td>-82.07 (***</td>
</tr>
<tr>
<td>F2</td>
<td>58.68 (.)</td>
<td>-331.44 (.)</td>
<td>-88.35 (.)</td>
<td>93.49 (***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIT (n = 388)</td>
<td>F1</td>
<td>-47.96 (***)</td>
<td>-70.95 (.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>-260.99 (.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOT (n = 380)</td>
<td>F1</td>
<td>-15.32 (.)</td>
<td>-132.97 (**)</td>
<td>43.91 (.)</td>
<td></td>
<td>93.49 (***</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>19.56 (**)</td>
<td>-156.86 (.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POOL (n = 82)</td>
<td>F1</td>
<td>-21.92 (.)</td>
<td>-67.19 (.)</td>
<td></td>
<td>-67.13 (.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>-193.24 (.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRUT (n = 548)</td>
<td>F1</td>
<td>-37.61 (***)</td>
<td>-103.87 (**)</td>
<td>27.81 (.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>-192.5 (**)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THOUGHT (n = 224)</td>
<td>F1</td>
<td>-23.4 (.)</td>
<td>-154.1 (**)</td>
<td></td>
<td>-47.65 (.)</td>
<td>57.68 (.)</td>
<td>-44.84 (.)</td>
</tr>
<tr>
<td>F2</td>
<td>-238 (***</td>
<td>-59.89 (**)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAP (n = 586)</td>
<td>F1</td>
<td>-159.87 (**)</td>
<td>-32.56 (.)</td>
<td>31 (**)</td>
<td>-30.41 (.)</td>
<td>60.69 (***</td>
<td>65.24 (***</td>
</tr>
</tbody>
</table>
Table 2: Significant model coefficients for linear mixed effects models for each vowel class, with Lobanov-normalized formant as dependent variable.

<table>
<thead>
<tr>
<th>Vowel class</th>
<th>Formant</th>
<th>Recording (Zoom)</th>
<th>Gender (M)</th>
<th>Gender (M) × Recording (Zoom)</th>
<th>Logged duration</th>
<th>Headphones and int mic × Recording (Zoom)</th>
<th>No headphones and int mic × Recording (Zoom)</th>
<th>Host (Y) × Recording (Zoom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRESS</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td>0.368 (***</td>
<td></td>
<td></td>
<td>-0.227 (***</td>
</tr>
<tr>
<td>(n = 582)</td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLEECE</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 520)</td>
<td>F2</td>
<td>-0.582 (*)</td>
<td>0.229 (*)</td>
<td></td>
<td>0.279 (***</td>
<td>0.26 (**)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOAT (n = 540)</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>0.166 (.)</td>
<td>-0.224 (***)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOOSE</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.366 (**)</td>
</tr>
<tr>
<td>(n = 246)</td>
<td>F2</td>
<td>-0.285 (*)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIT (n = 388)</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td>0.25 (**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>-0.049 (.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOT (n = 380)</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.234 (***</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POOL (n = 82)</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td>0.264 (.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRUT</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td>0.167 (**)</td>
<td></td>
<td></td>
<td>-0.096 (.)</td>
</tr>
<tr>
<td>(n = 548)</td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THOUGHT</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.309 (</td>
</tr>
<tr>
<td>(n = 224)</td>
<td>F2</td>
<td>-0.138 (**)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.172 (*)</td>
</tr>
<tr>
<td>TRAP (n = 586)</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td>0.185 (**</td>
<td></td>
<td></td>
<td>-0.214 (</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>-0.099 (**)</td>
<td>0.091 (**)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Is Zoom viable for vocalic analysis?
Table 3: Significant model coefficients for linear mixed effects models for each vowel class, with WFM-normalized formant as dependent variable.

<table>
<thead>
<tr>
<th>Vowel class</th>
<th>Formant</th>
<th>Recording</th>
<th>Gender (M)</th>
<th>Gender (M) × Recording</th>
<th>Log ed duration</th>
<th>Headphones and int mic × Recording</th>
<th>No headphones and int mic × Recording</th>
<th>Host (Y) × Recording</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRESS (n = 582)</td>
<td>F1</td>
<td>-0.066 (**)</td>
<td>0.119 (***)</td>
<td>0.056 (*)</td>
<td>0.061 (*)</td>
<td>-0.099 (***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td>0.018 (*)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLEECE (n = 520)</td>
<td>F1</td>
<td>-0.054 (***)</td>
<td></td>
<td></td>
<td>0.084 (***)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td>0.065 (*)</td>
<td></td>
<td></td>
<td>0.287 (***)</td>
<td></td>
</tr>
<tr>
<td>GOAT (n = 540)</td>
<td>F1</td>
<td>-0.054 (***)</td>
<td></td>
<td></td>
<td>0.138 (***)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td>-0.064 (***)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOOSE (n = 246)</td>
<td>F1</td>
<td></td>
<td></td>
<td>0.036 (.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td>-0.074 (.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIT (n = 388)</td>
<td>F1</td>
<td></td>
<td></td>
<td>0.076 (***)</td>
<td>0.068 (**)</td>
<td>0.049 (.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOT (n = 380)</td>
<td>F1</td>
<td>0.052 (**)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.071 (.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.053 (***)</td>
<td>-0.175 (***)</td>
<td></td>
</tr>
<tr>
<td>POOL (n = 82)</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRUT (n = 548)</td>
<td>F1</td>
<td>0.039 (*)</td>
<td></td>
<td></td>
<td>0.062 (*)</td>
<td>0.151 (***)</td>
<td>-0.086 (***)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THOUGHT (n = 224)</td>
<td>F1</td>
<td></td>
<td></td>
<td></td>
<td>0.071 (***)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAP (n = 586)</td>
<td>F1</td>
<td>0.07 (***)</td>
<td></td>
<td></td>
<td>0.062 (**)</td>
<td>0.083 (**)</td>
<td>0.094 (**)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAP (n = 586)</td>
<td>F2</td>
<td>-0.021 (***)</td>
<td></td>
<td></td>
<td>0.036 (***)</td>
<td></td>
<td>0.114 (***)</td>
<td></td>
</tr>
</tbody>
</table>
space. In other words, the recording method seems to have stronger effects on F1 the lower the F1 value is and F2 the higher the F2 value is, suggesting a stretching of the spectral space on Zoom.

Beyond recording method, gender is shown to often significantly predict raw formant values, with men exhibiting lower values than other speakers. Finally, logged duration emerged as significant for multiple vowel classes, such that longer tokens are less centralized in the vowel space. A number of significant interactions between factors also emerged as significant. Interactions between gender and recording method reveal that discrepancies between Zoom and in-person recordings were larger for men for FLEECE F2, GOAT F2, and TRAP F1, while discrepancies were larger for women for GOAT F1, as suggested by Figure 3. In addition, a number of interactions reveal that discrepancies are often larger for speakers recording without headphones and with computer-internal microphones than for speakers wearing headphones and recording with external microphones, as can be seen in Figure 5. Finally, as shown in Figure 7, hosts of the Zoom conversation exhibited larger discrepancies between recording methods than speakers who were not hosts of the conversation. This is an unexpected finding, as the hosts’ data is recorded directly onto their own computers, while the data from speakers who were not hosts was recorded secondhand on the conversation host’s computer. However, the speakers recording without headphones often exhibited the largest discrepancies between recording methods, as shown in Figures 5 and 6, and both of these speakers were also the hosts of their Zoom conversations, so it is possible that the speakers without headphones are driving the larger discrepancies among speakers who hosted the Zoom conversations than among those who were not hosts.

Table 2 shows that recording method does not emerge as a significant for Lobanov-normalized formants, and nearly all formant values are statistically comparable across recording methods. There is a trend for KIT F2 to be lower in the Zoom recording, but this trend does not reach statistical significance. In addition, gender significantly predicts FLEECE F2, such that men exhibit backer normalized FLEECE than women and nonbinary speakers, again suggesting a more compressed vowel space for men, once formants are normalized. Duration remains significant for many Lobanov-normalized values, suggesting that after normalization, longer tokens are still less centralized in the vowel space. A number of interactions also emerge as significant. The interaction between gender and recording method significantly predicts Lobanov-normalized FLEECE F2, GOAT F1, and TRAP F2, such that the difference between recording methods is larger for men. Recording setup interactions emerge for a number of Lobanov-normalized formants; in general, discrepancies between recording methods are larger for those recording on Zoom without headphones and with computer-internal microphones, than for those recording on Zoom with headphones and external microphones, as illustrated in the left-hand plots in Figure 6. Finally, as shown in the left-hand plots in Figure 8, some of the formant discrepancies are slightly larger for those who were hosts of the Zoom conversation than for those who were not hosts. As for the non-normalized values, the larger discrepancies among Zoom hosts are likely largely driven by the two speakers who recorded on Zoom without headphones, and who were both the hosts of their respective Zoom calls.

Table 3 shows the significant effects predicting WFM-normalized F1 and F2 for each vowel class. Overall, recording method emerges as a significant predictor of fewer WFM-normalized formants than for raw formants, suggesting that WFM normalization may ameliorate some of the discrepancies between Zoom and in-person recordings. However, FLEECE F1, GOAT F1, and KIT F1 are lower for Zoom than for the recorder, LOT F1, STRUT F1, and TRAP F1 are higher for Zoom than for the recorder, and STRUT F2 and TRAP F2 are lower for Zoom than for the recorder. Duration effects are retained from non-normalized values, such that longer tokens are less centralized. Significant interactions between gender and recording method emerge, but in non-linear ways. Recording setup significantly interacts with recording method, such that speakers recording on Zoom without headphones and with computer-internal microphones exhibit larger discrepancies between Zoom and in-person measurements than do speakers recording on Zoom with headphones and an external microphone, as illustrated in the right-hand plots in Figure 6 (though it must be acknowledged that the sample is unbalanced: only three speakers recorded with headphones and internal microphones, and only two recorded without headphones). Finally, hosts of the Zoom call exhibit lower values on Zoom than in-person for a number of WFM-normalized values than do non-hosts.
Figure 5: Mean formant values in raw hertz for speakers who recorded using headphones and external microphones (top, n = 13), headphones and computer-internal microphones (middle, n = 3), and computer-internal microphones without headphones (bottom, n = 2) during the Zoom call.
4 Conclusions

Overall, for raw formants, Zoom recordings appear to stretch the spectral space compared to the in-person recording, such that F1 is lower in Zoom recordings and F2 is higher for Zoom recordings.\(^9\) The effects are greater for F1 the lower the F1 of the vowel class, and greater for F2 the higher the F2 of the vowel class. It appears that around the 1,200 Hz range is where the two recording methods are most similar, and the farther away the measurements are from this frequency, the larger the discrepancies become. In a parallel study where we explore the effect of recording method on sibilants, which occupy a spectral range much higher than vowel formants, we find the discrepancies between Zoom and in-person measurements are even wider for sibilants, further illuminating the stretching of the spectrum on Zoom recordings (Calder and Wheeler in press).

\(^9\) F2 is statistically significantly higher in the Zoom recordings specifically for the vowels DRESS, GOOSE, and LOT, though many other vowels exhibit similar F2 patterns in particular conditions in interactions with other factors.
One possible explanation for this stretching could be the difference in sampling rates between the two recording methods. While the in-person recordings were collected at a sampling rate of 44.1 kHz, Zoom recordings are automatically done at a sampling rate of 32 kHz (with no option for a higher sampling rate at the time of this writing). Acoustically, the maximum frequency that can be represented in a recording is half of the sampling rate (i.e., the Nyquist frequency; Audacity 2020): for 44.1 kHz, this is about 22,050 Hz, and for 32 kHz, it is about 16,000 Hz. It is possible that a greater higher frequency range for the in-person recordings collected at 44.1 kHz influences where formant resonances are acoustically concentrated, when compared to the Zoom recordings at a lower sampling rate. Other work has also found sampling rates to affect formant frequencies, with lower sampling rates showing greater differences between formant values than higher sampling rates, suggesting a stretching of the vowel space (Wagner et al. 2017). More work is needed to probe whether Zoom and in-person recordings differ when sampling rates are controlled for; one possibility would be to compare in-person recordings recorded at 32 kHz with Zoom recordings to explore whether differences between recording

Figure 7: Mean formant values in raw hertz for speakers who hosted (top, n = 9) and did not host (bottom, n = 9) the Zoom call.
methods still exist. Interestingly, despite the spectral stretching exhibited by Zoom, the degree of difference between Zoom and in-person recordings does not appear to be the same magnitude as differences found between Skype and in-person recordings a decade ago (Bulgin et al. 2010).

Lobanov normalization appears to mitigate the effects of the spectral stretching of the Zoom recording. For speakers recording with setups involving headphones and internal microphones, no significant differences between recording methods were observed for Lobanov-normalized values. However, speakers recorded without headphones or with computer-internal microphones did exhibit discrepancies for some vowel classes. On the other hand, it seems that WFM normalization does not retain the discrepancies between recording methods exhibited by the raw formants, but it introduces new discrepancies that do not always pattern in linear and predictable ways. Also, differences between Zoom and in-person recordings seemed to be greater for men than for women in some cases, especially for the raw and Lobanov-normalized formants, as shown in Figure 4. Despite these patterns for raw and Lobanov-normalized values, the interaction between gender and recording method patterned in a less clear way for WFM-normalized formants.

Furthermore, a recording setup involving the use of headphones and an external microphone on the Zoom call seems to minimize many of the discrepancies between recording methods that are exhibited to greater degrees among speakers who recorded on Zoom using less optimal setups, as shown in Figure 5. It is possible that a speaker recording on Zoom with an external microphone, like those on earbuds or a gaming headset, more closely mimics the traditional setup with a lapel mic, with the microphone inches away from the speaker’s mouth, as opposed to feet away, when the microphone is built into the computer. In addition, a speaker recording on Zoom without headphones may be more likely to record noise coming from the Zoom conversation, which could affect the speaker’s own formant values. Previous research has shown that recordings collected with the microphone farther away from the speaker affects the sound-to-noise ratio (SNR) of a recording (e.g., Svec and Granqvist 2010; Titze and Winholtz 1993). For example, a lower SNR reading resulting from a greater degree of ambient white noise can result in higher F2 measurements, and a lower SNR
reading resulting from overlapping speech can result in lower F1 measurements (DeDecker 2016). One possibility is that speakers recording with computer-internal microphones and without headphones produce Zoom recordings with lower SNR readings, resulting in the more dramatic stretching of the formant space in the Zoom recording. An exploration of the relationship between SNR readings and discrepancies between recording methods is one avenue for future research.

Overall, the patterns suggest that if a researcher wants to collect data from Zoom that is most comparable to in-person recordings, a recording setup where the subject wears headphones and records using an external mic (even if just the microphone on a pair of earbuds or a gaming headset) comes closest to yielding data comparable to an in-person setup.

Interestingly, in many cases Zoom hosts exhibited larger differences between Zoom and in-person recordings than did those who did not host the Zoom calls, as shown in Figure 7. This pattern is perhaps driven by the two speakers who recorded on Zoom without using headphones, as both were hosts of their respective Zoom calls. One avenue for future work is to probe the effects of headphone use on Zoom recordings in a more controlled way with a balanced sample.

Finally, our results diverge in a number of notable ways from the results found in Freeman and DeDecker’s (2021) comparison of Zoom and in-person recordings. While we found a stretching of the vowel space on Zoom, Freeman and DeDecker seemed to find a squishing in certain regions of the vowel space. Specifically, we found Zoom to lower F1, which raised most vowels (especially high vowels), stretching the vowel space upward. In contrast, Freeman and DeDecker found that Zoom primarily raised low vowels rather than high vowels, compressing the lower vowel space upward. In addition, while we found that men exhibited greater differences between recording methods for some formants than women did, Freeman and DeDecker found greater differences for women. Our results may differ from previous findings due to a difference in recording methods: while the Zoom data was recorded directly into an iPad Air in a sound-attenuated booth for Freeman and DeDecker’s study, our Zoom data was recorded outside of a sound-attenuated booth with a variety of setups. In addition, our in-person recordings involved a lavalier microphone, while speakers in the aforementioned study recorded directly into a H4n Pro recorder, placed 30–40 cm from the speaker. Freeman and DeDecker suggest that recording hardware may be a more important consideration than the recording application, and the difference in hardware used between the two studies is a likely source of the difference in patterns differentiating recording methods. Further work is necessary to probe how different recording setups influence differences between Zoom and in-person recordings.

Before concluding, we want to acknowledge a few limitations of this study. For one, some of our statistical findings should be taken with caution, due to small sample sizes for some of the vowel classes (especially POOL). For some recording factors, our sample was unbalanced as well. For example, future work may explore the effect of various aspects of the recording setup (e.g., headphones, microphones) with equal numbers of speakers in each condition. Also, while this was primarily an intra-speaker study, we want to acknowledge the paired nature of our data, such that we had two recordings of each token for each speaker. We also recorded our Zoom conversations with default settings, meaning that the automatic noise-reduction filter built into the software was enabled. While all speakers recorded in quiet environments, future studies would be more controlled without this automatic noise-reduction feature enabled. In addition, our microphones were not studio-grade, and some of our portable recorders produced files that had to be converted from MP3 to WAV format, although this did not seem to significantly affect formant measurements in preliminary statistical models. Finally, while we had half of the participants record the Zoom conversations in order to explore whether formant measurements were affected by whether the recording was collected on one’s own computer, or secondhand through another computer, a better method may have been to have everyone record the Zoom conversations on their own computer, in order to compare the exact same speech on their own recording versus the recording collected from their conversation partner’s computer.

We now return to the point that, when investigating the social factors conditioning sociophonetic variation, we want to ensure that we are filtering out as many extraneous effects as possible, in order to know whether the effects we observe in our data are really social and phonological, as opposed to artifacts introduced through technical limitations. To this effect, we should exercise caution in comparing raw formant
values across recording methods, as these seem to be quite significantly affected by the means of data collection. For one, when analyzing raw Zoom data, we may be tempted to assume that a speaker’s vowels are higher and fronter in the vowel space than they actually are, leading us to possibly make incorrect assumptions about a speaker’s participation in regional vowel shifts. Similar misinterpretations may arise when analyzing WFM-normalized Zoom data.

However, Lobanov normalization, along with a Zoom recording setup involving headphones and an external microphone, seems to yield values that are closest to what we would achieve with traditional in-person recordings. While the two recording methods did not align perfectly for every vowel class, even for Lobanov-normalized values, for no formant did differences between recording methods reach statistical significance. If we want to be sure that the significant differences we observe in data collected through Zoom are indicative of social influences, analyzing Lobanov-normalized data, recorded with minimal audio interferences, appears to be the best approach. Restricting comparisons between speakers to those recorded within the same recording method (e.g., Zoom) is another alternative. Collecting sociophonetic data via Zoom presents many benefits: it does not require unnecessary travel, it arguably costs less, and it allows participants to record in locations of their choosing. Even when face-to-face research once again becomes a safe possibility, Lobanov-normalized Zoom data may be a beneficial alternative to in-person recordings for sociophonetic analysis, as long as researchers acknowledge the medium’s imperfections.

References


Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Seyfarth Scott, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 2015. *FAVE (forced alignment and vowel extraction)*, version 1.1.3. ZENODO.


---

**Supplementary Material**: The online version of this article offers supplementary material (https://doi.org/10.1515/lingvan-2020-0148).