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

A comparison of acoustic correlates of voice quality across different recording devices: a cautionary tale

Joshua Penney, Andy Gibson, Felicity Cox, Michael Proctor, Anita Szakay

Centre for Language Sciences, Department of Linguistics, Macquarie University, Australia
{joshua.penney, andy.gibson, felicity.cox, michael.proctor, anita.szakay}@mq.edu.au

Abstract

There has been a recent increase in speech research utilizing data recorded with participants' personal devices, particularly in light of the COVID-19 pandemic and restrictions on face-to-face interactions. This raises important questions about whether these recordings are comparable to those made in traditional lab-based settings. Some previous studies have compared the viability of recordings made with personal devices for the clinical evaluation of voice quality. However, these studies rely on simple statistical analyses and do not examine acoustic correlates of voice quality typically examined in the (socio-) phonetic literature (e.g. H1-H2). In this study, we compare recordings from a set of smartphones/laptops and a solid-state recorder to assess the reliability of a range of acoustic correlates of voice quality. The results show significant differences for many acoustic measures of voice quality across devices. Further exploratory analyses demonstrate that these differences are not simple offsets, but rather that their magnitude depends on the value of the measurement of interest. We therefore urge researchers to exercise caution when examining voice quality based on recordings made with participants' devices, particularly when interested in small effect sizes. We also call on the speech research community to investigate these issues more thoroughly.

Index Terms: voice quality, non-modal phonation, creaky voice, COVID-19, home recordings, smartphone recordings, online research

1. Introduction

As ownership of smartphones and personal laptop computers has become near universal and the technology used in such devices continues to advance, a growing number of researchers have begun to collect speech data remotely through the use of participants' own personal devices (e.g. [1, 2, 3, 4, 5]). This approach has a number of advantages: it may lead to increased participation which in turn reduces the potential for sampling bias [3]; it enables access to participants and populations that may otherwise be difficult to record due to geographical or social barriers, which may assist in the documentation of endangered or minority languages/dialects [5, 6, 7]; and from a clinical perspective, it may facilitate increased screening and early identification of treatable voice disorders [8].

The trend towards utilizing participants' personal devices to collect speech data was accelerated in 2020 by the COVID-19 pandemic, as researchers sought to manage data collection in the face of restrictions on travel and face-to-face contact. This has seen many projects shift to online/hybrid recording sessions in which speech data are captured by recordings on participants' personal devices, often while being supervised by a researcher via a video call [9, 10]. This approach raises several important issues for data analysis. First, as each participant is recorded on a separate device, this may have an effect on acoustic measurements examined across participants (e.g. [11]). Also, it is not yet well understood to what extent recordings made on personal devices may affect the quality of the speech signal and whether such recordings are comparable to recordings made with high-end devices that have traditionally been used in lab-based recordings. It is possible that the personal devices may vary in the way they capture and process data that could have an effect on the acoustic measures typically used in linguistic studies of speech (e.g. F0, F1, F2). In an early investigation into whether recordings made with personal devices were suitable for sociolinguistic analysis, [12] concluded that iPhones and MacBook Pro laptops were suitable for analyses of F1 and F2 (despite reporting some deviation from the reference device), but not of F3.

More recently, as a result of the increase in studies relying on virtual data collection due to COVID-19, a few studies have examined the reliability of recordings made simultaneously on high-end reference devices typically used in lab-based recordings and a range of personal devices [13, 14, 15]. [13] visually compared vowel space plots of F1 and F2 measurements and found that personal devices were suitable for comparison of formants, but that smartphones distorted formant measures in the area between 750-1500Hz. [14] compared recordings made with smartphones and Zoom video conferencing software against a reference device. They found no significant differences between the smartphone recordings and the reference device for F0, F1, F2, or F3, although they did find that recordings made with the Zoom software had an effect on F1 for low vowels, and on F3 for front vowels. [15] examined a number of measures commonly included in phonetic analyses on smartphones and laptop computers. They found some measures (e.g. vowel/consonant duration and intensity) appeared to exhibit a significant offset (in either a positive or negative direction) compared to the reference device, but they found no significant differences for F0 or F1. They did, however, find that F2 was significantly lower with a MacBook laptop, significantly higher with an iPad, and marginally higher with an iPhone. The authors suggest that these results may be driven by high/mid-high diphthongs whose trajectories were not well tracked.

In addition, there is a small body of literature that has explored the viability of recordings made with personal devices for clinical evaluation of voice quality (see [8] for a summary). These studies have focused primarily on measures such as F0, jitter, shimmer, and noise measures such as CPP and HNR. In general, these studies have found that recordings of the same utterance on different devices are highly correlated for the acoustic measures examined [16, 17], which is to be expected,

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-729

but the results also suggest that F0 is the only measure consistently found not to differ significantly between devices (note though that some studies also found no difference in either CPP, jitter, or shimmer) [8, 18, 19, 20]. Importantly, the majority of these studies relied on simple statistical analyses, and did not account for effects of multiple speakers. In addition, they did not analyze the acoustic correlates of voice quality/non-modal phonation that are typically examined in the (socio-)phonetic literature, e.g. measures of harmonic amplitude such as H1-H2, H2-H4 (e.g. [21, 22, 23]). To date, few studies utilizing recordings made on participants' devices have examined voice quality with such measures (though see [24]). Therefore, it is important to understand if, and how, the acoustic measures of voice quality are affected by choices of device type. In their study, [15] included an analysis of H1-H2 and found that this measure was significantly lower in Android devices and marginally higher in iPhones relative to a reference device. The authors speculate that this may be due to discrepancies in how smartphones capture lower/higher frequencies. It is worth noting that smartphone recordings in [15] were made in M4A formant, so data compression issues may have contributed to these discrepancies.

In this paper, we compare audio recordings captured by a set of commonly used personal devices with those captured using a standard high-end recorder to assess the reliability of acoustic correlates of voice quality.

2. Methods

2.1. Participants

Data were produced by four phonetically trained participants: two female, two male. All of the speakers were L1 speakers of English: one male and one female were speakers of Australian English; one male and one female were speakers of New Zealand English.

2.2. Data collection and processing

Data for this study were recorded in a sound treated room in the Department of Linguistics at Macquarie University. Participants were seated comfortably in front of a desk on which two laptop computers were positioned next to each other at a 45 degree angle and two smart phones were positioned in between the two laptop monitors on smartphone tripods, such that the microphones of all four devices were pointed towards the participant at an approximately equivalent height. The laptop computers were a MacBook Pro 2020 running macOS Catalina 10.15.7 and a Dell Mobile Precision 7530 running Windows 10 1809 LTSC. The smart phones were an iPhone SE (second generation) running iOS 13.4.1 and a Samsung Galaxy S8OS running Android 9. These devices were chosen as they are broadly representative of personal devices frequently used by participants in our studies. All recordings were made simultaneously in uncompressed WAV format. Recordings were made from the two laptop devices using an online speech recorder (https://mmig.github.io/speech-to-flac/) accessed through Google Chrome with 48 kHz sampling rate at 16-bit depth. Recordings were made from the iPhone device using the Rode Reporter speech recording application with 48 kHz sampling rate at 16-bit depth. Recordings were made from the Galaxy device using the Easy Voice Recorder speech recording application with 44.1 kHz sampling rate at 16-bit depth. In addition, each participant was recorded to a reference Zoom H6 recorder through a Rode HS2 headset microphone with 44.1kHz sampling rate at 16-bit depth.

Each participant produced each of the AusE/NZE monophthongs in a standard /hVd/ frame, a set of selected sentences, and 'The Boy who Cried Wolf' reading passage ([25, cf. [2]). Only the reading passage is included in the analyses below.

All files for each speaker were trimmed and processed with a Matlab [26] function to time-align all of a speaker's files, resample files to 44.1 kHz (if necessary), and to normalize intensity. The reference file for each speaker was then force-aligned through MAUS [27] and the segment boundaries in the resulting TextGrid files were hand corrected for all vowels. Figure 1 shows two waveforms recorded simultaneously of the vowel /æ/ produced by a female speaker. The upper panel is taken from the reference recording; the lower panel is taken from the MacBook recording. It can be seen that substantial differences between the waveforms are apparent, although F0 appears broadly similar.

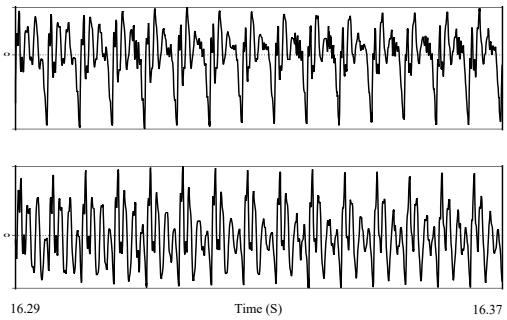

Figure 1: *Waveforms of the vowel /æ/ recorded simultaneously on a Zoom H6 reference device (upper panel) and on a MacBook laptop (lower panel).*

2.3. Acoustic measures

The data were processed using VoiceSauce [28] to extract the following acoustic measurements: F0, F1, F2, H1, H2, H4, H2kHz, H5kHz, H1-H2, H2-H4, H4-H2kHz, H2kHz-H5kHz, CPP, HNR05, HNR15, HNR25, HNR35. F0 was estimated using the default Straight pitch tracker [29]. F1 and F2 measures were calculated with Praat [30] using standard settings for male and female speakers. Note that we extracted measures of harmonic amplitudes that were not corrected for the effect of formants [31] (e.g. H1 rather than H1*) as the correction algorithm requires accurate formant estimation and it is not clear what effect device type has on formant frequency measures. In addition, as we compare across multiple recordings of the same utterance, the quality of the vowels is matched across recordings.

All measures were extracted for each vowel segment, then averaged over three equal subsections of the vowel. In the analyses below, we report average values taken from the middle subsection of each vowel, to reduce any influence of neighbouring segments.

2.4. Statistical Analyses

Sections 3 and 4 below present the results of a series of linear mixed effect models fitted with the lme4 package [32] in R [33], designed to examine whether the acoustic measures of interest differed between the reference device and the other devices.

3. Analysis 1

In this first analysis, we fitted separate models for each measure of interest. In each model the measure (e.g. H1) was the dependent variable and the device (Zoom H6, iPhone, MacBook, Dell, Galaxy) was included as a fixed factor, with the Zoom H6 as the reference. Random intercepts were included for Speaker and for Vowel nested within Speaker.

Table 1 contains a summary of results comparing each device to the reference device for each selected acoustic measure. As can be seen, F0 is the only measure where none of the devices differed significantly from the reference device, which is in line with previous findings [8, 18, 19, 20] (though see comments in section 4 below). Additionally, H2-H4 did not differ from the reference device for any of the devices with the exception of the Dell laptop. H2 and H4 did not show a difference for either of the smartphones, but did differ for both of the laptops. In many cases, it appears that all of the devices show an offset in the same direction; e.g. for all devices H1 is higher than the reference, and for CPP all devices are lower than the reference. However, in other cases (e.g. H1-H2) there is less consistency between devices. While some of the effects are small and unlikely to be audible, others are much greater (e.g. model estimates for H1-H2 differences ranged from 0.79 to 1.66 dB, but for H2K ranged from 2.62 to 7.79 dB). Additionally, across a large corpus even small differences could potentially accrue to cause meaningful biases in an analysis of these measures.

Table 1: *Results of lmer models for a range of acoustic measures comparing four devices to the Zoom H6 reference recorder. ***=p<.001; **=p<.01; *=p<.05; NS=not significant. Symbols in parentheses show if direction from baseline is positive (+) or negative (-).*

Measure	iPhone	Mac	Dell	Galaxy
F0	NS	NS	NS	NS
F1	** (-)	* (-)	*** (+)	** (-)
F2	*** (+)	*** (+)	* (-)	*** (+)
H1	*** (+)	*** (+)	*** (+)	*** (+)
H2	NS	*** (+)	*** (+)	NS
H4	NS	*** (+)	*** (+)	NS
H2kHz	*** (+)	*** (+)	*** (+)	NS
H5kHz	*** (+)	*** (+)	*** (+)	*** (+)
H1-H2	NS	*** (-)	*** (-)	** (+)
H2-H4	NS	NS	*** (-)	NS
H4-H2kHz	*** (-)	*** (-)	NS	NS
H2kHz-H5kHz	*** (-)	*** (+)	*** (+)	*** (-)
CPP	*** (-)	*** (-)	*** (-)	*** (-)
HNR05	*** (-)	*** (-)	*** (-)	*** (-)
HNR15	*** (-)	*** (-)	*** (-)	*** (-)
HNR25	*** (-)	*** (-)	*** (-)	*** (-)
HNR35	*** (-)	*** (-)	*** (-)	*** (-)

4. Analysis 2

Further inspection of the raw data suggested that the differences found in the models above may not be simple offsets. For example, while the model above found a higher measurement of H1 for the iPhone compared to the reference device, we observed that this difference appeared to be driven by the higher amplitude values, whereas for lower amplitude values the two devices were more closely aligned. Previous studies have not examined whether inter-device differences are sensitive to the value of the measurements themselves. However, many forms of audio processing respond dynamically to the signal (e.g. amplitude compressors reduce the amplitude for all signals above a certain threshold, with the degree of reduction increasing as the signal level increases; multi-band compressors work similarly, but apply different thresholds and rates of reduction in multiple frequency bands). While a detailed examination and comparison of dynamic signal processing used by the different devices is beyond the scope of this paper, it is reasonable to assume that this is likely to be at work in (some of) these devices [34]. The different devices may also utilize microphones that differ in terms of their frequency response and dynamic range [35].

Therefore, in this section we report on a series of exploratory linear mixed effect models, designed to investigate whether the differences between devices are consistent across the range of each measure, or whether between-device differences involve more dynamic relationships (i.e. whether they occur more strongly in certain ranges). Due to limitations of space we report only a subset of the results for these analyses, for the measures of H1, H2, H1-H2, and H2-H4. Note that we do not report models for F0 in these analyses as the residual variance in these models was unequal, with much greater residuals for data points at the lower end of the F0 range. This in itself is a concerning finding that warrants further study, and suggests that caution should be applied when examining F0 across different devices.

First, we calculated the difference between each device and the reference device for each measurement (e.g. the difference in H1 for iPhone is calculated as the iPhone H1 value minus the reference H1 value). We refer to this difference value as *Error*, which we included as the dependent variable in each model. The value of each measurement as recorded by the reference device was centred and included as fixed factor in the models. We refer to this factor as *Amplitude*, as we focus here on measures of harmonic amplitude. A separate model was fitted for each device per measure (e.g. for H1 a separate model was fitted for iPhone, MacBook, Dell, and Galaxy respectively). As above, we also included random intercepts for Speaker and for Vowel nested within Speaker. If differences between devices were consistent across the range of the measurement, we would expect the intercept in these models to be significant, and no significant differences to be present for *Amplitude*. For example, for H1 for iPhone, which in section 3 was found to be significantly higher than the reference device, these models should show a significant positive intercept and no effect for *Amplitude* if the differences were consistently higher across the range.

The intercept was not significantly different to zero in any of the models (i.e. there is no simple offset for any of the devices for these measures). Table 2 summarizes the results for the effect of *Amplitude* on *Error* for each of the models. As can be seen, the amplitude of the measure significantly predicts the magnitude of the difference between the two devices in many of the models. That is, in many cases, what appeared to be a simple offset compared to the reference device in section 3 is now non-significant, with the variance in the data being better explained by signal-dependent slopes. This suggests that many of the differences between devices cannot be considered to be simple offsets that are either higher or lower than the reference

device. Rather, these differences may be greater or smaller according to the value of the measurement of interest.

In addition, some measurements that were not found to significantly differ from the reference device in section 3 were found to have a significant effect of *Amplitude* in these models. This is illustrated in Figure 2 using the example of H2 in the Galaxy device (note that values on the x-axis in this figure are raw values, not centered values as were included in the model). At low amplitudes of H2, the difference between the Galaxy and the reference is positive (i.e. the Galaxy is higher than the reference). As amplitude increases, the values on the plot appear closer the zero line (i.e. there is less difference between the devices). At greater amplitudes, the difference between the devices becomes negative (i.e. the Galaxy is lower than the reference). The results of the first set of models (in section 3) suggested that the Galaxy was equivalent to the reference device, as there was no significant difference found. This might have led to the conclusion that this device is trustworthy for this measure. However, the second set of models (in this section) shows that there is a systematic bias in measuring H2 with this device, which researchers studying voice quality measures across different devices may need to be aware of.

Table 2: *Significant effects of Amplitude on Error for lmer models according to device and acoustic measure. ***=p<.001; **=p<.01; *=p<.05; NS=not significant. Symbols in parentheses show if the direction of the effect for Amplitude is positive (+) or negative (-).*

Measure	iPhone	Mac	Dell	Galaxy
H1	*** (-)	*** (+)	* (+)	* (-)
H2	*** (-)	NS	NS	*** (-)
H1-H2	*** (-)	*** (-)	NS	NS
H2-H4	*** (-)	*** (+)	NS	*** (-)

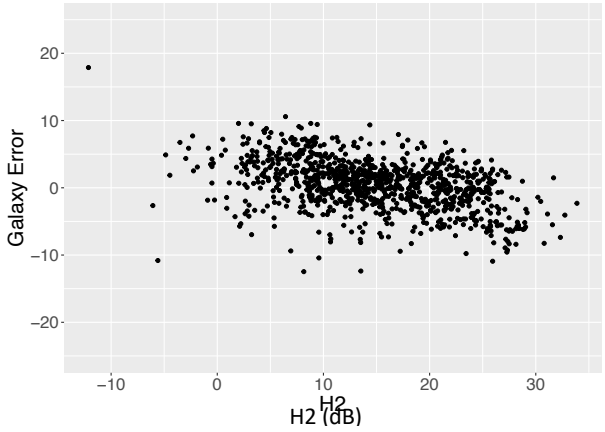

Figure 2: *Error in Galaxy recordings of H2 according to Amplitude of H2 from reference device.*

5. Discussion

The results of the analyses presented above suggest that researchers should exercise caution when interpreting acoustic measures of voice quality recorded with participants' personal devices. Analysis 1 showed that significant differences exist for a range of acoustic measures commonly used in the analysis of voice quality when compared to a reference device. Analysis 2 demonstrated that these differences should not be thought of as simple offsets, i.e. it is not the case that a particular value recorded with a particular device will be higher or lower than a reference device and that this effect will be consistent. Rather, such differences may be greater or smaller according to the value of the measurement of interest. This entails that it will be difficult to account for such differences by means of a simple normalization procedure blanketly applied to all data or by including an offset value in statistical models.

While the results presented here are sobering for researchers interested in examining voice quality in data collected with participants' personal devices, it is not our aim to suggest that such procedures be abandoned. Indeed, it is our belief that the advantages offered to speech research by the availability and ubiquity of smartphone/laptop ownership largely outweigh the disadvantages, and at any rate it is probable that more researchers will make use of such methods in the future, particularly in the climate of an ongoing pandemic. Rather, our aim is to raise awareness of potential issues involved in such an approach, and to stimulate discussion of best practices to mitigate such problems going forward. We hope that this study serves as a cautionary tale and a call to explore these issues more thoroughly. There are, of course, a number of limitations to this study. For example, the analyses are restricted to a small set of speakers recorded in laboratory conditions. There is also a wide range of potential devices that participants may use, and this analysis examined only a few of these, and only one exemplar of each device type. Other devices, or other exemplars of the same device type, may yield different results. In addition, we observed some individual speaker differences in the direction of the variance for some of the measures; we therefore aim to explore speaker specific error patterns in our future analyses.

Finally, although the findings detailed above suggest that caution is needed when investigating voice quality using recordings made with participants' personal devices, it should be borne in mind that these issues may not be relevant for all studies relating to voice quality and will depend on the magnitude of the effects under investigation (cf. [15]). As an example, we manually labelled all occurrences of creaky voice in the files examined above, and found that both F0 and H1-H2 clearly differentiated creaky items from non-creaky items for all speakers across all of the devices. This demonstrates that data collected from participants' devices may be less affected by the issues raised in this paper for certain measures such as a binary classification of creak/non-creak described above that is determined by a relatively large within-speaker difference in the acoustic measures. It is also possible that automatic methods of identifying creaky voice, such as [36], which make use of acoustic cues to classify speech as either creaky or not creaky, would be robust with such data. This, however, remains to be assessed. For more detailed or fine-grained differences or for measures averaged across multiple speakers, we recommend caution and careful selection of devices until strategies for managing the issues raised above are developed.

6. Acknowledgements

We thank Marcus Ockenden for technical assistance, Serje Robidoux for statistical advice, and Louise Ratko and Hannah White for supplying their voices. This research was supported by Australian Research Council Grant DP190102164 and Australian Research Council Future Fellowship Grant FT180100462 to the third author.

7. References

[1] A. Leemann, M.-J. Kolly, J.-P. Goldman, V. Dellwo, I. Hove, I. Almajai, and D. Wanitsch, "Voice Äpp: a mobile app for crowdsourcing Swiss German dialect data," in *Proceedings of INTERSPEECH 2015*, Dresden, Sep. 2015, pp. 2804–2808.

[2] A. Leemann, M.-J. Kolly, and D. Britain, "The English Dialects App: The creation of a crowdsourced dialect corpus," *Ampersand*, vol. 5, pp. 1–17, 2018.

[3] A. Leemann, "Apps for capturing language variation and change in German-speaking Europe: Opportunities, challenges, findings, and future directions," *Linguistics Vanguard*, vol. 7, no. s1, pp. 1–12, Jan. 2021.

[4] N. Entringer, P. Gilles, S. Martin, and Christoph Purschke. "Schnëssen: Surveying language dynamics in Luxembourgish with a mobile research app," *Linguistics Vanguard*, vol. 7, no. s1, pp. 1–15, Jan. 2021.

[5] N. H. Hilton, "Stimmen: A citizen science approach to minority language sociolinguistics," *Linguistics Vanguard*, vol. 7, no. s1, pp. 1–15, Jan. 2021.

[6] S. Bird, F. R. Hanke, O. Adams, and H. Lee, "Aikuma: A mobile app for collaborative language documentation," in *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages,* June 2014, pp. 1–5.

[7] S. Bird, "Designing Mobile Applications for Endangered Languages," in *The Oxford Handbook of Endangered Languages*, K. L. Rehg and L. Campbell, Eds. New York: Oxford University Press, 2018.

[8] 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, Mar./Apr. 2019.

[9] A. Gibson, F. Cox, L. Buckley, and J. Penney, "Child Speech, Community Diversity: Baseline Data for Pre-nasal TRAP-raising." Paper presented at the *51st Annual Conference of the Australian Linguistic Society,* Dec. 2020.

[10] 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, pp 1–16, Sep. 2020.

[11] T. Rathcke, J. Stuart-Smith, B. Torsney, and J. Harrington, "The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetic real-time studies," *Speech Communication*, vol. 86, pp. 24–41, 2017.

[12] P. De Decker and J. Nycz, "For the record: Which digital media can be used for socio-phonetic analysis?" *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, pp. 51–59, 2011.

[13] V. Freeman, P. DeDecker, and M. Landers, "Suitability of self-recordings and video calls: Vowel formants and nasal spectra. *The Journal of the Acoustical Society of America,* vol. 148, p. 2714, 2020.

[14] 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, p. 2717, 2020.

[15] C. Sanker, S. Babinski, R. Burns, M. Evans, J. Kim, S. Smith, N. Weber, and C. Bowern, "(Don't) try this at home! The effects of recording devices and software on phonetic analysis," lingbuzz/005748, 2021.

[16] C. Manfredi, J. Lebacq, G. Cantarella, J. Schoentgen, S. Orlandi, A. Bandini, and P. H. DeJonckere, "Smartphones Offer New Opportunities in Clinical Voice Research," *Journal of Voice*, vol. 31, no. 1, pp. 111.e1–111.e7, 2016.

[17] T. Kojima, S. Fujimura, R. Hori, Y. Okanoue, K. Shoji, and M. Inoue, "An Innovative Voice Analyzer "VA" Smart Phone Program for Quantitative Analysis of Voice Quality," *Journal of Voice*, vol. 33, no. 5, pp. 642–648, 2019.

[18] A. P. Vogel, K. M. Rosen, A. T. Morgan, and S. Reilly, "Comparability of modern recording devices for speech analysis: smartphone, landline, laptop, and hard disc recorder," *Folia Phoniatrica et Logopaedica*, vol. 66, pp. 244–250, 2014.

[19] V. Uloza, E. Padervinskis, A. Vegiene, R. Pribuisiene, V. Saferis, E. Vaiciukynas, A. Gelzinis, and A. Verikas, "Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening," *European Archives of Oto-Rhino-Laryngology*, vol. 272, pp. 3391–3399, 2015.

[20] 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, pp. 9–14, 2016.

[21] P. Callier and R. J. Podesva, "Multiple Realizations of Creaky Voice Evidence for Phonetic and Sociolinguistic Change in Phonation." Paper presented at *New Ways of Analyzing Variation (NWAV) 44*, Nov. 2015.

[22] M. Garellek, "The phonetics of voice," in *The Routledge Handbook of Phonetics*, W. F. Katz and P. F. Assmann, Eds. pp. 75–106, 2019.

[23] M. Garellek and S. Seyfarth, "Acoustic differences between English / t/ glottalization and phrasal creak," in *Proceedings of INTERSPEECH 2016*, San Francisco, Sep. 2016, pp. 1136–1140.

[24] B. Gittelson, A. Leemann, and F. Tomaschek, "Using Crowd-Sourced Speech Data to Study Socially Constrained Variation in Nonmodal Phonation," *Frontiers in Artificial Intelligence*, vol. 3, pp. 1–9, 2021.

[25] D. Deterding, "The North Wind versus a Wolf: Short texts for the description and measurement of English pronunciation, Journal of the International Phonetic Association, vol. 36, pp. 187–196, 2006.

[26] Mathworks, "MATLAB version *R2020a,"* https://www.mathworks.com/products/matlab.html

[27] T. Kisler, U. D. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer. Speech and Language,* vol. 45, pp. 326–347, 2017.

[28] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proceedings of the International Congress of Phonetic Sciences*, Hong Kong, Aug. 2011, pp. 1846–1849.

[29] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[30] P. Boersma and D. Weenink, "*Praat: Doing phonetics by computer*," version 6.1.16, 2020.

[31] M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex, and vowel dependencies of acoustic measures related to the voice source," Journal of the Acoustical Society of America, vol. 121, pp. 2283–2295, 2007.

[32] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[33] R Core Team, "*R: A Language and Environment for Statistical Computing*," version 4.0.2, 2020.

[34] B. M. Faber, "Acoustical Measurements with Smartphones: Possibilities and Limitations," *Acoustics Today*, vol. 13, no. 2, pp. 10–17, 2017.

[35] J. G. Švec and S. Granqvist, "Guidelines for Selecting Microphones for Human Voice Production Research," *American Journal of Speech-Language Pathology*, vol. 19, pp. 356–368, Nov. 2010.

[36] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech and Language*, vol. 28, no. 5, pp. 1233–1253, Sep. 2014.