Robert Marcelo Sevilla*

Yiyang Xiang Vowel Quality: Comparability Across Two Recording Mediums

Abstract: Current events have necessitated the sacrifice of some degree of recording quality in order to reach inaccessible or far-away areas; for instance, using video conferencing software like Zoom (Zoom Video Communications) for recording over traditional in-person microphone or sound booth recording. This then leads to the question: can Zoom-recorded data be used more-or-less interchangeably with standard recording procedures? The present research is an analysis of vowel acoustics in the Yiyang dialect of Xiang (Sinitic), comparing across two recording mediums: one online (Zoom) and another in person (Sound Booth). Research of Xiang varieties has been made increasingly difficult during the pandemic. This study analyzes two recordings retelling the events of the 'Pear Stories' video, performed by a speaker of Yiyang Xiang (Female, 24, college educated) one recorded in the sound-booth at the University of Hong Kong and another recorded through Zoom using a laptop microphone. Acoustic features analyzed include F1, F2, and F3. Preliminary findings suggest that while F1 is fairly comparable, the higher two formants are altered in ways that question the comparability of Zoom-recorded vs. Sound booth-recorded vowels. However, a good degree of comparability is possible for most of the vowels if results are collected by hand.

Key Words: Xiang Chinese, recording methodology, acoustic phonetics, phonology

* The University of Hong Kong; E-mail: u3545575@connect.hku.hk; https://orcid.org/0000-0001-6869-1651

Acknowledgments: I would like to thank Viktorija Kostadinova and Matt Gardner at the Getting Data Working Group (https://gettingdata.humanities.uva.nl/) for their feedback and support in exploring this topic; their drive and motivation in establishing this group in response to the pandemic is truly inspiring. I would also like to thank Dr. Jonathan Havenhill for his comments on certain methodological issues; all errors are my own of course.

1 Introduction

As an increasing number of people become more accustomed to working remotely due to the pandemic, it has become increasingly difficult to meet in person with subjects for elicitation tasks in sound attenuated booths. This poses particular problems for phonetic research, which often requires a certain degree of acoustic quality in recordings which can only be guaranteed in a sound booth. In some cases, elicitation tasks have to take place on Zoom video conferencing software (Zoom Video Communications, San Jose, CA), the acoustic quality of which may largely depend on the informant's laptop microphone. While this may cause a variety of problems for acoustic analysis, it causes particular problems for field research on relatively isolated or localized varieties, such as Xiang Chinese, which is spoken in Central Hunan, China.

Given the recent difficulties associated with getting subjects into sound booths, let alone visiting rural areas in China for fieldwork, the present study seeks to query whether the acoustic quality of Zoom recordings can be treated as comparable with more standard methods of data collection such as sound booth recording. To this end, we explore the acoustic quality of vowels in a localized dialect of Xiang Chinese (Yiyang). Data was gathered from two Yiyang Xiang recordings retelling the events of the 'Pear Stories' video (see Chafe 1980), performed by one speaker (Female, 24, college educated) recorded in the sound-attenuated booth at the University of Hong Kong and recorded through Zoom using a laptop microphone. Vowels were analyzed for the first three formants (F1, F2, and F3), which were extracted in Praat (Boersma and Weenink 2021) and analyzed in R (R Core Team 2021). The following questions are raised: What are the general phonetic characteristics of vowels in the sample? Is there a noticeable difference in the acoustic quality of Zoom-recorded vs sound booth-recorded vowels? What limitations exist (if any) and how significant are they? Is Zoom a viable tool for acoustic research on lesser-studied or hard-to-reach varieties? The research hopes to explore these questions to the extent allowed by the data available.

Research related to the quality of audio recorded on Zoom has been conducted on English (Freeman and De Decker 2021), Mandarin (Ge et al. 2021), and vowel quality in various languages (Zhang et al. 2021) in the recent past, which attempt to answer similar questions to those posed in this study. This study attempts to broaden the perspective by focusing on an understudied language, whose speakers are difficult to reach outside China. The motivations for this research can be divided into two broad categories: practical and typological. Firstly, this research intends to inquire as to the feasibility of doing acoustic research over Zoom (in line with previous research on this topic), and whether acoustic data gathered in this way is still usable for conducting fieldwork from a distance. Secondly, this research hopes to expand on existing knowledge of Sinitic typology, shedding light on an understudied subgroup.

2 Background

This section introduces background on acoustic quality of video conferencing software (2.1) and Yiyang Xiang phonology (2.2), and establishes the basis for the present research. Most sources seem to agree that use of Zoom for fieldwork (and video conferencing software generally) is not issue-free, particularly due to its acoustic compression algorithm; this will be elaborated on below.

2.1 Recording medium and acoustic quality

Online remote data collection for acoustic research is particularly sensitive to two confounding factors, which do not apply to data collected in-person: packet loss and audio compression (Freeman and De Decker 2021; Zhang et al. 2021), both of which are possible concerns when using Zoom to record informants. Zhang et al. (2021) mention that recording device and file format, while generally leaving F0 unaffected, can alter formants in various ways. Freeman and De Decker (2021) note that, depending on Internet quality, "packet loss will destructively alter the sound quality of a digital recording that arrives on the researcher's

computer", and that the 'lossy' audio compression conducted by the 'Opus' audio codec (https://opus-codec.org/), used by Zoom, may remove certain acoustic features from the source signal (Salomon 2007). Both of these have potential to alter the signal in such a way as to complicate comparison with audio recordings done in-person. In particular, however, lossy audio compression is problematic, as it irreversibly alters the original signal by removing acoustic information (Bulgin et al. 2010; Salomon 2007), albeit ideally with loss of redundant or unimportant data. Packet loss due to changes in internet speeds can also reduce the amount of acoustic information transmitted (Freeman and De Decker 2021; Mayorga et al. 2003).

Bulgin et al. (2010) reported that lossy compression to an MP3 format does not change the first two formants or F0 appreciably, although it does produce noticeable alteration of F3 and F4. However, they do report considerable distortion of the vowel space when the signal is run through Skype's (Microsoft, Redmond, WA) compression algorithm, and advise against its use. De Decker and Nycz (2011) similarly find that lossy compression methods alter the vowel space substantially relative to 'lossless' compression, such as that used for .m4a recording on smartphones. Of course, since these were on the 2010-11 version of the software, much will have changed since then, including advances in the ways VoIP technologies like Skype and Zoom perform audio compression. More recent research by Ge et al. (2021) discusses the relative reliability of F0 and F1 measurements on Zoom, while the higher formants like F2 and F3 are altered in unpredictable ways.

Aside from conducting most research on English (Bulgin et al. 2010; De Decker and Nycz 2011; Freeman and De Decker 2021), most of the studies above perform tests on audio quality in highly controlled environments; for instance, in environments where it is possible to record the same informant using both traditional and online methods (Freeman and De Decker 2021). In addition, there is the added complication that many of the studies have tightly-controlled or non-naturalistic tasks, such as instructing informants to sustain cardinal vowels for several seconds in the case of Zhang et al. (2021). For a variety of reasons, these are not realistic given time constraints and informant availability, as well as the nature of the task relative to natural speech. There is little doubt that the compression method used by Zoom will alter audio in such a way as to remove acoustic information (particularly of the higher formants); the question is whether the loss is acceptable when conducting speaker extrinsic comparison, in a naturalistic linguistic elicitation environment such as narrative production.

2.2 Yiyang Xiang

Yiyang is a variety of Xiang Chinese (ISO: hsn; Glot: xian1251) spoken in Northern Hunan province. The dialect represents a variety of what is commonly called 'New' Xiang, referring to those varieties that are relatively more innovative and influenced by Northern varieties of Chinese such as Mandarin (Norman 1988; Wu 2005). It is also classified within the 'Changyi' subgrouping, owing to its similarity to the Changsha variety of Xiang, which represents the speech of the provincial capital.

In terms of Sinitic typology, Yiyang is fairly typical regarding its vocalic inventory, which includes seven contrastive vowels (Cui 1998); one feature of note is a contrast between a central schwa and a mid-front [e], where many varieties lack one or the other as a contrastive monophthong, or have a vowel closer to [ɤ] as the open variant of [ə]. The so-called apical 'vowel', found in many Sinitic languages, is also usually included in the vocalic inventory but is here excluded.[1] The contrastive monophthongs of Yiyang are shown below in Figure 1 (IPA chart from http://ilg.usc.es/ipa-chart/keyboard/).

[1] The apical vowel [ɿ] is listed by Cui and is traditionally treated as a vowel, but actually shares many features with syllabic consonants (as a voiced apical fricative) and is therefore not included.

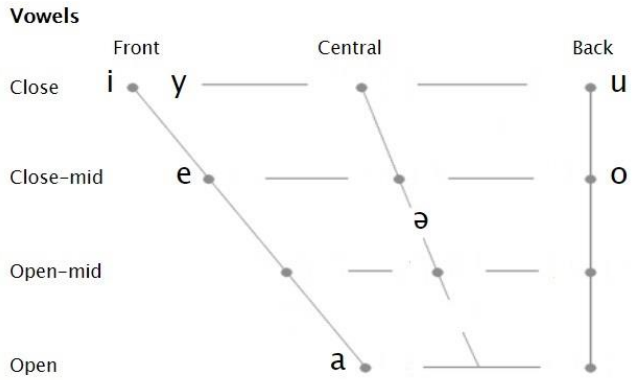

Figure 1. Yiyang monophthongs, following Cui (1998)

Each of these vowels occur in open CV syllables, excluding /e/: *ha1* [xa˧] 'he, she, it'; *li2* [li˨] 'pear'; *nu3* [ŋy˥] 'woman, female'; *ke2* [kʰə˨] 'to go'; *gu1* [ku˧] 'girl'; *go2* [ko˨] 'numeral classifier for humans'. The mid-front vowel /e/ occurs either preceding or following a glide, or both. There are no instances of it occurring in CV syllables; for instance: *xue5* [ɕɥe˩] 'to speak', *xuei3* [ɕɥei˥] 'water'. Cui also lists five lexical tones in Yiyang, with three contrasting heights: low, mid, and high. The five tones are: Mid-Level Tone 1, Low-Rising Tone 2, High-Falling Tone 3, Low-Level Tone 4, and High-Level Tone 5. In addition to the five lexical tones, Yiyang also has a contextually-supplied default tone, which primarily occurs in affixes and clitics.

3 Methodology

The data for this case study involved two 4"24 and 1"13 recordings (referred to as yy_booth, and yy_zoom, respectively) of a native speaker of Yiyang Xiang performing the Pear Stories paradigm (Chafe 1980), which involves watching a video depicting a child stealing pears from a farmer and then relating the events from memory into a microphone. The participant is 24 years of age, female, and lived in Yiyang for most of her life (18yrs), then going on to study in Changsha for four years (the provincial capital of Hunan province), and then going on to pursue postgraduate studies in Hong Kong for two years. The recording for yy_booth was conducted in person at the HKU sound attenuated booth (Sennheiser-MD46 microphone), recording through Praat (WAV file). yy_zoom was recorded over Zoom (v. 5.4.9) with the informant's laptop microphone, using Zoom's recording function (m4a file), over HKU's wifi connection (in order to ensure some degree of signal strength and reliability).

Acoustic analysis took place in Praat (Boersma and Weenink 2021); items analyzed included vowel tokens in (C)(G)V(G) syllables, extracting between 50-100ms of the steady-state portion of the vowel. Tokens were extracted from the recordings using the Praat Vocal Toolkit (Corretge 2012; http://www.praatvocaltoolkit.com/), and were then labelled and corrected manually to include the steady-state portion of the vowel. A modified version of a Praat script written by Stanley (2019; https://joeystanley.com/downloads/191002-formant_extraction.html) was used to measure formants at the vowel midpoint. Six vowel qualities were extracted from the lexical set, following the phonological analysis of Cui (1998; see Section 2.3).[2] In total, 308 monophthongal vowel tokens were extracted (yy_booth = 219, yy_zoom = 89; see Table I).

[2] The high front rounded vowel [y] was excluded from the analysis for lack of tokens in the sample.

Table I. Vowel tokens in the sample

	/a/	/e/	/i/	/ə/	/u/	/o/	Total
yy_booth	67	17	48	18	33	36	219
yy_zoom	29	9	20	6	7	18	89
Total	96	26	68	24	40	54	308

Acoustic features extracted included F1, F2, and F3; the first two formants were plotted using the NORM Vowel Normalization and Plotting Suite (Thomas and Kendall 2007) in order to define the acoustic vowel space in Yiyang; since F3 does not function as a distinguishing acoustic feature for any of the vowels in the sample (except perhaps for /o/, since /ə/ seems to have a back allophone [ɤ]), it was not included in the analysis of vowel space. All statistical analyses were conducted using R (R Core Team 2021).

4 Results

Results are summarized in Figure 2 (created with R). The two conditions seem to differ only slightly in terms of F1, and it is only in the higher formants that the difference begins to become noticeable, especially for F3. While the standard deviations are quite large in the higher formants, this is expected given the logarithmic increase in frequency from F1 to F3. The general trend seems to be for F2 and F3 to be lower in the Zoom condition. These are also potentially attributed to errors of the Praat script, which occasionally confuses F2 and F3; however, these errors were never substantial enough to blur the lines between vowels for yy_booth (see Figure 4), unlike in yy_zoom, where the errors were much more substantial and debilitating to the analysis.

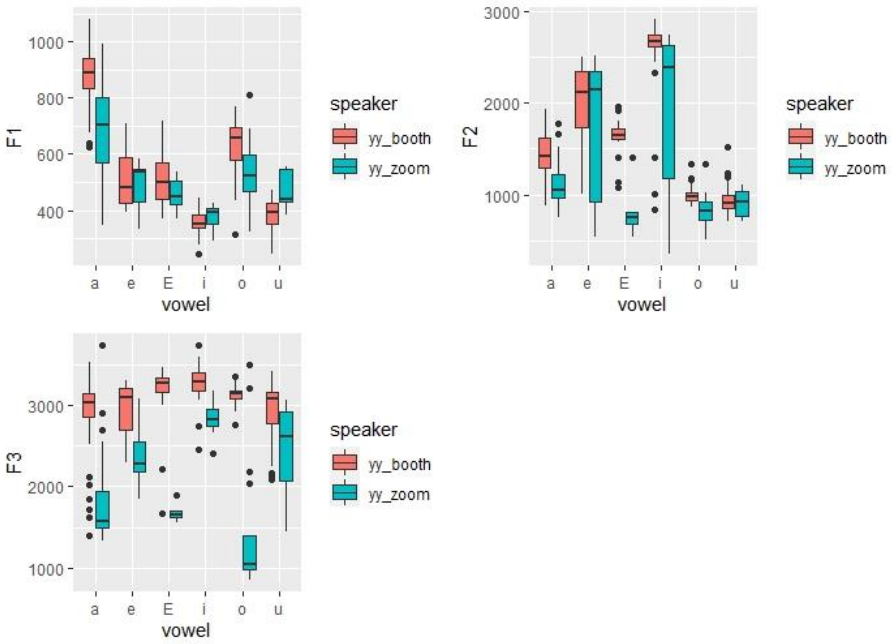

Figure 2. Data distribution for items in the sample ('E' = /ə/)

Figure 2 shows that comparability of means worsens as one moves through the formants; this was primarily due to an increase in the number of mis-identification errors. For the most part, average frequency of F1 is as expected for each of the target vowels; however, here we start to see increased confusion of F2-F3, which becomes substantial enough that F3 averages for /a/, /o/, and /ə/ are clearly too low, and the range of values for F2 is clearly too high. These errors, as Figure 4 in Section 4.1 shows, dramatically alter the vowel space for yy_zoom and blur the lines between /u/, /o/, and /ə/. Owing to the small sample of vowels for this recording, this effect might be a sampling error; however, errors increase dramatically with the Zoom recording, particularly for F2-F3, which has implications for our characterization of the vowel space (at least for F2), covered in the next section. If the errors are hand-corrected, however, results improve considerably (Section 4.2). The next section also covers linear models run comparing each vowel's F1-F2-F3 across recording conditions.

4.1 Vowel space comparison

Figure 3 shows the results for raw F1 and F2 values extracted from the two recordings.[3] Vowel qualities are represented with different colors, while recording conditions are differentiated by symbols (filled dot for yy_booth, empty square for yy_zoom); the six target vowel qualities are /i u e ə o a/. Of note is the wide distribution of values for /i/ in green for yy_zoom; this has altered the means considerably for this item, and resulted in considerable distortion of the vowel space (see Figure 4). This is also true for /ə/ ('E') and to a lesser extent /o/.

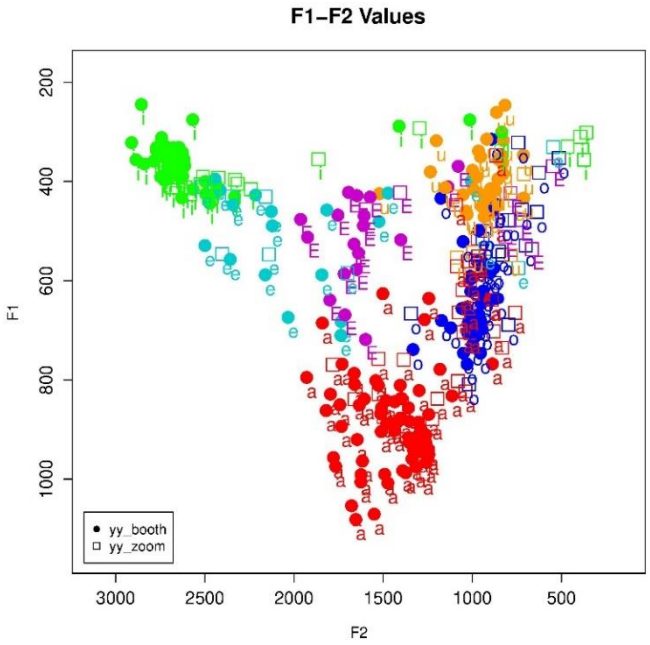

Figure 3. Raw vowel tokens in the sample ('E' = /ə/)

For the most part, a good degree of overlap is apparent across conditions, but the erroneous clustering along the upper right hand corner in the yy_zoom condition causes issues of comparison. Errors are also apparent for yy_booth, but these are not as dramatic or numerous as for the latter. If the means are calculated (at 1

[3] All vowel plots made with NORM (Thomas and Kendall 2007).

standard deviation), the vowel space resembles what we see in Figure 4, with expected results for yy_booth with a few errors; the errors for yy_zoom, however, are consistent enough to dramatically alter results.

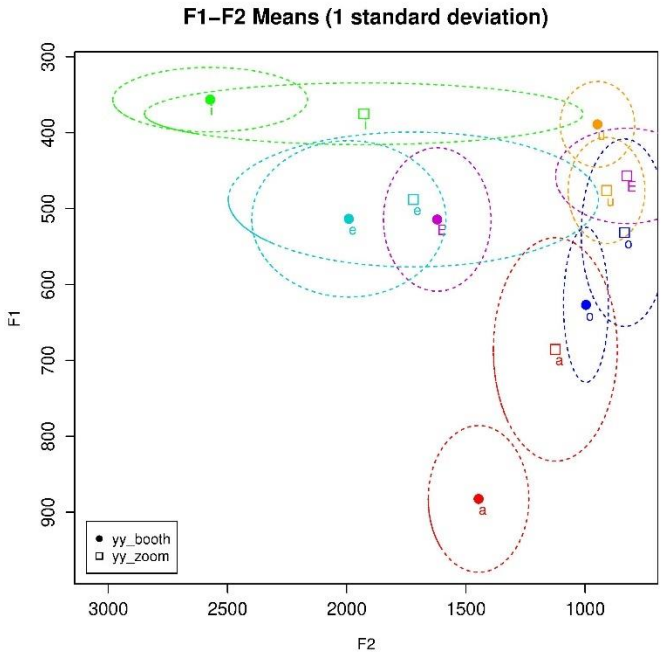

Figure 4. Raw vowel space for the two recordings (SD = 1)

The positioning of /ə/ for yy_zoom is most likely due to a sampling error (low number of tokens for /u/ and /ə/); the degree of overlap of /u ə o/ for yy_zoom means that these cannot be compared, as this would suggest that this speaker does not reliably distinguish these acoustically, even though this is clearly the case for yy_booth.

Linear models were run for each vowel on raw F1, F2, and F3, with recording condition as the fixed effect (yy_booth as intercept). Tables with the results of these tests can be viewed in the appendix (Tables I through VI). For the vowels /i/ and /ə/ (Tables III and IV) highly significant results clustered around the higher formants (F2 and F3) while F1 had no significant change. In two cases all three formants had highly significant differences (vowels /a/ and /o/, Tables I and VI); in fact, F1 is significantly different in three out of the six vowels (/a/, /o/, and /u/). The vowel /u/ is an outlier in that F2 did not have a significant effect while F1 did (Table V). The vowel /e/ (Table II) had significant effects only for F3, which is unexpected since F2 was significant for all vowels minus /u/. However, the general trend across all of these, excepting /u/, is for there to be increasing significance as one moves from F1 to F3, and for all yy_zoom formants to be lower than their yy_booth equivalents. For instance, consider the trend for /e/ (Table II): F1 [t = -0.628, p = 0.536], F2 [t = -1.175, p = 0.251], F3 [t = -3.832, p = 0.000804 ***]. While results are only significant for F3, the general trend is for significance to increase. Notice also that in each case the estimate is negative, indicating that the formants are lower in frequency than in the Booth condition; this is the case for every vowel and formant, with a single exception being the F1 of /i/.

Differences in F3 are almost without exception highly significant, the odd one out again being /u/, with its F3 being only slightly significant: [t = -2.454 , p = 0.0188 *]. This high discrepancy for the two conditions can be attributed to misidentification of F3 for F2 or vice versa (assumed to be an F3 under 2000Hz), which occurred in 46 out of 89 items in yy_zoom; compare this with 6/219 for yy_booth. This is most likely due to a script error and are almost exclusive to the vowels /a/, /ə/, and /o/, but it is telling that the Zoom

condition had errors for over half of the items.[4] Because of this high error rate, all of the items were checked and corrected manually for comparison; the results for the clean tokens are presented in the following section.

4.2 Results after hand-correction

The method for hand-correction involved taking mean F1-F2-F3 across the steady-state portion of each vowel token; two data points off each end were excluded in avoidance of contextual effects. Results improve considerably if the data are cleaned manually (see Tables VII through XII): the only significant results are F3 of /e/ [t = -2.431, p = 0.0229 *]; F1 [t = 4.979, p = 4.86e-06 ***], F2 [t = -2.929, p = 0.00466 **], and F3 [t = -4.534, p = 2.5e-05 ***] of /i/; and F1 of /u/ [t = 3.443 p = 0.00141 **] and /o/ [t = 3.262, p = 0.00195 **]. This is summarized in Figure 5.

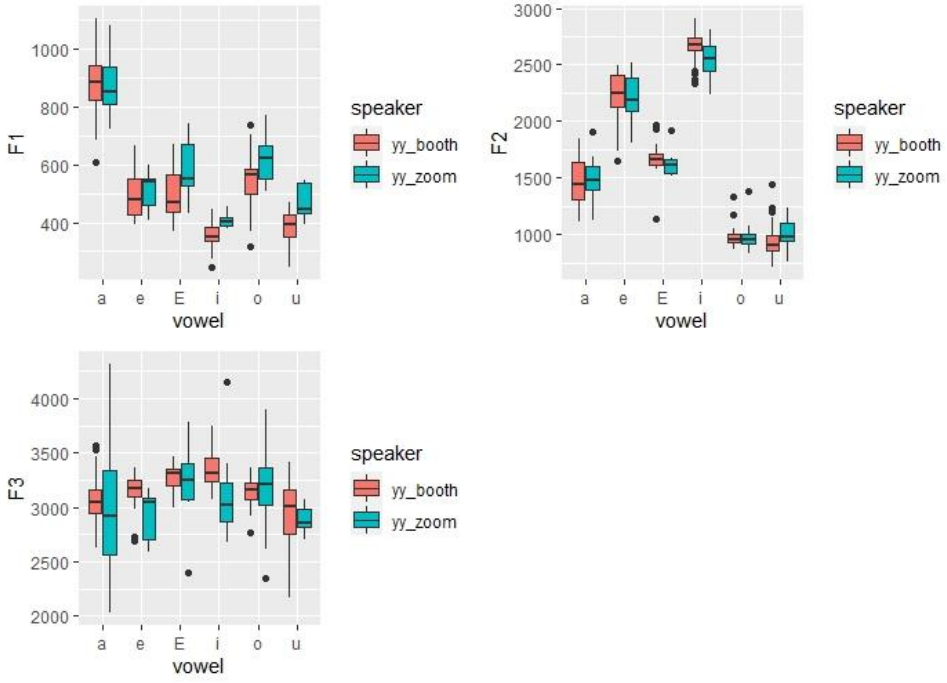

Figure 5. Means and distribution after hand-correction ('E' = /ə/)

Notice that while higher formants of /e/ and /i/ in yy_zoom all display decreases in average frequency (as occurred with the raw measurements in Section 4.1), average F1 *increases* relative to yy_booth; this is not in line with any predicted outcome, and in fact contradicts findings from a variety of sources (Ge et al. 2021; Zhang et al. 2021). F3, while generally being lower in yy_zoom, is only significantly so in two cases. Other than these cases however, values were generally as expected, with no clear distortion of higher formants in the Zoom condition. The values for the corrected vowel tokens can be seen below in Figure 6.

[4] For /a/ and /o/, this trend seems to be due to the closeness of F1 and F2 which is generally seen in these vowels; the low F2 of /o/ might overlap with F1, while the high F1 of /a/ might overlap with F2. For /ə/, the reason for the confusion is unknown, but probably also has to do with the closeness of F1 and F2.

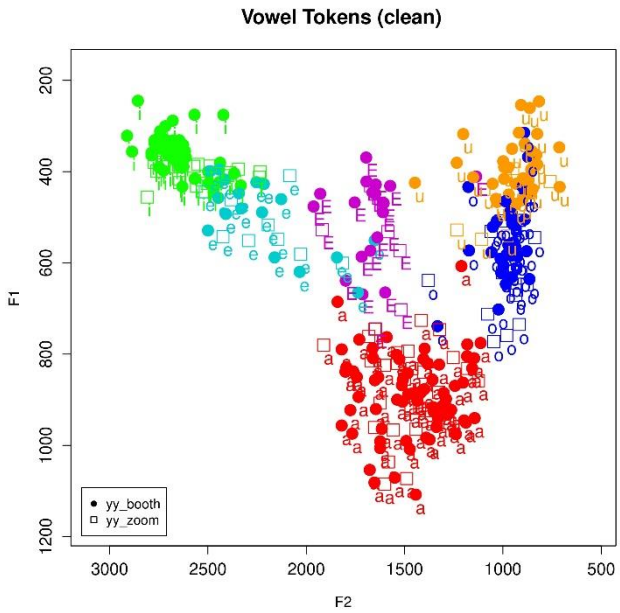

Figure 6. Clean vowel tokens ('E' = /ə/)

Figure 6 no longer displays the wide range of F2 values seen in Figure 3, nor the overlap of /a/, /o/, and /ə/. While there are still a few outliers, they seem to be evenly distributed across both conditions. A snapshot of the vowel space can be observed in Figure 7.

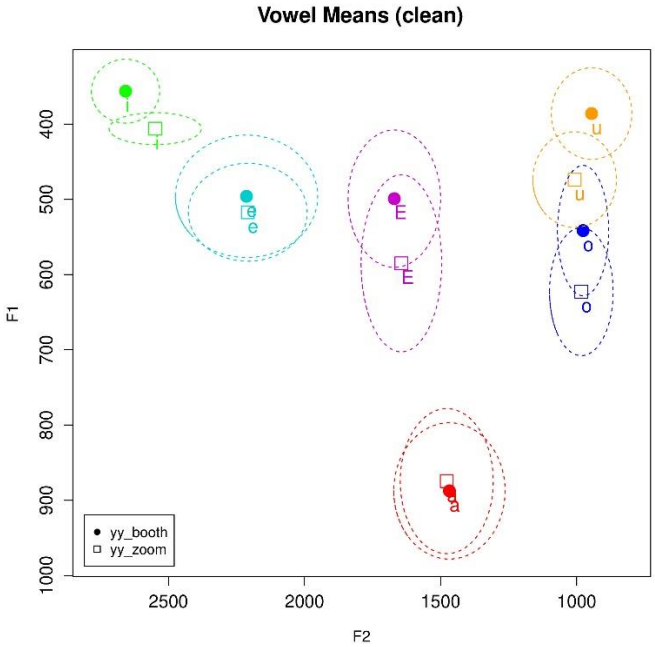

Figure 7. Vowel space for clean vowel tokens

There is a good amount of overlap for equivalent vowel categories across conditions, with the locations being as expected (this can be clearly contrasted with Figure 4). Notice also the increase in F1 in yy_zoom, which appears relatively consistent across most vowels. At the same time, the standard deviation for F1 and

F2 is not too large, especially when compared to that seen in Figure 4. In short, the two conditions are largely comparable as long as values are verified by hand.

5 Discussion and conclusions

For the raw data, the wide divergences seen for F2-F3 and the relative faithfulness of F1 are predicted in previous work by Ge et al. (2021), who observe effects on the higher formants; perhaps the main difference in this instance is that F1 performed worse than expected. The Zoom condition seems to display general distortion of the of the higher formants across most vowels (and all formants in at least two cases: /a/ and /o/), which necessitates hand-correction to improve comparability. In terms of the vowel space, the degree of overlap of /a/, /ə/, and /o/ is troublesome for comparison across conditions, and seems to be due to errors in formant identification in some way related to the recording medium. In the Zoom condition, mis-identification of F2 for F3 caused considerable lowering of F3, resulting in low similarity when compared to yy_booth.[5] For instance, average F3 for /o/ in yy_zoom is far below the expected average, at just below 1.5 kHz. This is attributed to the exaggerated number of errors in F3 identification in the Zoom condition (for yy_zoom accounting for more than half of the sample), mentioned in the preceding section.

These two distortion effects (mis-identification of F2 and F3, altering of F2) are predicted based on how compression algorithms operate on acoustic data; acoustic compression favors lower frequencies (F0-F1) at the expense of higher frequencies (F2-F3) sacrificing detail at these frequencies as they are less likely to pose issues for speech perception. However, it should be noted that here even F1 is seriously affected, although not for all vowels equally; for instance, it is still fairly comparable for /e/ and /i/. Suboptimal conditions (low internet speeds, low microphone quality) may magnify these effects and cause additional complications for analysis; this is why hand-correction of the data is necessary. In addition to altering of higher frequencies, there appears to be some targeting of particular vowels. In condition yy_zoom there is significant F2 distortion for /a/, /ə/, /o/, and /i/ while for /e/ and /u/ it is insignificant. The reasons for this are somewhat obscure.

Manual formant extraction seems to be a necessity for data usability, although even here results are not without issue. As with the previous results, particular vowels (/e/ and /i/) seem to be targeted, and the increase in F1 is unexpected. The poor performance of the script in extracting reliable measurements in the Zoom condition may be due to a variety of factors, not all of which may be Zoom-related (microphone quality, internet speed), which thus necessitates this additional step.

These results can be compared to those found in Ge et al. (2021), Zhang et al. (2021), and Freeman & DeDecker (2021). In the first pass without manual correction, the results matched those of Ge et al. (2021), where F2 seemed to be affected by substantial lowering (particularly for /i/ but not for /u/). This is also what is found for front vowels in Zhang et al. (2021). However, once the data had been cleaned manually, results were more like those of Freemand & DeDecker (2021), who find that the changes to formants are mostly minimal and are only of concern when conducting analyses of vowel overlap (recall that /i/ is still problematic across all formants even after the data is hand-corrected, and that F1 seems to increase). The targeting of low and back vowels they find is not identified in the present research.

One clear limitation is the lack of simultaneous recording of the speaker; recording yy_zoom took place at a later date only on Zoom. Most other studies on using conferencing software for acoustic research perform simultaneous recording across several different devices and software (Freeman & DeDecker 2021; Ge et al. 2021, among others), which eases comparability. It is quite possible that results were altered by the low sample size of yy_zoom relative to yy_booth, which would have been avoided if the exact same vowels

[5] Since F3 does not factor in on the distinction between these six vowel qualities (F1-F2 sufficing lacking contrasts in rounding), it does not alter the vowel space (see Figure 4).

were being analyzed. Also, it cannot be ruled out that the final unexpected results, such as increase in average F1, might be in some way due to this methodological flaw, given that none of the studies cited report this effect. However, the goal of this research was to investigate the usability of Zoom in a naturalistic recording environment (or as a method to simplify fieldwork), where one may not be able to record an informant both in the sound booth and online to check for inconsistencies. Additionally, given the high-comparability following from the correction in Section 4.2, it does not appear that the vowels are substantially altered despite being recorded at different times.

Comparability across Zoom and Sound Booth recordings is not without issue, and manual correction is necessary owing to the high degree of formant identification errors in the Zoom conditions. For this reason, Zoom's usefulness as an instrument for remote data collection (at least for acoustic data) is called into question for vowel faithfulness (particularly for F2-3). As other studies point out (Zhang et al. 2021; Freeman & DeDecker 2021; Ge et al. 2021), using software like Zoom is still possible for studies on prosody, tone, or those that do not require a high level of vowel faithfulness. How much detail a researcher is willing to sacrifice will depend on a variety of factors and should be considered on a case by case basis.

6 Appendices

Table I. Linear Model for raw low /a/ (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	882.42	13.95	94	63.26	< 2e-16 ***
yy_zoom	-196.95	25.38	94	-7.76	1.02e-11 ***

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	1446.78	27.73	94	52.180	< 2e-16 ***
yy_zoom	-321.94	50.45	94	-6.382	6.56e-09 ***

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	2924.43	57.33	94	51.01	<2e-16 ***
yy_zoom	-1074.75	104.31	94	-10.30	<2e-16 ***

Table II. Linear Model for raw mid-front /e/ (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	513.36	23.91	24	21.469	<2e-16 ***
yy_zoom	-25.53	40.64	24	-0.628	0.536

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	1991.1	135.5	24	14.696	1.7e-13 ***

yy_zoom	-270.6	230.3	24	-1.175	0.251

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	2959.98	90.74	24	32.621	< 2e-16 ***
yy_zoom	-591.05	154.23	24	-3.832	0.000804 ***

Table III. Linear Model for raw high-front /i/ (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	356.243	6.056	66	58.827	<2e-16 ***
yy_zoom	18.678	11.166	66	1.673	0.0991

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	2572.64	86.96	66	29.58	< 2e-16 ***
yy_zoom	-644.57	160.35	66	-4.02	0.000152 ***

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3289.55	29.61	66	111.103	< 2e-16 ***
yy_zoom	-431.71	54.59	66	-7.907	3.88e-11 ***

Table IV. Linear Model for raw mid-central /ə/ (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	514.18	20.84	22	24.677	<2e-16 ***
yy_zoom	-57.40	41.67	22	-1.377	0.182

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	1620.35	57.61	22	28.125	< 2e-16 ***
yy_zoom	-796.08	115.22	22	-6.909	6.15e-07 ***

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3123.12	95.03	22	32.865	< 2e-16 ***
yy_zoom	-1443.18	190.05	22	-7.593	1.39e-07 ***

Table V. Linear Model for raw high-back /u/ (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	388.83	10.24	38	37.983	< 2e-16 ***
yy_zoom	87.09	24.47	38	3.559	0.00102 **

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	948.16	27.36	38	34.651	<2e-16 ***
yy_zoom	-38.10	65.41	38	-0.582	0.564

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	2901.98	78.09	38	37.163	<2e-16 ***
yy_zoom	-458.04	186.67	38	-2.454	0.0188 *

Table VI. Linear Model for raw mid-back /o/ (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	626.70	18.28	52	34.285	<2e-16 ***
yy_zoom	-95.09	31.66	52	-3.004	0.0041 **

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	996.3	21.6	52	46.136	< 2e-16 ***
yy_zoom	-162.0	37.4	52	-4.331	6.79e-05 ***

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3134.29	77.12	52	40.64	< 2e-16 ***
yy_zoom	-1699.08	133.57	52	-12.72	< 2e-16 ***

Table VII. Linear Model for low /a/, manually corrected (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	887.53	11.29	94	78.64	<2e-16 ***
yy_zoom	-12.94	20.53	94	-0.63	0.53

F2	**Estimate**	**SE**	**df**	***t***	***p***

Intercept (yy_booth)	1467.545	23.848	94	61.538	<2e-16 ***
yy_zoom	9.752	43.390	94	0.225	0.823

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3075.40	42.91	94	71.673	<2e-16 ***
yy_zoom	-50.65	78.07	94	-0.649	0.518

Table VIII. Linear Model for mid-front /e/, manually corrected (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	496.01	18.53	24	26.772	<2e-16 ***
yy_zoom	21.23	31.49	24	0.674	0.507

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	2213.061	60.210	24	36.756	<2e-16 ***
yy_zoom	-4.594	102.337	24	-0.045	0.965

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3126.22	48.55	24	64.395	<2e-16 ***
yy_zoom	-200.58	82.51	24	-2.431	0.0229 *

Table IX. Linear Model for high-front /i/, manually corrected (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	356.243	5.427	66	65.637	<2e-16 ***
yy_zoom	49.827	10.008	66	4.979	4.86e-06 ***

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	2657.40	20.16	66	131.819	<2e-16 ***
yy_zoom	-108.89	37.17	66	-2.929	0.00466 **

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3348.94	31.47	66	106.428	<2e-16 ***
yy_zoom	-263.06	58.02	66	-4.534	2.5e-05 ***

Table X. Linear Model for mid-central /ə/, manually corrected (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	499.11	23.09	22	21.616	2.6e-16 ***
yy_zoom	85.95	46.18	22	1.861	0.0761

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	1670.06	38.96	22	42.866	<2e-16 ***
yy_zoom	-25.99	77.92	22	-0.334	0.742

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3274.00	56.78	22	57.661	<2e-16 ***
yy_zoom	-83.64	113.56	22	-0.736	0.469

Table XI. Linear Model for high-back /u/, manually corrected (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	385.99	10.66	38	36.206	<2e-16 ***
yy_zoom	87.75	25.48	38	3.443	0.00141 **

F2	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	945.22	25.95	38	36.424	<2e-16 ***
yy_zoom	62.34	62.34	38	1.005	0.321

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	2929.19	52.36	38	55.94	<2e-16 ***
yy_zoom	-41.27	125.17	38	-0.33	0.743

Table XII. Linear Model for mid-back /o/, manually corrected (F1/F2/F3 ~ Recording Condition)

F1	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	541.57	14.36	52	37.712	<2e-16 ***
yy_zoom	81.15	24.87	52	3.262	0.00195 **

F2	**Estimate**	**SE**	**df**	***t***	***p***

Intercept (yy_booth)	976.440	17.009	52	57.406	<2e-16 ***
yy_zoom	6.449	29.461	52	0.219	0.828

F3	**Estimate**	**SE**	**df**	***t***	***p***
Intercept (yy_booth)	3140.19	44.50	52	70.560	<2e-16 ***
yy_zoom	57.63	77.08	52	0.748	0.458

7 References

Boersma, Paul & David Weenink. 2021. *Praat: Doing phonetics by computer*. Available at: http://www.praat.org/.

Bulgin, James, Paul De Decker, & Jennifer Nycz. 2010. Reliability of formant measurements from lossy compressed audio. Paper presented at the British Association of Academic Phoneticians Colloquium, University of Westminster, 29-31 March.

Chafe, Wallace (ed.). 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.

Corretge, Ramon. 2012. *Praat Vocal Toolkit*. Available at: http://www.praatvocaltoolkit.com/.

Cui, Zhenhua. 1998. *Yiyang fangyan yanjiu* [A study of the Yiyang dialect]. Changsha: Hunan Education Press.

De Decker, Paul & Jennifer Nycz. 2011. For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics*, 17(2). 51–59.

Ge, Chunyu, Yixuan Xiong, & Peggy Mok. 2021. How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements. *Proc. Interspeech 2021*. 3984-3988.

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

Mayorga, Pedro, Laurent Besacier, Richard Lamy, & J.-F. Serignat. 2003. Audio packet loss over IP and speech Recognition. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding* (IEEE Cat. No.03EX721). 607-612.

Norman, Jerry. 1988. *Chinese*. Cambridge University Press.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Salomon, David. 2007. *A Concise Introduction to Data Compression*. Springer Science & Business Media.

Stanley, Joey. 2019. *Automatic Formant Extraction in Praat*. Available at: https://joeystanley.com/downloads/191002-formant_extraction.html/

Thomas, Erik & Tyler Kendall. 2007. *NORM: The vowel normalization and plotting suite*. Eugene, OR: University of Oregon. Available at: http://ncslaap.lib.ncsu.edu/tools/norm/.

Zhang, Cong, Kathleen Jepson, Georg Lohfink, & Amalia Arvaniti. 2021. Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America*, 149(6), 3910-3916.