Accepted draft 01/07/2019

Reliability of clinical voice parameters captured with smartphones – measurements of added noise and spectral tilt

Felix Schaeffler[1,2], *Stephen Jannetts*[1], *Janet Beck*[1,2]

[1]Clinical Audiology, Speech and Language (CASL) Research Centre, Queen Margaret University Edinburgh, Scotland, UK

[2]Fitvoice CIC, Business Innovation Zone, University Drive, Musselburgh, Scotland, UK

fschaeffler@qmu.ac.uk, sjannetts@qmu.ac.uk, jbeck@qmu.ac.uk

Abstract

Smartphones have become powerful tools for data capture due to their computational power, internet connectivity, high quality sensors and user-friendly interfaces. This also makes them attractive for the recording of voice data that can be analysed for clinical or other voice health purposes. This however requires detailed assessment of the reliability of voice parameters extracted from smartphone recordings. In a previous study we analysed reliability of measures of periodicity and periodicity deviation, with very mixed results across parameters. In the present study we extended this analysis to measures of added noise and spectral tilt. We analysed systematic and random error for six frequently used acoustic parameters in clinical acoustic voice quality analysis. 22 speakers recorded sustained [a] and a short passage with a studio microphone and four popular smartphones simultaneously. Acoustic parameters were extracted with Praat and smartphone recordings were compared to the studio microphone. Results indicate a small systematic error for almost all parameters and smartphones. Random errors differed substantially between parameters. Our results suggest that extraction of acoustic voice parameters with mobile phones is not without problems and different parameters show substantial differences in reliability. Careful individual assessment of parameters is therefore recommended before use in practice.

Index Terms: acoustic voice parameters, reliability, smartphone, clinical voice analysis

1. Introduction

Smartphones are increasingly used to collect speech data for acoustic analysis, and several authors have suggested that smartphones have sufficient reliability for voice analysis with a clinical purpose [1]–[6]. [7] took a less optimistic stance on this and argued that previous analyses partly rested on false assumptions about the nature of the problem. For example, some studies have used non-significance between a smartphone recording and a reference recording as evidence that smartphone measurement are reliable. [7] have argued that this approach can only ever assess systematic error and that the random error of a measurement is completely ignored with this approach, or even worse, a high random error supports a non-significant result and thus a 'positive' conclusion. If both systematic and random error are quantified, then some acoustic parameters show worrying levels of random error.

[7] investigated four acoustic parameters with high relevance for clinical voice analysis, mean F0, smoothed Cepstral Peak Prominence (CPPS), jitter (RAP) and shimmer %, across four popular smartphone devices (two Samsung and two Apple models) and for two different types of voice material, i.e. sustained vowels and the reading of a passage. The smartphone recordings were compared to simultaneous recordings with a Neumann U89i studio microphone.

F0 measurements were generally judged to be sufficiently reliable. All phones showed a systematic error of less than 2 Hz and a random error of ± 5Hz for passage data, which was deemed acceptable for most practical purposes.

CPPS also showed a small bias across all phones, never exceeding 1 dB. The random error was also below ±1 dB, but assessing the relevance of this amount of random error is more difficult than for F0. [7] suggested relating the random error to the range of values observed in a sample and to clinical thresholds, if available. In the study sample, the random error of CPPS corresponded to about 10% of the total range of values. This was preliminarily accepted as sufficiently reliable, especially as data from other studies suggests that the range of CPPS values in the population is actually much wider.

Both Jitter (RAP) and Shimmer % showed large random errors across recording devices and were therefore not deemed reliable enough for practical purposes.

The study by [7] therefore suggests that smartphone reliability for acoustic voice assessment very much depends on the parameter under question. The study presented here is a continuation study of [7], extending the analysis to the quantification of added noise and glottal pulse efficiency as reflected in the spectral tilt of the voice spectrum.

We analysed five acoustic parameters with high relevance in studies of clinical voice assessment: 'Harmonics-to-noise ratio' is a well-established measure of the level of noise added at the glottis. There are various implementations of this measure. We used the standard implementation in Praat. Glottal noise excitation ratio (GNE) attempts to measure added noise independent of shimmer- and jitter-type deviations from periodicity [8][9].

The second aspect of glottal characteristics considered here is the effectiveness of the glottal pulse in generating acoustic energy. Various parameters quantify the decrease of energy over frequency (often referred to as spectral tilt), with steeper slopes indicating less overall effectiveness of the glottal pulse.

Different measures of spectral tilt vary in the frequency range they take into account. Some measures, especially those

related to estimates of vocal fry, mainly consider the energy difference between the first two harmonics (H1 and H2). In this paper we used both an uncorrected estimate of H1 minus H2 (H1-H2) as well as a version that aims at correcting for vocal tract influence (H1*-H2*) [10]. Simpler spectral slope measures consider energy differences between frequency regions, for example the 'tilt' and 'slope' measures suggested in [11], which calculate spectral slope by comparing the energy in two frequency bands, either directly or from a regression line.

2. Method

2.1. Speakers and procedure

Recording procedures followed the procedures described in [7] and used the same data. We recorded 12 women and 10 men in a sound-treated recording studio. All participants were recorded simultaneously with five devices, four smartphones and one studio microphone. The smartphones used were a Samsung Galaxy S8+ (SG8), an iPhone 6s (i6s) and an iPhone 7 (ip7).

The smartphones were arranged in a semi-circular array directly below the studio microphone to ensure comparable microphone-to-mouth distance (~20 cm). The position of smartphones was systematically changed between participants, so that each smartphone took one of the five positions (including the central position) in turn.

Participants produced sustained [a] vowel sounds and read a shortened version of the phonetically balanced passage 'The Dog and Duck Story' [12]. Prompts were displayed on the phone screens, using the smartphone app 'Fitvoice' [13]. Reliability of passage data was generally higher in our previous study, therefore only passage data is presented here.

2.2. Acoustic analysis

Acoustic analysis was performed with Praat [14]. Voiced segments were extracted following the procedure described in [11]. 'Slope' and 'tilt' were extracted using the procedures from the same source. GNE was extracted with the Praat 'To Harmonicity (gne):' command, with minimum frequency set at 500 Hz, maximum frequency set at 5000 Hz, bandwidth set at 3000 Hz and step at 300, following recommendations by [9]. H1-H2 and H1*-H2* were calculated following the procedure described in [10] and implemented in Praat.

2.3. Statistical analysis

As in [7], Bland-Altman analysis was performed using R 3.4.0 (R Core Team, 2017) with R package 'BlandAltmanLeh' [15]. This package calculates a 95% confidence interval for the systematic error. Systematic errors (bias) were deemed significant if the confidence interval did not include zero.

The random error was derived from the 'limits of agreement' of the difference between the two measures, i.e. ± 1.96 SD around the mean [16]. The absolute value of this range divided by two is referred to as the 'critical difference'. The critical difference describes the expected random error of any measured value of a certain parameter.

As described above, the assessment of a random error with respect to its practical relevance requires some meaningful quantity to relate it to. Here we relate the random error to the total range of values, as measured with the studio microphone. As a preliminary and somewhat ad-hoc guideline, we suggest that random errors that do not exceed 10% of the total range could be acceptable.

3. Results

Here we present systematic and random error for the six parameters (cf. Tables 1 to 6). Significant systematic errors are marked with an asterisk (*). In addition, we provide Bland-Altman (BA) plots for all phones and parameters. The Bland-Altman plots show males (red dots) and females (blue triangles) separately. This allows us to assess whether separate calculations for male and female values would be likely to change the random or systematic error calculations.

3.1. Slope

All phones, with the exception of the Samsung Galaxy S8, showed significant bias for spectral slope measures. All devices had critical differences that constituted between 13.5 and 16.9% of the range.

Table 1: *Systematic and random error for slope measures*

Phone	bias	Crit. Diff.	Range%
gs8	0.29 dB	2.27 dB	14.4
sj3	1.38* dB	2.68 dB	16.9
ip7	0.56* dB	2.21 dB	14.0
i6s	0.79* dB	2.13 dB	13.5

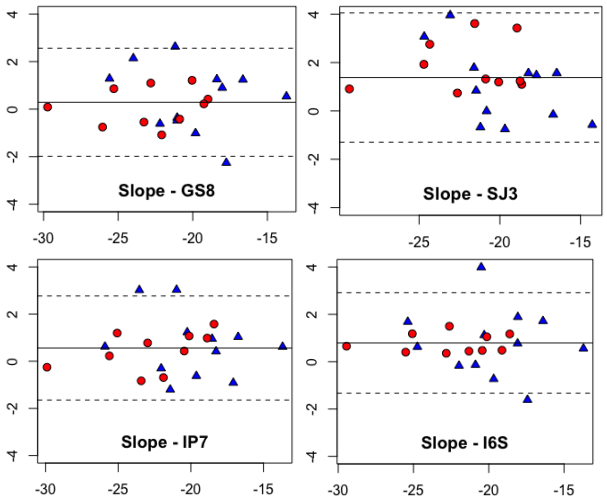


Figure 1: *BA plot for slope parameter (unit dB). Solid line: mean, dashed lines: limits of agreement (± 1.96SD). Blue triangles: female values; red dots: male values.*

3.2. Tilt

All devices showed significant negative bias in the range of 0.41-2.45 for spectral tilt measures. The critical difference for the Samsung devices was in the range of 47.4-51.3%. The critical difference for the iPhone devices was in the range of 21.6-27.8%.

Table 2: *Systematic and random error for tilt measures*

Phone	bias	Crit. Diff.	Range%
gs8	-1.23* dB	1.37 dB	47.4
sj3	-2.45* dB	1.49 dB	51.3
ip7	-0.55* dB	0.63 dB	21.6
i6s	-0.41* dB	0.80 dB	27.8

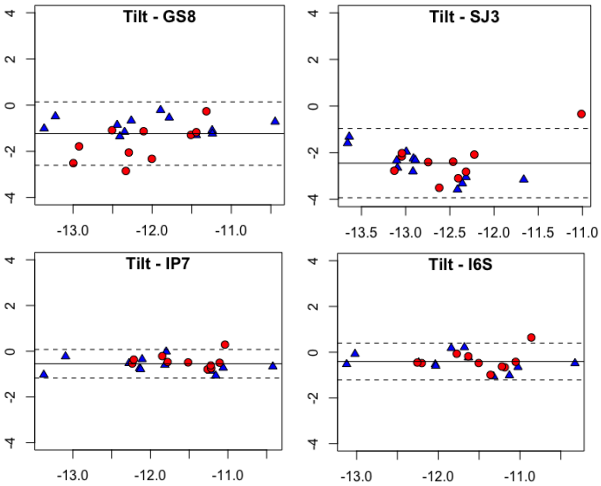


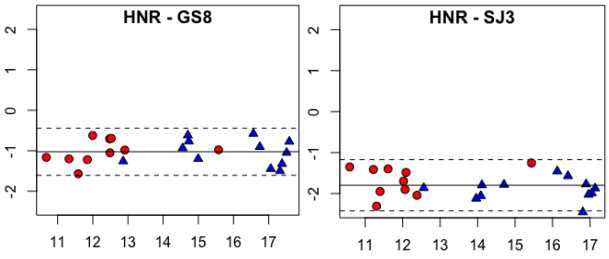
Figure 2: *BA plot for tilt parameter (unit dB). Solid line: mean, dashed lines: limits of agreement (± 1.96SD). Blue triangles: female values; red dots: male values.*

3.3. HNR

All devices showed significant negative bias in the range of 1.02-1.80 for HNR measures. Critical differences were in the range of 0.58-0.75 dB. The two Samsung devices had random error values <10% of the range, whereas the iPhone devices were between 10-11% of the range.

Table 3: *Systematic and random error for HNR measures*

Phone	bias	Crit. Diff.	Range%
gs8	-1.02*	0.58 dB	8.6
sj3	-1.80*	0.63 dB	9.2
ip7	-1.10*	0.71 dB	10.5
i6s	-1.22*	0.75 dB	11.0



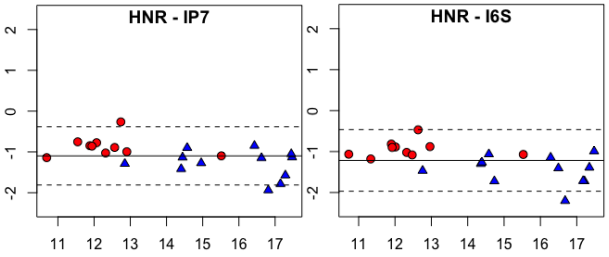


Figure 3: *BA plot for HNR parameter (unit dB). Solid line: mean, dashed lines: limits of agreement (± 1.96SD). Blue triangles: female values; red dots: male values.*

3.4. GNE

All devices showed significant negative bias in the range of 0.025-0.04 for GNE measures. The critical difference for all devices constituted between 13.9-20.8% of the range.

Table 4: *Systematic and random error for GNE measures*

Phone	bias	Crit. Diff.	Range%
gs8	-0.025*	0.032	16.7
sj3	-0.04*	0.040	20.8
ip7	-0.026*	0.030	15.7
i6s	-0.030*	0.027	13.9

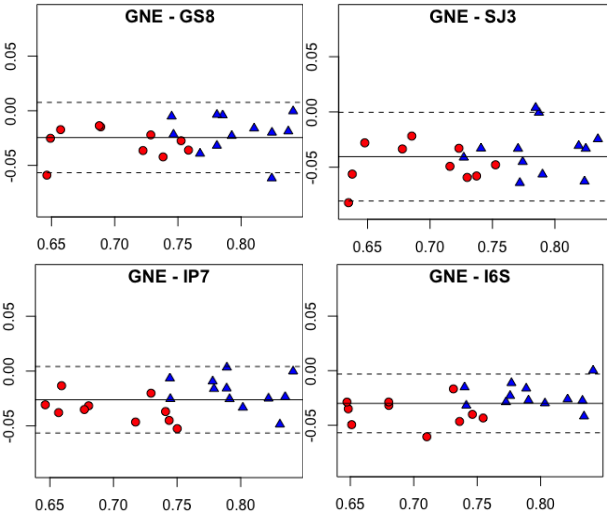


Figure 4: *BA plot for GNE parameter (no unit). Solid line: mean, dashed lines: limits of agreement (± 1.96SD). Blue triangles: female values; red dots: male values.*

3.5. H1-H2

All devices showed significant negative bias in the range of 0.53-5.09 dB for uncorrected H1-H2 measures. Critical differences were in the range of 22.1-36%.

Table 5: *Systematic and random error for H1-H2 measures*

Phone	bias	Crit. Diff.	Range%
gs8	-2.91* dB	2.17 dB	28.2
sj3	-5.09* dB	2.78 dB	36.0
ip7	-0.54* dB	2.10 dB	27.2
i6s	-0.53* dB	1.70 dB	22.1

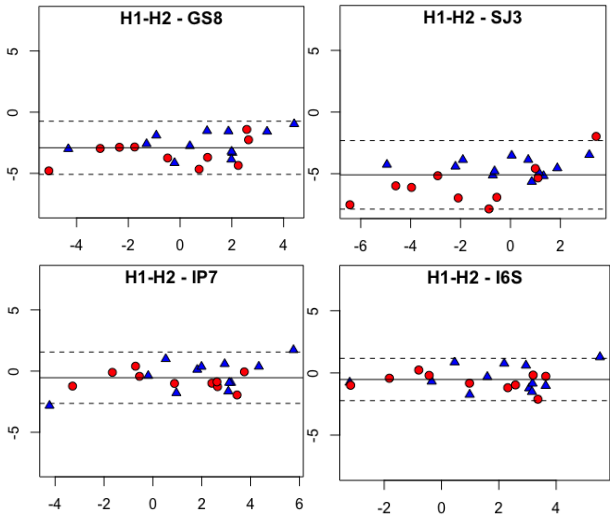


Figure 5: *BA plot for H1-H2 parameter (unit dB). Solid line: mean, dashed lines: limits of agreement (± 1.96SD). Blue triangles: female values; red dots: male values.*

3.6. H1*-H2*

All devices showed significant negative bias for H1*-H2* measures in the range of -0.51 to -4.12 dB. The critical differences were between 1.38 dB and 4.68 dB constituting between 16.2% and 54.5% of the reference range. The Samsung devices showed a higher critical difference than the Apple devices.

Table 6: *Systematic and random error for H1*-H2* measures*

Phone	bias	Crit. Diff.	Range%
gs8	-2.10* dB	3.29 dB	38.2
sj3	-4.12* dB	4.68 dB	54.5
ip7	-0.51* dB	1.86 dB	21.7
i6s	-0.51* dB	1.38 dB	16.1

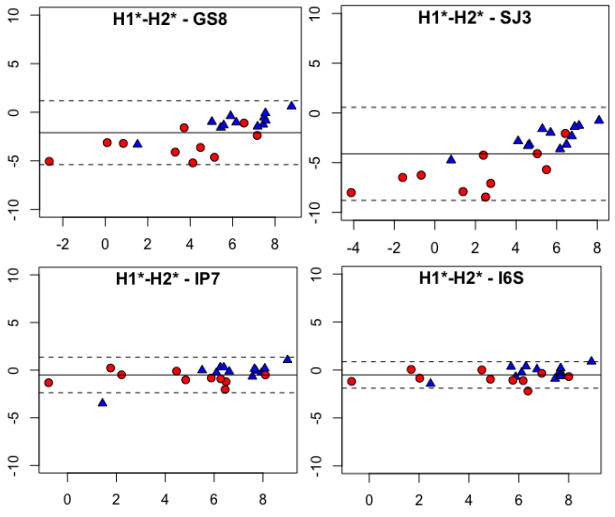


Figure 6: *BA plot for H1*-H2* parameter (unit dB). Solid line: mean, dashed lines: limits of agreement (± 1.96SD). Blue triangles: female values; red dots: male values.*

4. Discussion

Of the parameters tested here, only HNR fulfilled the 10% criterion. However, besides just considering the range of values observed in this sample it might be worth looking at ranges from other corpora. A study of acoustic parameters derived from the Kay-Pentax Disordered Voice Database (DVDB) [17] provided us with comparative values from a larger sample comprising healthy and disordered voices. For GNE, the range of values observed in this database is considerably larger (0.42 compared to 0.19 in the current sample). This results in a relative error below 10% for all phones.

For tilt, the parameter with the highest relative random error in the current study, the DVDB data has a range of 5.37 dB compared to 2.89 dB in the current sample. This would reduce the mean relative error across all phones to 20%, which is still substantial. As tilt is used in the popular Acoustic Voice Quality Index (AVQI) [18], this high random error warrants further investigation. The slope measure (also an AVQI component) shows a random error around 14%, and the range for the DVDB database is 21.73, reducing the mean relative random error to 10.7%.

Relative random errors for H1-H2 and H1*-H2* are also considerable. They reduce to a mean of 6.2% and 8.8% respectively if DVDB ranges are considered. However especially for H1*-H2* and the Samsung phones, the Bland-Altman plots suggest that the random error is not uniformly distributed across the range. This will require further exploration.

5. Conclusion

This study suggests that extraction of acoustic voice parameters with smartphones requires careful, parameter-specific consideration. While HNR, GNE and slope seem relatively unproblematic parameters, tilt has a very high random error, and both H1-H2 measures suggest issues with the distribution of the random error.

This is also a caveat against assuming that the random error will be uniform across a parameter's possible range. So far, we have analysed typical voices, therefore the extension to pathological voices is speculative and requires further study.

It is important to consider in which contexts the random and systematic errors are most relevant. For repeated recordings with the same device, the systematic error might be less important, especially if only within speaker variation is considered. As soon as comparisons to other devices are required, or a comparison to thresholds derived with studio equipment is planned, the systematic error should be considered, and the values presented here could provide a basis for calibration.

The random error, however, will also affect repeated recordings with the same device and should guide the interpretation of change. Value changes that fall within the range of the random error might not be reliable.

Future studies should investigate to what extent calibration can reduce systematic error and how far other field recording conditions (e.g. noise and room configuration) impact measurement. Furthermore, more general studies of typical and clinical ranges of acoustic parameters would support overall validity and reliability of clinical acoustic analysis.

6. References

[1] 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, no. 6, pp. 244–250, 2014.

[2] V. Uloza *et al.*, '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, no. 11, pp. 3391–3399, 2015.

[3] 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.

[4] C. Manfredi *et al.*, 'Smartphones Offer New Opportunities in Clinical Voice Research', *Journal of Voice*, vol. 31, no. 1, pp. 111.e1-111.e7, 2017.

[5] Y. Maryn, F. Ysenbaert, A. Zarowski, and R. Vanspauwen, 'Mobile Communication Devices, Ambient Noise, and Acoustic Voice Measures', *Journal of Voice*, vol. 31, no. 2, pp. 248.e11-248.e23, 2017.

[6] 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*, 2018.

[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*, 2019.

[8] D. Michaelis, T. Gramss, and H. W. Strube, 'Glottal-to-noise excitation ratio - a new measure for describing pathological voices', *Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[9] J. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, 'The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders', *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[10] M. Iseli and A. Alwan, 'An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation', in *ICASSP 04*, vol. 1, pp. 666–669, 2004.

[11] Y. Maryn and D. Weenink, 'Objective Dysphonia Measures in the Program Praat: Smoothed Cepstral Peak Prominence and Acoustic Voice Quality Index', *Journal of Voice*, vol. 29, no. 1, pp. 35–43, 2015.

[12] A. Brown and G. J. Docherty, 'Phonetic variation in dysarthric speech as a function of sampling task', *Eur J Disord Commun*, vol. 30, no. 1, pp. 17–35, 1995.

[13] M. Eichner, *fitvoice*, v1.1.0, [App], Available: Google Play & Apple App Store, Fitvoice CIC, 2019.

[14] P. Boersma and D. Weenink, *Praat: Doing Phonetics By Computer*. 2018.

[15] B. Lehnert, 'BlandAltmanLeh: plots (slightly extended) Bland-Altman plots', *R package version 0.1.0*, 2014.

[16] J. Martin Bland and D. Altman, 'Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement', *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[17] F. Schaeffler, M. Eichner, and J. Beck, 'Towards Ordinal Classification of Voice Quality Features with Acoustic Parameters', in *Proceedings of the 30th Conference on Electronic Speech Signal Processing (ESSV) 2019, Dresden, Germany.*, Dresden, Germany, 2019.

[18] Y. Maryn, M. Bodt, and N. Roy, 'The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders', *Journal of Communication Disorders*, vol. 43, no. 3, pp. 161–174, 2010.