INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements

Chunyu Ge , Yixuan Xiong , Peggy Mok

The Chinese University of Hong Kong, China

chunyuge@cuhk.edu.hk, yixuanxiong@cuhk.edu.hk, peggymok@cuhk.edu.hk

Abstract

The COVID-19 pandemic posed an unprecedented challenge to phonetic research. On-site collection of speech data is difficult, if not impossible. The advancement of technology in mobile devices and online conference platforms offers the opportunity to collect data remotely. This paper aims to answer the question of how reliable speech data collected remotely are based on controlled speech. Seven devices, including smartphones and laptops, were used to record speech simultaneously, locally or on the cloud using ZOOM, both in a sound-attenuated lab and a conference room. Common acoustic measurements were made on these recordings. Local recordings proved to be reliable in duration, but not for recordings made on the cloud. Different devices have comparable performances in F0 and F1. The values acquired by different devices differ a lot for F2 and higher formants, spectral moments, and voice quality measures. These differences can lead to erroneous interpretation of segmental and voice quality contrasts. The recordings made remotely by smartphones and locally using ZOOM can be useful in studying prosody, but should be used with care for segments.
Index Terms: smartphone, ZOOM, acoustic measurements

1. Introduction

Traditionally, much emphasis has been put on the recording quality for data collected for phonetic research. High quality recording made with professional recording devices in sound treated rooms (e.g., phonetics labs) was considered the norm. However, the COVID-19 pandemic has caused severe disruption to all aspects of life, including data collection, because of lockdowns and travel restrictions. Moreover, the airborne nature of COVID-19 has rendered the behavior of speaking into a common microphone highly risky. Before the pandemic, multiple online experimental platforms are already available for conducting perception or recognition experiments (e.g., Qualtrics, Gorilla, Lioness Lab), and there is also a proposal to use mobile devices to collect speech data for less accessible communities [1]. The advancement of the Internet and mobile technologies has opened new possibilities of collecting speech data. For example, smartphone applications can record in very high quality (44.1KHz, 16 bits or better), and they are very common and easy to operate. They can potentially provide a practical solution for remote recording. Moreover, the online conference platform ZOOM [2] not only allows real-time interaction with remote participants, but can also record the online meetings automatically on their cloud server or locally in users' computers. Thus, it seems quite feasible to collect speech data remotely, but how comparable the recordings are to those recorded in ideal conditions is a critical issue to investigate.

There have been some recent studies exploring the above question and several factors are at play. The first factor is the recording device. [3] discussed the general issues and challenges involved in using smartphones to collect speech data and [4] concluded earlier that iPhone recordings were usable to investigate vowel spaces. [5] found no significant differences across devices for voice quality measures (F0, jitter, shimmer and HNR). Recently, [6] systematically compared the recordings made by ten devices including professional recorders, a laptop, and smartphones of different brands (both iPhones and Android) in a phonetic lab and in a quiet room. They found that all devices had good frequency response. F0 data were relatively consistent, but some differences were found in F1, F2 and Center of Gravity (CoG). However, F3, shimmer, jitter and H1-H2 showed high variations, contrary to the findings in [5], but concurring with those in [7] who also found robust F0 data but problematic data for shimmer and jitter.

Software also have an influence on the acoustic measurements. While [8] explicitly advised against the use of ZOOM recording for phonetic analysis because of the variable devices used by the participants and the audio compression algorithms applied by ZOOM, [9] found some comparable acoustic measurements from recordings made using the ZOOM meeting application, non-lossy smartphone applications and reference recording using a solid-stated recorder. Similar to the above studies, F0 is the most robust across devices, followed by lower formants. The upper formants showed more variation. Both smartphones and ZOOM performed better for vowels than for pure tones. Interestingly, while duration was not measured by most previous studies, [9, 10] found that irregular waveforms and filtering algorithm artefacts caused considerable durational differences for ZOOM (~119ms), whereas durational measurements by smartphones are reliable. They did not take acoustic measures about voice quality, however. Recordings made locally may also be different from that recorded on the cloud, yet [9, 10] used local ZOOM recordings only.

Yet another factor is the audio format. For example, [11] suggested that audio compression is more optimized for male than for female speakers. Moreover, while different devices and software yield different audio formats (e.g., `mp3`, `m4a`, `aac`), usually the `wav` format is used by the common speech analysis tools. [6] used recordings of different formats, yet no information was provided on the conversion procedure. The present paper documents the conversion procedures in detail, which will facilitate reproducibility. But due to the technical intricacies involved in data compression and digitization, a full treatment of the issue is beyond the scope of this paper.

Our project builds on the aforementioned studies and focuses on the comparison between ZOOM recordings (both cloud and local) with those recorded by different smartphones (iPhones and Android) and a solid-state recorder simultaneously recorded in a phonetics lab and in a quiet conference room. Another inadequacy of the previous studies is that the materials used were either pure tones [9] or speech material with no control [6]. The present study thus focuses on controlled speech materials, which

http://dx.doi.org/10.21437/Interspeech.2021-1122

enables us to compare acoustic measurements and interpret the difference in a linguistically meaningful manner. Detailed comparison of the performance of different smartphones is beyond the scope of our study since there are many brands available on the market. Interested readers should consult the relevant studies reviewed above. We hope to see what is a viable protocol for collecting speech data remotely online.

2. Method

To compare common acoustic measurements used in phonetic research, a word list of Mandarin Chinese was compiled based on its inventory. All words are monosyllabic with CV structure, where V = /i u a/, and C covers stops, fricatives, affricates, and sonorants (nasals, laterals, and rhotics). Systematic gaps were ignored, and accidental gaps were filled in with *Pinyin*. There are 201 words in total.

Table 1: *Recording devices, formats, and the sampling frequencies (SF)*

Device	Format	ID	SF
Zoom H2N	wav	H2N	44100
iPhone 6s	m4a	IP6	48000
iPhone 8p	m4a	IP8	48000
Samsung Galaxy A50s	wav	SSG	22050
Oneplus Nord N10	aac	OPS	48000
DELL XPS15 (ZOOM local)	m4a	ZML	32000
MacBook Pro (ZOOM cloud)	m4a	ZMC	32000

One female speaker (the second author) was asked to read aloud the word list. The recording was conducted in a sound-attenuated lab and a quiet conference room. The word list was repeated twice for each condition. Seven devices were used to record the speech simultaneously. The details of these devices, together with the audio formats and sampling frequencies [1], can be seen in Table 1. Efforts were made to maintain an approximately similar distance between the devices and the speaker. The recordings of IP6, IP8, and OPS were made using the built-in recorder, while the app Smarter Recorder was used for SSG. For the laptop recording locally using ZOOM (denoted as ZML), and one of the iPhones (IP6), the speech was recorded through a plugged-in earphone, of which the microphone was held near the speaker's mouth. Another laptop (MacBook Pro) was used to record using the cloud server of ZOOM (thus denoted as ZMC), which was placed in another room. These settings were the same for the recordings in the lab and in the conference room. The settings were designed to simulate the realistic situation of recording by amateur participants and more sophisticated settings were left uncontrolled.

Due to the requirements of speech analysis tools (Praat and VoiceSauce), all the audio files were converted to `wav` format. The recordings made by ZOOM locally and on the cloud are in `m4a` format, which were converted to `wav` using iTunes, a built-in software of macOS (now the Music app). The recordings made by OPS are in `aac` format, which was converted to `m4a` using QuickTime Player, and then converted to `wav` format following the same procedure, together with the recordings of IP6 and IP8. The `aac` format cannot be directly converted to `wav` using iTunes, and QuickTime Player, which is also a built-in software of macOS, was used to make the influence of conversion as minimal as possible.

After conversion, all recordings were trimmed according to the first and last syllables. The durations of different devices are quite comparable (differences of a few ms), except for ZMC, of which the difference amounts to 30 ms at worst. Segmentation was carried out based on the recordings of the solid-state recorder H2N. Later inspection showed that the boundaries align neatly for recordings made by other devices, again except for ZMC. The boundaries differ wildly from the acoustic landmarks shown in the waveform and the spectrogram. The segmentations were corrected manually for ZMC, and the segmentations based on H2N were used for other devices. One possible reason for the discrepancy may be the delay of Internet connection, since the recordings of ZMC was made through the ZOOM cloud server.

F0, formants, and H1*-H2* values were measured using VoiceSauce [12]. Jitter, shimmer, and spectral moments [13] were measured using Praat [14]. F0, formants, and voice quality related measures were taken on the vowel, while spectral moments were only measured on fricatives. Data analyses were carried out in R [15], and linear mixed effects model was constructed using the `lmerTest` package [16]. For all models, acoustic measurements based on H2N were used as reference, i.e., data based on other devices were compared to those of H2N.

3. Results

3.1. Duration

As explained in the Method section, the same segmentation is applied to recordings made by all devices as they are very comparable, except for ZMC, which was manually corrected. Thus, only the durations based on recordings made by H2N and ZMC are compared here. The mean duration based on ZMC is significantly lower than that based on H2N ($\hat{\beta} = -7.5, t = -5.31, p < 0.01$). There is also an effect of segment types. The comparison of duration of different kinds of segments was made using the `emmeans` package [17]. The difference is greatest for fricative, which is 19.83 ms shorter for ZMC. That for vowel amounts to 7.51 ms shorter, yet given the relatively long duration of vowel, the difference is quite marginal. Those for stop, affricate, and sonorant are quite comparable (~ 1 ms). There is indeed a durational difference between recordings made locally and on the cloud. The difference is most pronounced for fricative.

3.2. F0

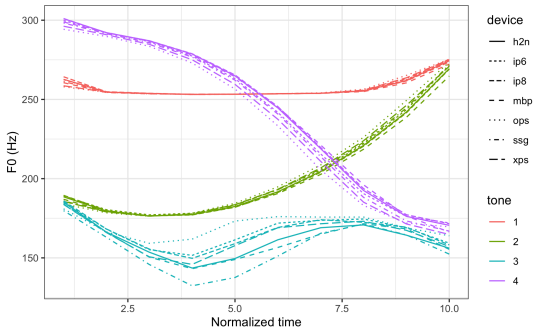

Figure 1: *Mean F0 trajectories of four Mandarin tones for different devices (combining the data of the lab and conference room, the same for Figures 2 and 3)*

[1] The sampling frequencies were achieved in Praat after the audio files were transformed to `wav` format.

F0 values were measured at ten equidistant points during the vowel portion. Figure 1 shows the mean F0 trajectories of four Mandarin tones for differnt devices. A separate linear mixed effects model was constructed for each tone to make the model easy to interpret. The influence of device is smallest for T4 (high falling), where only OPS is significantly lower than H2N on the intercept ($\hat{\beta} = -6.00, t = -3.85, p < 0.01$). It is similar for T3 (dipping), where the only significant effect is contributed by SSG ($\hat{\beta} = -9.12, t = -3.46, p < 0.01$), again lower on the intercept. A special note on T3 is that the speaker pronounced it with creaky voice (a common feature of T3), and the pitch tracking can be inaccurate. We will discuss voice quality measures further below. OPS again has a significant effect on the intercept of T2 ($\hat{\beta} = -3.69, t = -2.97, p < 0.01$), but it also has a significant effect on the slope ($\hat{\beta} = 1.05, t = 5.25, p < 0.01$), which is also the case for ZML ($\hat{\beta} = 0.57, t = 2.85, p < 0.01$) and SSG ($\hat{\beta} = 0.49, t = 2.46, p = 0.014$). T1 is similar to T2 in that the significant effects are also contributed by OPS, SSG, and ZML. What is unexpected is that ZMC performed better than ZML in F0, given its poorer performance in duration discussed above. The recording environment also has an effect on F0. F0 is a few Hz lower for recordings made in the conference room than in the lab (e.g., $\hat{\beta} = -9.02, t = -23.33, p < 0.01$ for T4). Although some of the devices and the environment influence F0, the effect is small. The difference between the devices and the reference (H2N) never exceeds 10 Hz, with only a few exceeding 5 Hz. OPS seems to have the worst performance, but the recording made by OPS is in `aac` format, which is later converted into `wav`. It is not clear if the influence is caused by the smartphone itself or the conversion algorithm.

3.3. Formants

Formants were taken at the fifth point out of ten during the vowel portion. Values greater than twice the standard deviation were removed. Since creaky voice is present in the middle portion of T3, which can lead to erroneous formant values, the data of T3 were also removed. The effect of devices on formants is remarkable, as can be seen in Figure 2, which plots the mean values of /i u a/ in the F1 × F2 plane, and the devices are marked by different colours.

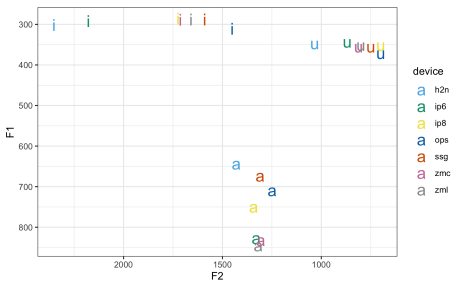

Figure 2: *Mean values of F1 and F2 of /i u a/ for different devices*

The most remarkable difference is in the high front vowel /i/. Only IP6 has relatively closer F2 to H2N, and other devices all result in lower F2, which drags the front vowel to a central position. In terms of the entire vowel system, the devices other than H2N drag the whole vowel triangle down and back. Again due to the high front vowel, the contrast in F2 is greatly undermined. The devices seem to be consistent in F1, especially for high vowels. This is confirmed by statistical analyses. None of the devices differ significantly from H2N in F1 for /i/, and only OPS has slightly higher F1 ($\hat{\beta} = 23.66, t = 3.74, p < 0.01$) for /u/. But for the low vowel /a/, all but SSG ($\hat{\beta} = 26.68, t = 1.64, p = 0.1$) have significantly higher F1 than H2N. The difference is between 100 to 200 Hz. Statistical analyses also corroborated the observation that only IP6 has similar F2 to H2N ($\hat{\beta} = -148.86, t = -1.43, p = 0.15$). It is followed by ZMC ($\hat{\beta} = -496.69$) and the difference is greatest for OPS ($\hat{\beta} = -798.68$). For /u a/, the difference in F2 is significant for all devices, but it is smaller, around 100 Hz for /a/ and 200 Hz for /u/. Again, iPhones (IP6 and IP8) have the closest value to H2N, and OPS the most different. Although almost all devices are significantly different from H2N in F3, the magnitude is comparable to those of F2. For the high back vowel /u/, the difference never exceeds 100 Hz for any device. Therefore, these devices have comparable performance in F1, but varies greatly for higher formants. The effect of recording environment is not consistent for vowels. All the formants of /i/ are not significantly different in the conference room from the lab, but recordings in the conference room had higher F1 and lower F3 for both /u a/.

3.4. Spectral moments

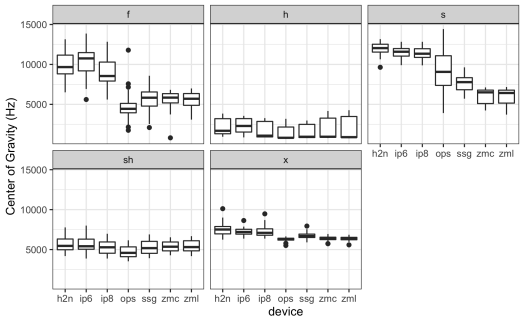

Figure 3: *Center of Gravity (CoG) of fricatives for different devices*

Center of Gravity (CoG), spectral skewness, kurtosis, and standard deviation were measured for the fricatives /f s ɕ ʂ h/. These measurements showed great variation among different devices, probably due to the differences in sampling frequency (Table 1). As an illustration, Figure 3 presents box-plots of CoG of fricatives for different devices. Each panel corresponds to a fricative, and the device is marked on the abscissa. /ɕ/ is represented by "x", and /ʂ/ by "sh" (i.e., *pinyin*). The patterns of /s f/ are distinct from other fricatives. For /s f/, H2N, and the two models of iPhone (IP6, IP8) group together, while other devices group together and is much lower than the former. Although such a grouping trend still holds for other fricatives, the difference between the two groups is much smaller. The data of CoG were thus divided into three groups (/ɕ/ is treated separately because it is slightly different from /h ʂ/), and a separate model was run for each subset. For /s f/, IP6 has similar values to H2N ($\hat{\beta} = -68.24, t = -0.29, p = 0.77$), and the values of IP8 are much lower ($\hat{\beta} = -756.51, t = -3.19, p < 0.01$). But the values of other devices are several thousand Hz lower than H2N. This is also the pattern for /h ʂ/, where the values of IP6 and IP8 are relatively similar to H2N compared to other devices. For /ɕ/, this is also the case where the difference between the two iPhones and H2N is around 300 Hz and that of other devices amounts to 1000 Hz.

The grouping of the two iPhones with H2N can also be seen in other spectral moments. For example, IP6 and IP8 are the only

devices of which the values of kurtosis, skewness, and spectral standard deviation are not significantly different from H2N for the fricative /s/. Sampling frequency is likely to be one of the causes of such an effect, since the sampling frequencies of H2N, IP6 and IP8 were all above 40,000 Hz, although OPS also had a similarly high sampling frequency.

The recording environment seems to have little effect on spectral moments. The difference between the conference room and the lab is only 40 Hz for /s f/, which is not significant ($p = 0.75$), and the inclusion of the interaction between device and environment only improves the model marginally ($\chi^2(6) = 14.97, p = 0.02$).

3.5. Voice quality measures

For H1*-H2*, the model containing the interaction between device and tone fails to converge. Therefore, the interaction is not included in the final model. While T1, T2, and T4 are all modal, T3 is creaky. This is confirmed by the model. Only T3 differs significantly from T1 in H1*-H2* ($\hat{\beta} = -3.97, t = -5.21, p < 0.01$). The negative H1*-H2* value is indicative of creaky voice. All devices but SSG ($\hat{\beta} = -0.13, t = -0.46, p = 0.64$) have significantly different H1*-H2* values from H2N. While H1*-H2* based on SSG is slightly lower than H2N, all other devices have higher H1*-H2* than H2N. This can be quite detrimental in distinguishing voice quality contrasts, given the difference between T1 and T3 is only -3.97 dB. For example, the difference between T1 and T3 is not significant ($\hat{\beta} = -1.01, t = -1.11, p = 0.27$) for IP6.

Although the devices have different performances in H1*-H2*, they seem to perform equally well in jitter. A model of jitter including the interaction between device and tone reveals significant difference of OPS from H2N in T1 ($\hat{\beta} = -0.15, t = -1.97, p = 0.048$), and the interaction of both IP8 and SSG with T3 ($\hat{\beta} = -0.33, t = -2.98, p = 0.002$ for IP8, $\hat{\beta} = 0.28, t = 2.54, p = 0.011$ for SSG). For shimmer, ZMC is most different from H2N for all tones. Jitter and shimmer is crucial in characterizing T3 as it is creaky voiced. Models based on the subset of T3 showed similar results. Only IP8 is significantly different from H2N in jitter ($\hat{\beta} = -0.35, t = -2.53, p = 0.011$), and ZMC in shimmer ($\hat{\beta} = -1.38, t = -3.77, p < 0.01$).

The influence of recording environment is also significant on voice quality measures. Jitter of the recordings made in the conference room is greater than in the lab by 0.23 ($p < 0.01$), which exceeds the influence of any device. This is also the case for shimmer ($\hat{\beta} = 0.46, t = 7.95, p < 0.01$). But for H1*-H2*, the influence of environment is smaller than the devices ($\hat{\beta} = 0.44, t = 2.97, p < 0.01$).

4. Discussion and Conclusions

This study compared common acoustic measurements using recordings made by different devices in a sound-attenuated lab and a quiet conference room. It differs from the previous studies in that we focus on the performance of the devices in capturing linguistically meaningful information in the acoustic signal, instead of merely comparing the measurements regardless of its linguistic implications. A variegated pattern is revealed by considering different measurements.

As has been confirmed by other studies, F0 is relatively reliable among different devices. The difference reported in this study is significant yet small, with OPS having the most different F0 value from H2N. Different devices came up with various formant values, in particular for F2. Although we did not measure formants higher than F3, the devices gave more different values in F2 than F3. What is also unexpected is that F2 of /i/ rather than /u/ was more influenced. The illustration of vowel space using /i u a/ indicates that the choice of device in recording is critical in the study of vowels. Some of the smartphones can give rise to misleading results, as is also the case for ZOOM recordings. OPS again has a poor performance.

Sampling frequency is perhaps the most important factor influencing spectral moments. For devices with sampling frequencies lower than 40,000 Hz, they failed to capture the energy concentration at high frequencies, which are important for some fricatives. Other factors may be also at play. Although the sampling frequency of OPS is the same as IP6 and IP8, the spectral moments of OPS group with other devices. As OPS differs from the two iPhones in that the recordings were originally made in `aac` format, it is likely to be introduced by different data compression and conversion algorithms. Further analysis is needed to see how comparable the spectral moments are if the recordings were all down-sampled to the same frequency. The devices have critical effects on voice quality measures too, especially for H1*-H2*. A point worth noting is that the values of H1*-H2* caused by the creaky voice of T3 is not captured by IP6. But for jitter and shimmer, these devices are more reliable.

Although there is a notable difference in duration for recordings made on the cloud using ZOOM (ZMC), it does not stand out in other acoustic measurements, of which values are neither as good as iPhones in formants and spectral moments, nor as bad as OPS in most of the measurements. This does not validate the use of the cloud server of ZOOM, though. As mentioned in the Introduction, [9] reported differences about 100 ms for ZOOM recordings. Our results (~ 19.83 ms) are better than theirs. But detailed inspection of the waveform and spectrogram suggests that the recordings made on the cloud can be misleading, even erroneous. A better solution is to record locally using the speaker's computer.

The present study, together with previous studies, by no means exhaust all the possible methods of collecting speech data remotely. The speech can also be recorded via the built-in recorder in the speaker's computer, or using Praat with online assistance of the researcher via ZOOM. Smartphones can be another alternative when computers are unavailable. But care should be taken in processing the collected data, and steps taken should be documented as clearly as possible. These recordings collected in a simulated remote setting can be useful in studying prosody, where F0 and duration are primarily concerned, but they can be problematic in studying segmental and voice quality contrasts. Formant tracking tends to be inaccurate for F2, which will lead to misinterpretation of the vowel system. Fricatives of energy concentration at high frequencies are also influenced, perhaps due to the low sampling frequency. The issue cannot be settled for ZOOM recordings, since the user have no control over sampling frequency. The devices gave similar values in jitter and shimmer, but differ a lot in H1*-H2*, thus can undermine voice quality contrasts. A last note is that the format of the recording also matters. OPS with the `aac` format performs worst in nearly all measurements.

5. Acknowledgements

This study was supported by the RGC General Research Fund 2019/20 Project Reference 14607619 awarded to the third author and the CUHK Research Fellowship Scheme 2019-20 awarded to the first author.

6. References

[1] S. Bird, "Designing mobile applications for endangered languages," in *Oxford Handbook of Endangered Languages*, K. L. Rehg and L. Campbell, Eds. Oxford University Press, Sep. 2018, pp. 842–861.

[2] ZOOM, "Zoom," Zoom Video Communications Inc, 2021.

[3] N. H. Hilton and A. Leemann, "Editorial: Using smartphones to collect linguistic data," *Linguistics Vanguard*, vol. 7, no. s1, Jan. 2021.

[4] P. De Decker and J. Nycz, "For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?" *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, pp. 51–59, Jan. 2011.

[5] E. U. Grillo, J. N. Brosious, S. L. Sorrell, and S. Anand, "Influence of Smartphones and Software on Acoustic Voice Measures," *International Journal of Telerehabilitation*, vol. 8, no. 2, pp. 9–14, 2016.

[6] Y. Guan and B. Li, "Usability And Practicality of Speech Recording by Mobile Phones for Phonetic Analysis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Jan. 2021, pp. 1–5.

[7] S. Jannetts, F. Schaeffler, J. Beck, and S. Cowen, "Assessing voice health using smartphones: Bias and random error of acoustic voice parameters captured by different smartphone types," *International Journal of Language & Communication Disorders*, vol. 54, no. 2, pp. 292–305, 2019.

[8] A. Leemann, P. Jeszenszky, C. Steiner, M. Studerus, and J. Messerli, "Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing," *Linguistics Vanguard*, vol. 6, no. s3, Sep. 2020.

[9] C. Zhang, K. Jepson, G. Lohfink, and A. Arvaniti, "Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2717–2717, Oct. 2020.

[10] ——, "Comparing acoustic analyses of speech data collected remotely," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 3910–3916, Jun. 2021.

[11] I. Siegert and O. Niebuhr, "Case Report: Women, Be Aware that Your Vocal Charisma can Dwindle in Remote Meetings," *Frontiers in Communication*, vol. 5, 2021.

[12] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proceedings of ICPhS XVII*, Hong Kong, China, 2011, pp. 1846–1849.

[13] C. DiCanio, "Spectral moments of fricative spectra script in Praat," Haskins Laboratories & SUNY Buffalo, 2017.

[14] P. Boersma, "Praat: Doing phonetics with computers," 2020.

[15] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2021.

[16] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, vol. 82, no. 13, 2017.

[17] R. V. Lenth, *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*, 2021.