Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types

Accepted (peer-reviewed) version

Journal:	*International Journal of Language & Communication Disorders*
Manuscript ID	TLCD-2018-0042.R4
Wiley - Manuscript type:	Special Issue
Keywords:	voice monitoring, voice care, occupational voice disorder, smartphone, acoustic analysis
Authors and affiliation:	Stephen Jannetts, Felix Schaeffler, Janet Beck, Steve Cowen, CASL Research Centre, Queen Margaret University Edinburgh, Scotland, UK Telephone: 0131 474 0000 Fax: 0131 474 0001 Email: sjannetts|fschaeffler|jbeck|scowen@qmu.ac.uk

ABSTRACT

BACKGROUND: Occupational voice problems constitute a serious public health issue with substantial financial and human consequences for society. Modern mobile technologies like smartphones have the potential to enhance approaches to prevention and management of voice problems. This paper addresses an important aspect of smartphone-assisted voice care: the reliability of smartphone-based acoustic analysis for voice health state monitoring.

AIM: To assess the reliability of acoustic parameter extraction for a range of commonly used smartphones by comparison with studio recording equipment.

METHODS AND PROCEDURES: Twenty-two vocally healthy speakers (12 female; 10 male) were recorded producing sustained vowels and connected speech under studio conditions using a high-quality studio microphone and an array of smartphones. For both types of utterances, Bland-Altman-Analysis was used to assess overall reliability for Mean F0; CPPS; Jitter (RAP) and Shimmer %.

OUTCOMES AND RESULTS: Analysis of the systematic and random error indicated significant bias for CPPS across both sustained vowels and passage reading. Analysis of the random error of the devices indicated that that mean F0 and CPPS showed acceptable random error size, while jitter and shimmer random error was judged as problematic.

CONCLUSIONS AND IMPLICATIONS: Confidence in the feasibility of smartphone-based voice assessment is increased by the experimental finding of high levels of reliability for some clinically relevant acoustic parameters, while the use of other parameters is discouraged. We also challenge the practice of using statistical tests (e.g. t-tests) for measurement reliability assessment.

What this paper adds:

1. What is already known on this subject: There is emerging evidence that smartphones can be used reliably for acoustic voice assessment, but quantification of measurement error is lacking and statistical testing is often used incorrectly for reliability assessment.
2. What this study adds: We present analysis of systematic and random errors for acoustic parameters derived from smartphone recordings to evaluate the reliability of acoustic voice assessment with smartphones.
3. Clinical implications of this study: The results of this paper extend the evidence base for acoustic health-related monitoring of voices and thus facilitate preventative intervention in populations at risk of voice problems.

1 BACKGROUND AND MOTIVATION

Occupational voice problems and disorders are a long-standing public health issue with substantial financial and human consequences for society (Hazards, 2004, Verdolini and Ramig, 2001). This paper sits in the context of ongoing efforts to use modern mobile technologies (e.g. smartphones) for better voice problem prevention and management. The success of such efforts relies on a number of prerequisites. One important prerequisite is that any technology used for this purpose must provide reliable data. Acoustic monitoring of voices constitutes a crucial, but also technologically ambitious aspect of our approach to voice care. This paper reports on one part of an ongoing experimental evaluation of the reliability of acoustic voice parameter extraction using smartphones.

Occupational voice disorders affect a substantial proportion of the population, and it is suggested that as many as 5 million workers in the UK suffer from various forms of voice impairment (Hazards, 2004).

Some groups of workers, such teachers, are identified as being particularly at risk (Martin and Darnley, 2004), and over the last decades new high-risk occupations have emerged. These include call-centre workers and fitness instructors (Dejonckere, 2001) and workers in the rapidly expanding voiceover industry.

Voice problems pose a major problem for individuals, employers, and national healthcare systems. In teachers alone, the overall cost of voice disorder in the US has been estimated at $2.5 billion per year Verdolini & Ramig, 2001), and estimates for the UK are about £200m (Hazards, 2004), but the impact on individuals cannot be measured purely in financial terms. The consequences can be devastating if the ability to work and to communicate effectively are compromised with severe effects on mental health and well-being.

While these problems have long been recognised, the situation remains largely unchanged, even in teachers who are by far the most extensively studied professional group in this respect, and recent studies still report a high prevalence of voice problems in this profession (Van Houtte et al., 2011). This is a public health issue that requires both preventative strategies and early identification linked to effective intervention.

It is frustrating that the majority of occupational voice disorders are behavioural and hence, by definition, preventable. In spite of evidence that voice-care training can reduce the risk of voice problems, many teachers still receive no formal voice training as part of their education (Hazlett et al., 2011) and without integration of voice care practices into everyday routine any benefits of training may not last.

Early identification and intervention is thought to be effective in reducing the incidence of serious and persistent voice disorders, but we still lack tools for identification of early stage voice problems, and current approaches to intervention are often costly and difficult to access. At present, perceptual analysis

of voice remains the gold standard in assessment of voice quality. However, it may be that perceptual rating scales used by speech and language therapists are not sensitive enough to the subtle changes in voice quality that occur with the development of behavioural voice problems.

Clinical application of acoustic analysis has developed greatly over the last 40 years, but acoustic differentiation between healthy and disordered voices is still inexact (Awan *et al.* 2015). As with the perceptual rating scales, the acoustic patterns which might identify those most at risk of developing voice disorders or signal the earliest stages of behavioural voice problems are probably too subtle ever to show up reliably in 'snapshot' comparisons with population thresholds. The rationale for longitudinal monitoring of voices in addition to comparison with norms is that unusual patterns of within-speaker fluctuation, or changes from an individual baseline, may signal the earliest stages of voice disorder (whether behavioural and/or structural) well before any acoustic parameters exceed population thresholds for disorder. We are beginning to establish typical patterns of vocal fluctuation (Schaeffler and Beck 2017) and there are good theoretical reasons to hypothesise that atypical fluctuation could signal that a speaker is at risk. For example, lowering of F0 during the second part of the day (rather than the typical rise) could reflect inflammatory changes within the larynx (Hirano and Bless, 1993).

2 USING TECHNOLOGY FOR VOICE HEALTH INTERVENTION

Smartphones are an attractive choice for technology-enabled intervention due to their popularity, internet connectivity, intuitive user interface, computational power, and mobile availability. Their relatively small screens limit the length and complexity of information that can be displayed, but smartphone-based intervention can easily be combined with non-mobile web-browser-based intervention, as most smartphone users will also have access to devices with larger screens.

2.1.1 Acoustic voice monitoring for voice health intervention

Acoustic analysis is an established component of voice health assessment. Audio recordings through smartphones could help evaluate a client's vocal health status. Smartphones could provide an accessible, convenient, cost-effective and efficient way to track voice stability, deterioration and improvement, especially if parts of the process could be automated. Acoustic analysis for vocal health purposes has so far mainly been investigated in clinical environments, but a range of recent studies, summarised in Table 1, have suggested that the audio quality of smartphone recordings might be sufficient for meaningful and reliable acoustic analysis.

When interpreting the results of these studies it is important to consider the statistical methods used to analyse the data. The statistical methods used by these studies in assessing agreement of smartphones include: correlation (Pearson's correlation, Intraclass Correlation Coefficient (ICC), and Spearman's correlation); comparison of means (student t-test, ANOVA, Friedman repeated-measures); and Bland-Altman analysis.

Several of these statistical methods are inappropriate for assessing agreement (Altman and Bland, 1983; Daly and Bourke, 2000). Correlation only measures the strength of the linear association between variables; it does not assess agreement. This problem is somewhat reduced when using ICC, as it averages the correlations among all possible orderings of the pairs of values. However, ICC is highly influenced by the variance between subjects. When using comparison of means from two devices to assess agreement, authors frequently assume that *non-significant* results reflect high agreement between two devices (e.g. Grillo et al. 2016). However, this is not the case.

Statistical tests assess whether systematic differences are meaningful and any random error is only considered with reference to that. Significance tests like t-tests assess a difference by dividing the difference between two samples through a measure of the random error found in the samples, e.g. the

standard error. High absolute values of the test statistic (e.g. t for the t-test) will yield significant results. As the random error serves as the denominator of the fraction, high random errors will lower the test statistic and thus yield non-significant results. In other words, highly unreliable measurements are more likely to yield non-significant results. Statistical tests are therefore only useful to assess the systematic error of a measurement. Systematic errors are not a matter of reliability but of validity (e.g. Drost 2011) and can be accounted for by calibration. The assessment of the reliability of a measure will require proper consideration of the random error.

To overcome these issues, Bland and Altman proposed the assessment of agreement between two devices is done through analysis of both the systematic error or bias (i.e. mean of the differences between measurement methods) and the random error expressed in the 95% limits of agreement (i.e. bias ± 1.96 standard deviations, Bland and Altman, 1986). The overall agreement of two measurement methods is a combination of the bias and the 95% limits of agreement, the former representing any systematic error and the latter representing the random error. A scatter plot of the difference between two measurement methods against the average of the methods is now known as a Bland-Altman plot. In the following we will describe the systematic error as 'bias', and the random error as the 'critical difference'. The critical difference is defined as half the range between bias + 1.96 standard deviations and bias - 1.96 standard deviations.

Of the studies so far comparing smartphone devices with studio microphones, only two have included Bland-Altman plots (Uloza, *et al.* 2015; Manfredi *et al*. 2016). Both of these studies interpreted the plots visually but did not use the bias and limits of agreement in their assessment of agreement.

When assessing random errors an important question arises with respect to the acceptability of their size. All measurements come with random errors, and therefore it has to be decided what level of random error would be acceptable for a certain purpose. The size of the error therefore needs to be assessed with

reference to certain criteria and for acoustic voice parameters these criteria still have to be developed. In this study we will report random error size in absolute terms as well as with reference to the total range of measurements made with the reference device. The latter measure is motivated by the fact that the typical range of a measure will inform acceptable tolerance levels. For example, if any measured quantity usually varies between, say, 10 and 20, then a random error of ±1 (i.e. 10% of the total range) would probably be considered as more problematic than if the same quantity typically varied between 10 and 1010, where a random error of ± 1 unit would constitute 0.1% of the range. We will return to this question in the discussion section.

Insert Table 1 here

Table 1 - Summary of studies assessing reliability of smartphone use in voice quality analysis

Authors	Year	# of subjects	Speaking task	Recording setup	Devices used	Acoustic measures	Statistics	Results
Vogel *et al*.	2015	15 normal voices	Sustained vowel and reading passage	Simultaneous recordings with all devices	HDD + table microphone; Landline telephone; iPhone; PC + head-mounted microphone	F0 (mean, min, max, SD); jitter %; shimmer %; NHR; CPP; CPPs	Spearman's correlation; RMSE with 10% threshold	11 out of 14 comparisons between iPhone and PC recordings were correlated. Only F0 and CPP had error below 10% RSME.
Uloza *et al.*	2015	118 (34 normal; 84 pathological voices)	Sustained vowel	Simultaneous recordings with all devices	Studio microphone; Samsung Galaxy Note 3	F0; jitter %; shimmer %; NNE; HNR; SNR	Student t-test; Pearson correlation; Bland-Altman plot	Significant differences for all acoustic measures except F0 and shimmer %.
Grillo *et al*.	2016	10 normal voices	Sustained vowel and sentence	Simultaneous recordings with all devices	Head-mounted microphone; iPhone 5 and 6s; Samsung Galaxy S5	F0; SD of F0; jitter %; shimmer %; NHR; CPP; AVQI	ANOVA main effects and interactions	With regards to main effect of the device on acoustic measures, there were only no significant effects for F0.
Manfredi *et al*.	2016	Synthesised voices	Sustained vowel	Devices recorded simultaneously through loudspeaker	Studio microphone; HTC One; Wiko Smart2	F0; jitter %; shimmer %; NHR	ICC; Bland-Altman plots	All measures had high ICC – lowest correlation value was 0.81.
Maryn *et al*.	2017	50 (12 normal; 38 pathological voices)	Sustained vowel and reading passage	Devices recorded individually through loudspeaker	Studio microphone; iPad 2; Google Nexus 9; iPhone 5s; Samsung Galaxy s5; Nokia Lumia 520	Median F0; jitter (% and RAP); shimmer (% and dB); HNR; GNE; CPPs	Freidman repeated-measures; Wilcoxon signed-rank	No significant difference for F0 across all devices. No significant difference for jitter % and RAP, GNE, and CPPs for specific devices.
Kojima *et al*.	2018	6 normal voices	Sustained vowel	Simultaneous recordings with all devices	Studio microphone; MediaPad M3	F0; SNR	Spearman's correlation	High correlations for both acoustic measures (F0 1.00; SNR 0.89)

Unknown precision and accuracy of measurement and unclear criteria for acceptable levels of error are amongst the reasons that make acoustic tracking of voice health, while appealing, technologically ambitious and there are substantial knowledge gaps that inhibit the development of a smartphone based 'acoustic voice health tracker'. We still have rather limited information about typical or clinical patterns of variation.

Our current empirical research efforts focus on acoustic voice monitoring as the most technically ambitious and innovative aspect of the different types of interventions outlined above. Our group has developed a smartphone app "Fitvoice" that allows capture of voice-related audio and survey data (see Grillo *et al* 2016 for an alternative concept). Initial field tests have been promising (Schaeffler and Beck, 2017), but a better understanding of the capabilities and limitations of smartphones for monitoring acoustic parameters is still required. We have therefore recorded a dataset of healthy voices across five simultaneous channels, four popular smartphones and one studio-quality microphone.

The present study focussed on the relation between voice parameters derived from smartphone recordings and studio quality recordings, in order to establish whether parameter values extracted from smartphone recordings deviate from those extracted from studio microphone recordings, and if any deviations can be described as random or systematic errors.

3 METHOD

3.1.1 Speakers

22 vocally healthy woman and men (12 women; 10 men, aged 20-64y) produced two trials of sustained phonation of the vowel sound /a/ and read aloud a shortened version of the phonetically balanced text "The Dog and Duck Story" (Brown & Docherty 1995).

The vocal health of the participants was determined by the participants themselves and confirmed perceptually by the researchers during conversational speech on the day of the recording.

3.1.2 Reference Microphone

The microphone used as a reference was a Neumann U89i (Georg Neumann GmbH, Berlin, Germany) using a hypercardiod polar pattern. This microphone has a frequency range from 20Hz-20kHz and has a flat frequency response from around 40Hz-15kHz. It has an equivalent noise level of 17 dB_A and reaches total harmonic distortion of 0.5% at 134 dB_{SPL}. The microphone was connected to a Digidesign Digi 003 audio interface (Avid Audio, California, US) and recorded on an iMac using ProTools LE version 8.0.3 (Avid Audio, California, US). Recordings were made at 44.1 kHz sampling rate and 32-bit resolution.

3.1.3 Smartphones

Five smartphone devices were initially selected for use in the study representing a range of devices from the two large manufactures in terms of market share in the UK in 2017 – Apple and Samsung. The devices are Samsung Galaxy S8+ (SG8), iPhone 6s (i6s), iPhone 7 (i7). Samsung Galaxy J3 (SJ3) and iPhone SE. The iPhone SE device did not transmit voice recordings of 4 speakers, and therefore was excluded from further analysis.

3.1.4 Recording procedure

Recordings were made in a soundproof booth with ambient noise levels of < 45 dB_A (measured using a Casella CEL-254 sound level meter).

All smartphones were placed in semi-circular array directly adjacent to each other with the devices angled at approximately 45^o to replicate holding the device and keeping the microphone to mouth distance across phones relatively constant. As positioning a microphone to the side instead of directly in front might have some effect on recordings, the order of the smartphones was systematically cycled

across participants so that phone positioning was balanced across participant recordings. The reference microphone was placed in the centre of the array. Speakers stood 20cm from the centre of the array during recordings (see figure 1). All utterances were recorded simultaneously on all devices.

The smartphone devices used the Fitvoice™ app for recordings. This app is freely available for Android and iOS devices, records at 44.1 kHz sampling rate, saves recordings as uncompressed .wav (PCM), and allows recordings to be uploaded securely to a server.

Figure 1- The experimental setup with an array of four smartphone devices and reference microphone. *Insert Figure 1 here.*

3.1.5 Acoustic analysis

The sustained vowel and passage reading recordings were analysed separately for all 220 recordings (22 participants x 5 devices x 2 trials). The acoustic analysis was completed using Praat (Boersma & Weenink, 2015) and automatized through a Praat script. The extraction of acoustic parameters

proceeded in two steps. First, voiced segments in the recordings were extracted using a modified version of the *Praat* script by Maryn and Weenink (2015) and then four acoustic parameters were measured on the voiced segments only. The following acoustic parameters were extracted: Mean F0; smoothed cepstral peak prominence (CPPS); Jitter (RAP) and Shimmer %. CPPS and Shimmer % were extracted following Maryn and Weenink's (2015) procedures, while mean F0 was extracted following recommendations published in the Praat manual, using a cross-correlation algorithm. Jitter (RAP) was extracted using the Praat "voice report" functionality. The choice of parameters was motivated by their popularity and proven usefulness. Mean F0 is frequently extracted from speech signals for clinical and non-clinical purposes. Shimmer and Jitter are historically important parameters for clinical voice assessment, and Shimmer % is also a sub-component of the Acoustic Voice Quality Index (AVQI, Maryn and Weenink, 2015). CPPS was chosen because it is a sub-component of the AVQI and has generally shown good performance in studies of discrimination of healthy and disordered voices. Jitter (RAP) was chosen as it is one of the recommended jitter measures in the MDVP manual (KayPentax 2008) and clinical thresholds have been published for this measure.

3.1.6 Statistics

Statistical analysis was performed using R 3.4.0 (R Core Team, 2017) and IBM SPSS Statistics 21. Bland-Altman analysis was performed with R package "BlandAltmanLeh" (Lehnert 2015). This package provides the 95% confidence interval for the bias. A bias was judged as significant if the confidence interval excluded zero.

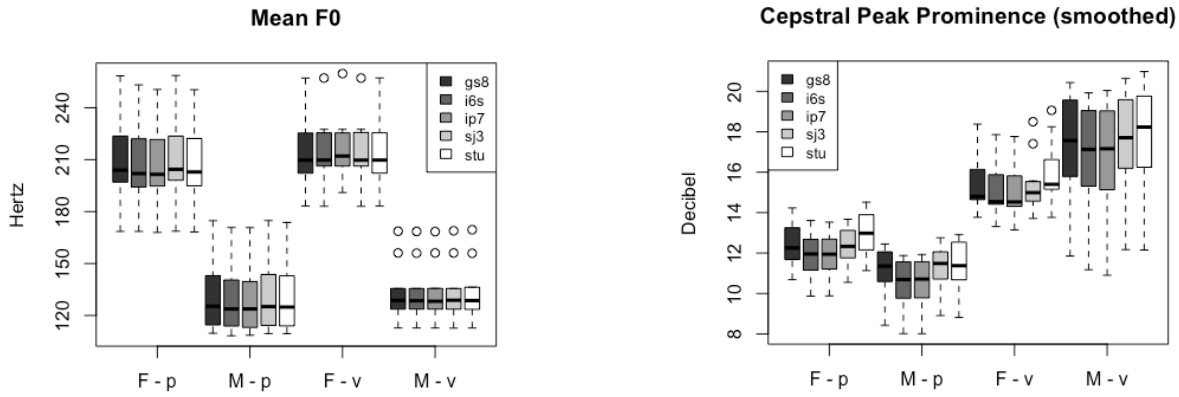
We used BA analysis to compare the means between the two recording samples from the reference microphone with the means from the four smartphones. As mentioned above, the size of the critical difference (random error) is given in two ways: as an absolute value and as a percentage of the total range of the respective parameter, as measured with the reference microphone.

4 RESULTS

4.1 DESCRIPTIVE DATA

Figure 2 provides boxplots for all parameters, arranged by gender, speech material and device type (four mobile phones and one studio microphone). The figure indicates that spread and level is more similar for some parameters than for others. Mean F0 shows a fairly similar distribution across devices for both vowel and passage data. CPPS shows a tendency for lower measurement values on the phones compared to the studio microphones, for both vowel and passage data and both genders, although male and female values seem to behave differently for vowel and passage data. CPPS is lower in the male sample for vowel data, but higher for the passage data. Shimmer % values tend to be higher on smartphones than on the studio microphone, especially for the passage data; the vowel data is less easy to interpret. Jitter (RAP) does not show clear tendencies overall, although values seem to be lower on smartphones for the male passage data.

Insert Figure 2 here



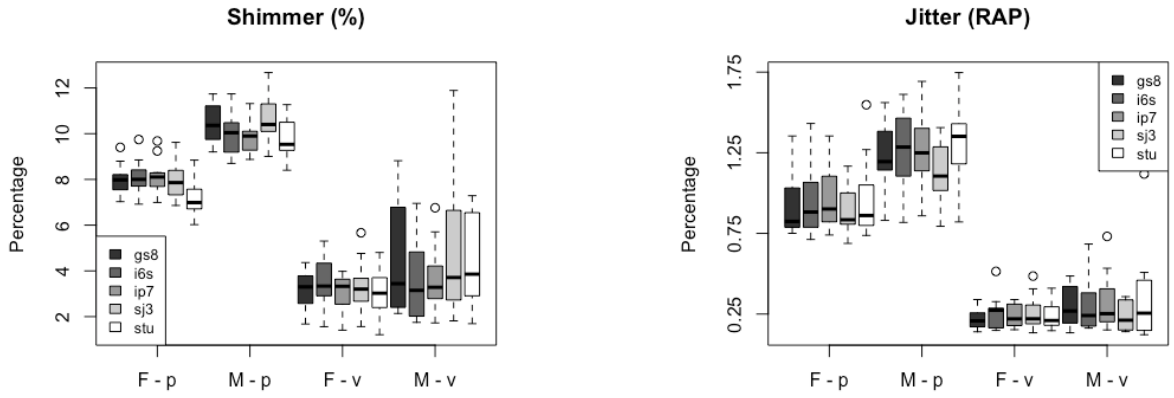


Figure 2: Boxplots for four acoustic parameters across female (F) and male (M) speakers, reading passages (p) and sustained vowels (v) and five devices, Samsung Galaxy GS8+ (gs8), iPhone 6s (i6s) and iPhone 7 (ip7), Samsung J3 (sj3) and a studio quality microphone, Neumann U89i (stu).

4.2 BIAS AND CRITICAL DIFFERENCE FOR ACOUSTIC MEASURE MEANS WITH REFERENCE TO STUDIO MICROPHONE

Bias and critical difference were calculated for all devices against the reference microphone. Tables 2 to 5 summarise the results for the four acoustic parameters for both sustained vowel recordings and passage reading recordings and Figures 3 to 6 provide the corresponding Bland-Altman (BA) plots. The left figure always shows the vowel parameters, the right figure the passage parameters. The tables provide the critical difference in absolute terms and as a percentage of the overall range of a parameter as derived from the studio microphone measurements. Bias and critical difference were calculated for male and female samples combined. The BA plots show data points marked by gender to assess any influence of gender on overall bias and critical difference. As mentioned above, significance tests were performed for bias values using the 95% confidence interval.

4.2.1 Mean F0

Table 2 summarises bias and random error data for mean F0 and Figure 3 shows corresponding BA plots.

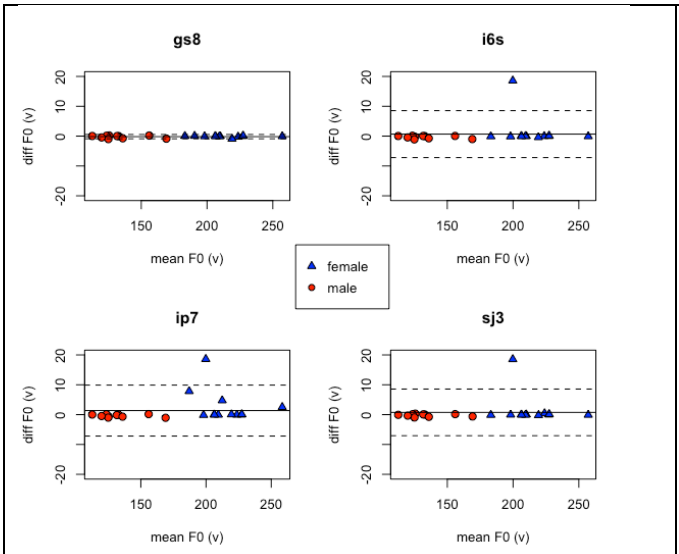
4.2.1.1 Vowels

There was a very small but significant negative bias for the gs8 in the vowel data, and a critical difference of less than 1 Hz. The other phones showed no significant bias, and higher critical differences of about 7-9 Hz, constituting around 5% to 6% of the range. An inspection of the Bland-Altman-plots (see Figure 3a) indicates that the higher errors are related to the female data. The plots also indicate an influence of a small number of outliers rather than a general reliability difference between phones. If the most extreme outlier is removed from the i6s, ip7 and sj3 data, the critical differences become 0.69 Hz (0.5%), 4.09 Hz (2.8%) and 0.64 Hz (0.4%) respectively, which is much closer to the error measured with the gs8.

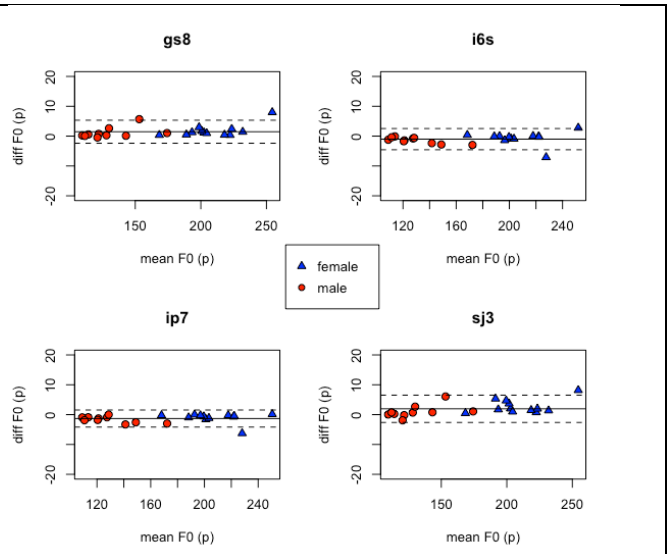
4.2.1.2 Passage

All phones showed a significant bias in the range of -1.31 to 1.95 Hz for mean F0 from the passage data. The bias was positive for the Samsung phones and negative for the iPhones (cf also Figure 2). All phones showed similar critical differences in the range of 2.86 to 4.57 Hz (2%-3.2%). The iPhone 7 showed the smallest error (which is interesting as this phone showed the largest error with the vowel data, even after outlier removal), and the SJ3 the largest, but all phones were well below the 10% criterion for this parameter. BA plots show less influence of outliers compared to the vowel data, but variation seems to be somewhat larger in the female data.

Add figure 3 about here



(a)



(b)

Figure 3: Bland-Altman plot for comparison of F0 measurements obtained from each smartphone device and the reference microphone. The plots on the left (a) are for the vowel speech task. The plots on the right (b) are for the passages.

Insert table 2 here

Table 2: Mean difference and 95% limits of agreement between the reference microphone and each smartphone device for mean fundamental frequency (F0).

		Mean difference (bias)	95% Limits of Agreement (random error)			
			Lower limit	Upper limit	Critical difference	Range %
Samsung Galaxy S8+	*vowel*	**-0.17***	-0.89	0.55	0.72	0.50
iPhone 6s	*vowel*	0.67	-7.20	8.54	7.87	5.40
iPhone 7	*vowel*	1.35	-7.19	9.89	8.54	5.90
Samsung J3	*vowel*	0.75	-7.06	8.56	7.81	5.40
Samsung Galaxy S8+	*passage*	**1.47***	-2.39	5.33	3.86	2.70

iPhone 6s	*passage*	**-1.01***	-4.57	2.54	3.56	2.50
iPhone 7	*passage*	**-1.31***	-4.17	1.55	2.86	2.00
Samsung J3	*passage*	**1.95***	-2.62	6.52	4.57	3.20

* indicates significance based on the 95% confidence interval of the mean difference.

4.2.2 Smoothed Cepstral Peak Prominence (CPPS)

Table 3 summarises bias and random error data for smoothed cepstral peak prominence (CPPS) and Figure 4 shows corresponding BA plots.

4.2.2.1 Vowels

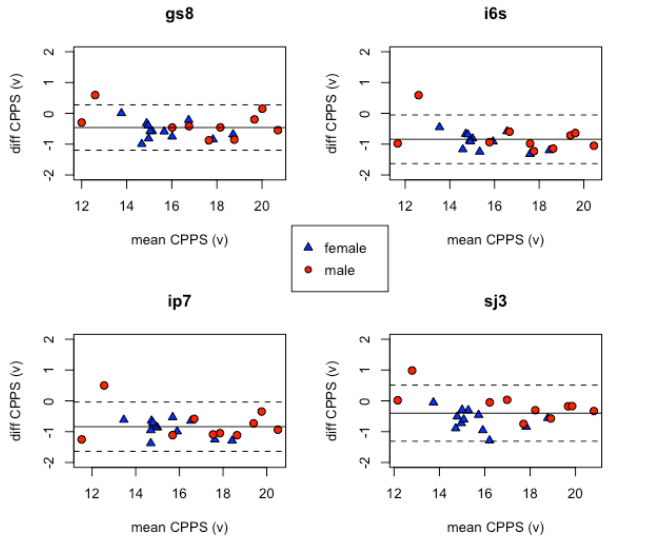
All phones showed a significant negative bias for the CPPS vowel data between -0.84 and -0.4 dB, suggesting that CPPS measures from phone recordings were overall somewhat lower than those from the reference microphone (see also boxplots in Figure 2). There was a clear tendency for the absolute amount of bias to be higher in the iPhones than in the two Samsung models, by about a factor of two.

The critical difference lay between 0.74 and 0.91 dB, constituting between 8.3% and 10.3% of the reference range. The differences between phones were small, but the Samsung J3 showed the highest critical difference. Comparison of gender groups suggests that the male data showed somewhat higher random error variation for the vowel data of this parameter.

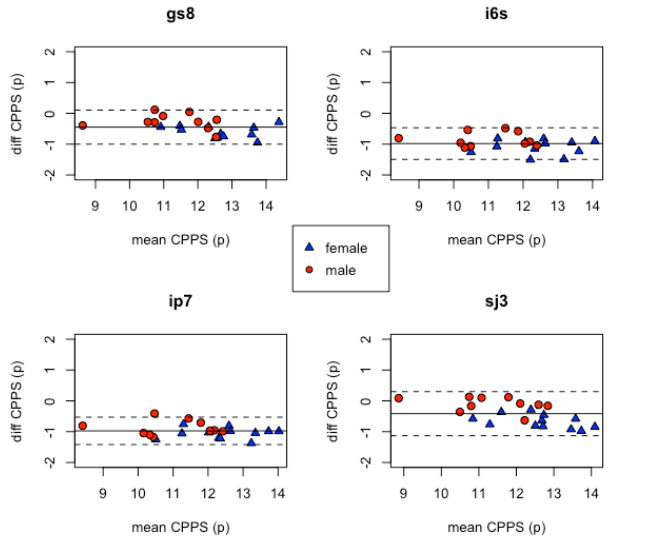
4.2.2.2 Passage

All phones showed a significant negative bias between -0.98 and -0.45 dB for the CPPS passage data, thus approximately the same order of magnitude as with the vowel data. The trend for a lower absolute bias for Samsung phones was confirmed, again by about the factor two. Critical differences lay between 0.44 and 0.72dB, constituting 9 to 12.6% of the reference range. Again, the Samsung J3 performed a bit

worse on this measure. Comparison between genders does not suggest large differences for this sample, however there is a tendency for higher means to show more negative bias values (compare especially the sj3 data in Figure 4b). This could be a gender or a value size effect as female values show both more negative differences as well as higher mean values in this instance.



(a)



(b)

Figure 4: Bland-Altman plot for comparison of CPPS measurements obtained from each smartphone device and the reference microphone. The plots on the left (a) are for the vowel speech task. The plots on the right (b) are for the passages.

Insert figure 4 here

Insert Table 3 here

Table 3: Mean difference and 95% limits of agreement between the reference microphone and each smartphone device for smoothed cepstral peak prominence (CPPS).

		Mean difference (bias)	95% Limits of Agreement (random error)			
			Lower limit	Upper limit	Critical difference	Range %
Samsung Galaxy S8+	*vowel*	**-0.46***	-1.20	0.28	0.74	8.30
iPhone 6s	*vowel*	**-0.84***	-1.63	-0.05	0.79	8.90
iPhone 7	*vowel*	**-0.84***	-1.64	-0.04	0.80	9.10
Samsung J3	*vowel*	**-0.40***	-1.31	0.51	0.91	10.30
Samsung Galaxy S8+	*passage*	**-0.45***	-1.00	0.11	0.55	9.70
iPhone 6s	*passage*	**-0.98***	-1.49	-0.47	0.51	9.00
iPhone 7	*passage*	**-0.97***	-1.42	-0.53	0.44	7.80
Samsung J3	*passage*	**-0.42***	-1.13	0.30	0.72	12.60

* indicates significance based on the 95% confidence interval of the mean difference.

4.2.3 Shimmer %

Table 4 summarises bias and random error data for Shimmer % and Figure 5 shows corresponding BA plots.

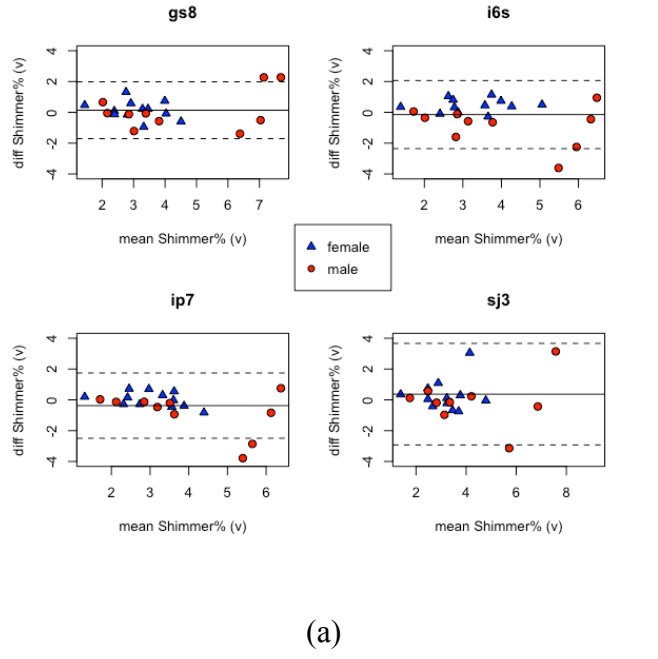
4.2.3.1 Vowel

None of the phones showed a significant bias for shimmer measurements for the vowel data. Critical differences were between 1.84 and 3.3%, constituting 30.3 to 54.2% of the reference range. The Samsung J3 showed the worst performance in this context, but overall all phones showed errors that were well beyond the 10% criterion. BA plots reveal that both types of errors are influenced by four outliers from the male data (Figure 5a) and maybe one female data outlier for the sj3.

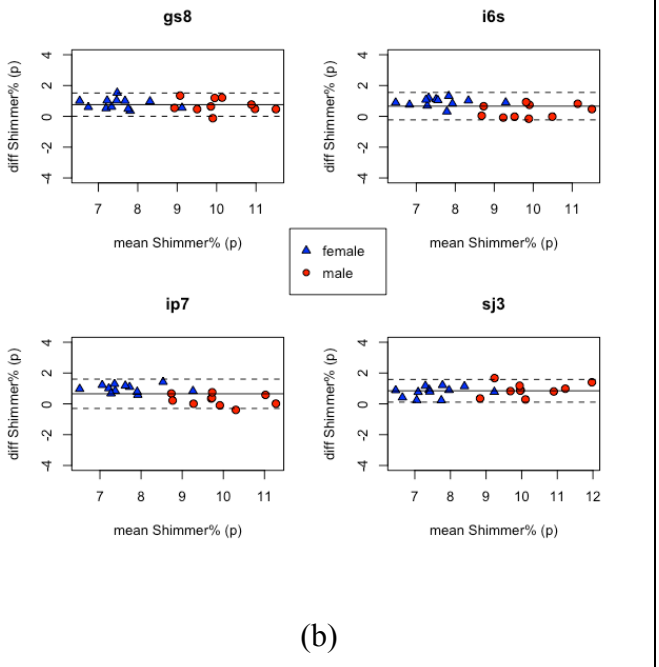
4.2.3.2 Passage

All phones showed a significant positive bias between 0.65% and 0.85%, suggesting that shimmer measures derived from phone recordings were somewhat higher than from studio microphone recordings,

although this trend could only be seen in the passage data. Critical difference values lay between 0.74 and 0.95%, comprising 14% to 18.2 % of the range. The larger critical difference for the Samsung J3 in the vowel data was not confirmed by the passage data. The male sample showed higher shimmer mean values for the passage data, without a major effect on random error. Both iPhones seem to show a small effect of decreasing difference values with higher mean values. As with CPPS above, this could be a gender or a value size effect as male values show both lower differences as well as higher mean values.



(a)



(b)

Figure 5: Bland-Altman plot for comparison of Shimmer % measurements obtained from each smartphone device and the reference microphone. The plots on the left (a) are for the vowel speech task. The plots on the right (b) are for the passages.

Insert figure 5 here

Insert table 4 here

Table 4: Mean difference and 95% limits of agreement between the reference microphone and each smartphone device for Shimmer local (Shimmer%).

		Mean difference (bias)	95% Limits of Agreement (random error)			
			Lower limit	Upper limit	Critical difference	Range %
Samsung Galaxy S8+	*vowel*	0.14	-1.71	1.98	1.84	30.30
iPhone 6s	*vowel*	-0.15	-2.36	2.06	2.21	36.30
iPhone 7	*vowel*	-0.37	-2.49	1.75	2.12	34.90
Samsung J3	*vowel*	0.37	-2.93	3.67	3.30	54.20
Samsung Galaxy S8+	*passage*	**0.76***	0.01	1.51	0.75	14.40
iPhone 6s	*passage*	**0.66***	-0.23	1.55	0.89	17.00
iPhone 7	*passage*	**0.65***	-0.30	1.61	0.95	18.20
Samsung J3	*passage*	**0.85***	0.11	1.59	0.74	14.00

* indicates significance based on the 95% confidence interval of the mean difference.

4.2.4 Jitter (RAP)

Table 5 summarises bias and random error data for jitter (RAP) and Figure 6 shows corresponding BA plots.

4.2.4.1 Vowel

None of the phones showed a significant bias for jitter (RAP) for the vowel data. The critical difference lay between 0.28% and 0.36%, constituting 27.6% to 35.9% of the range[1]. BA plots indicate that the critical difference is influenced by an outlier in the male data. If this outlier is removed, errors change to 0.09, 22.7% (gs8), 0.12, 31.3% (i6s), 0.12, 31.3% (sp7), 0.17 43.5% (sj3)[2].

4.2.4.2 Passage

The Samsung phones showed a significant negative bias for the jitter passage data, with a higher absolute amount for the sj3 than for the gs8. The critical difference was similar for the gs8 and the two iPhones (0.16-0.18%, comprising 15.4% to 17.7% of the range), while the error of the sj3 was higher by about a factor of two (0.32%, 31.7% of the reference range). There is no clear effect of gender identifiable in the BA plots, but as with some other parameters there is a tendency of decreasing difference values with increasing mean values, most visible in the sj3 data.

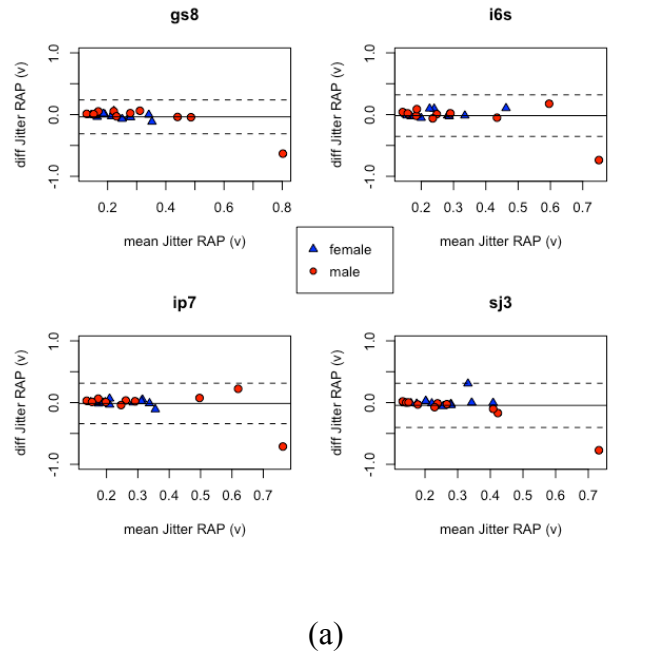
---

[1] The similarity between critical difference and percentage of reference range stems from the fact that the jitter (RAP) values in our sample ranged from 0.12 to 1.12% in the vowel data.

[2] Note that removal of this outlier also has an influence on the range of the parameter, so absolute error changes more than relative error.

Insert figure 6 here.

Insert Table 5 here



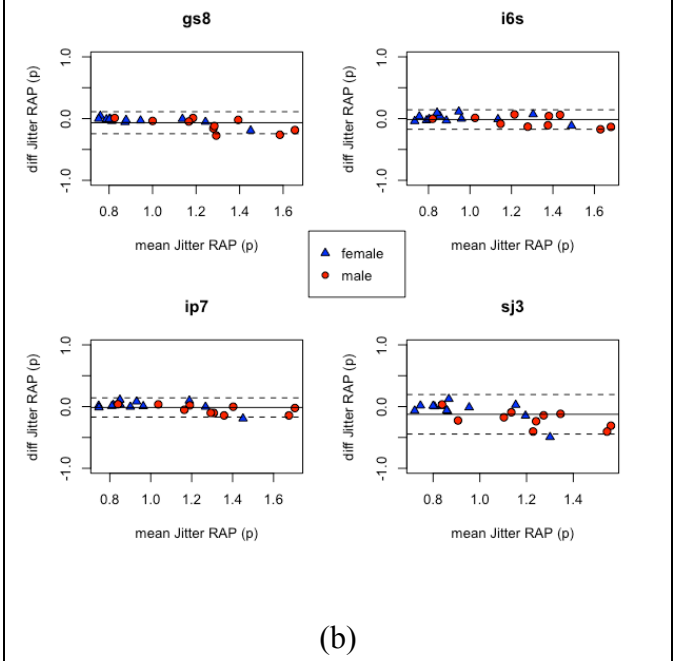


Figure 6: Bland-Altman plot for comparison of Jitter (RAP) measurements obtained from each smartphone device and the reference microphone. The plots on the left (a) are for the vowel speech task. The plots on the right (b) are for the passages.

Table 5: Mean difference and 95% limits of agreement between the reference microphone and each smartphone device for relative average perturbation Jitter (RAP).

		Mean difference (bias)	95% Limits of Agreement (random error)			
			Lower limit	Upper limit	Critical difference	Range %
Samsung Galaxy S8+	*vowel*	-0.04	-0.31	0.24	0.28	27.60
iPhone 6s	*vowel*	-0.02	-0.35	0.32	0.34	33.80
iPhone 7	*vowel*	-0.01	-0.34	0.31	0.33	32.90

Samsung J3	*vowel*	-0.05	-0.40	0.31	0.36	35.90
Samsung Galaxy S8+	*passage*	**-0.07***	-0.25	0.11	0.18	17.70
iPhone 6s	*passage*	-0.02	-0.17	0.14	0.16	15.60
iPhone 7	*passage*	-0.01	-0.17	0.14	0.16	15.40
Samsung J3	*passage*	**-0.13***	-0.45	0.20	0.32	31.70

* indicates significance based on the 95% confidence interval of the mean difference.

5 DISCUSSION

The present study investigated whether acoustic parameter values extracted from smartphone recordings deviate from those extracted from studio microphone recordings.

Overall, our study suggests that acoustic parameters can be measured with smartphones with varying reliability. When random error size is related to the total range of a parameter, F0 and CPPS show relatively small random errors, which Jitter (RAP) and Shimmer % show relatively large random errors. This finding is only in parts comparable to other studies investigating the clinical feasibility of smartphone devices in voice analysis. This is partly due to the fact that other studies have used statistical approaches that focus on systematic error and ignore the role of random error, which in our view leads to rather optimistic assessment of measurement reliability. Most of the previous studies have either not used Bland-Altman analysis, or only included visual interpretation of the Bland-Altman plots without further quantification. This makes direct comparison between the results presented here and the results of previous studies difficult.

This study is the first to report the bias and random error for each smartphone compared with a reference studio microphone using four acoustic parameters. In medical studies, Bland-Altman analysis ideally

uses random error ranges that are defined on the basis of knowledge about the parameter being measured and the normative ranges of the population being assessed. In the area of voice science, acoustic parameters are often described in terms of a threshold between normal and disordered voice quality, as compared with perceptual measures.

It is worth considering all four parameters here in turn and assessing the relevance (including clinical relevance) of their random error.

For F0 we observed a very small bias for on phone (GS8) for the vowel data, and small biases for the passage data for all phones. Interestingly, the Samsung phones produced F0 measurements that were slightly too high, while those from the Apple phones were slightly too low, but as the systematic errors never exceeded 2 Hertz, their practical relevance is probably limited. The biases should be irrelevant if recordings are compared for one and the same device across different time points. The random error is in the region of ±7-9Hz for the vowel data, apart for the gs8, which shows a much smaller random error. Random errors for the other phones are probably inflated by outliers but even without removal of outliers the random errors seem acceptable. ±7-9Hz is definitely above the human difference limens for pitch perception (Moore 1973) but not dramatically so. Users of mobile phone recordings should consider that F0 changes from recording to recording that do not exceed these error ranges might be artefacts. The random errors for connected speech are generally lower than those for vowels, never exceeding ±5Hz. Again this is above human difference limens thresholds but might not matter for many purposes.

For CPPS all phones showed a small but significant bias, with phone measures somewhat lower than reference measures. The bias tended to be smaller for the Samsung phones (around -0.5 dB) compared to the Apple phones (-0.8 to -1 dB), and biases were similar across speech tasks. For comparisons across devices, these bias values could provide guidance for calibration. Random error was in the region of

±0.7-0.9 dB for sustained vowels and ±0.4-0.7 dB for passage reading. There is a tendency for the sj3 – the cheapest phone in our sample – to perform somewhat worse than the other three phones. Errors below 1 dB give the impression of a small error, but it needs to be taken into account that the overall range of CPPS measures was not extremely high in our data so that this small error value corresponds to around 10% of the total range of values we measured. However the relevant range of CPPS values might be in fact much wider. Maryn et al (2015) report a range of CPPS values between 2.65 and 17.68 dB. Their sample includes pathological voices which explains the much wider range. If we assume that the range of CPPS measures that are relevant for clinical assessment is around 15 dB then an error of ±0.5 to ±1 dB is probably acceptable. Normative thresholds for CPPS are difficult to find, and published CPPS values can vary widely, even if we only consider CPPS measured with Praat. For example, Sauder et al (2017) report CPPS ranges between 16.47 to 22.99 dB for healthy voices and 14.71 to 20.31 dB for disordered voices. They suggest a cut-off value for 19.10 dB for pathology for CPPS measured with Praat, which neither fits our nor Maryn et al's (2015) values. This issue will require further investigation. For phone measures of CPPS researchers should be aware that changes below ±1 dB might be artefacts, even if measured with the same devices.

Shimmer % does not show bias for vowel measures but for passage measures, with phone measures exceeding studio microphone measures by 0.7 to 0.9 pp. Random errors are in the region of ±2 to ±3 pp for sustained vowels and ±0.8 to ±1 pp for the passage. If we compare this range to the total range of measures then we find that the error is between 30% and 50% of the range in the vowel data. The Praat manual provides 3.81% as threshold for pathology for Shimmer %. Our lowest measured value with studio equipment for vowels was 1.21%. The difference between our lowest value and the pathological threshold is thus 2.6%. An error of ±2% seems unacceptably high in this context and we therefore would not recommend using smartphone recordings for shimmer measurements.

Jitter (RAP) measurements do not show significant bias for sustained vowels. The two Samsung phones show significant bias for the passage data. Random errors are in the region of ±03 to ±0.4 pp for the vowel data. For all phones this constitutes around 30% of the measured range of values. The Praat manual provides a threshold of 0.68% for Jitter (RAP). Our lowest RAP value was 0.122%. The difference between our lowest value and the pathological threshold is thus 0.56pp. An error of 0.3pp seems unacceptably high in this context.

While the random error values allow decisions to be made about the reliability of smartphones as compared with a reference microphone, the bias value allows the readings for the smartphone to be adjusted, especially when comparing across different devices. If the bias values are significantly different from zero they are considered a fixed bias and therefore can be used to adjust the values from the smartphone accordingly. Given the results of the present study we suggest that a calibration approach could be beneficial, at least if comparisons between mobile phone recordings and studio microphone recordings are important for a certain application. Most calibration methods that relate a new, cheaper or field method to a standard use inverse regression to estimate the "true" reference value from the non-reference measurement (see e.g. Burke 2001). We aim at addressing this topic in a separate publication but would, for the moment, suggest that calibration with the Bland-Altman derived biases published here could be beneficial for some parameters, especially CPPS.

The use of smartphones in voice research is particularly useful due to the freedom for speakers to record their voice where it is convenient for them to do so. However, the effect of room acoustics and ambient noise becomes a significant issue that the current study did not address.

Maryn *et al.* (2017) investigated the combined effect of environmental noise and a variety of mobile recording devices on 10 acoustic parameters including all those used in this study. They found that only F0 was robust against both environmental noise and recording system. However, they also found that the

majority of acoustic markers differed significantly from the reference microphone values even under optimal ambient noise (20.5 dB_A). This is possibly due to the statistical methods used in this study – largely tests of mean differences. As the authors point out, however, acoustic parameters measured between commercially available recording systems that are considered standard clinical tools show considerable variability as well (Maryn *et al.* 2017).

Lebacq *et al.* (2017) utilised the same synthetic voice as in Manfredi *et al*. (2017) and added external noise to the recordings. They found that for jitter measures and NHR, good reliability remains up to ambient noise levels of 50 dB_A. Other acoustic parameters, however, may be more sensitive to noise than these measures. CPPS, for example, has consistently shown correlation – perceptual and acoustic – with noise, and therefore may be more sensitive to changes in noise level despite being consistently reliable in this study under studio conditions. Future work by the authors will be to conduct similar reliability tests under different ambient noise level conditions.

In our view a clear distinction of random and systematic error is paramount for future studies. Accurate description of various systematic errors including those caused by device type or background noise can form the basis for future calibration solutions and in this way increase reliability of acoustic voice parameters extracted from smartphone recordings. This in turn will allow more reliable judgment of voice condition from smartphone recordings.

6 CONCLUSIONS

Overall, this study adds to the emerging evidence that smartphones can be used for acoustic analysis if certain constraints are considered. If noise and microphone position can be controlled, F0 and CPPS

measures can be derived with smartphones with acceptable random error compared to a high-quality studio reference microphone, meaning that the devices used in this study can be used interchangeably for these measures. For CPPS, calibration is recommended as smartphone based CPPS measures were significantly lower than those derived from the reference microphone.

Other parameters should be treated with caution. Based on our current samples, the random error seems too high for shimmer and jitter to be of any practical use, even for within-device comparisons. Further work is required for the development of reliable reference ranges for acoustic parameters, even under studio conditions.

There were no striking differences in performance between the smartphones. The most affordable phone in our range, the Samsung SJ3, performed somewhat less well for some parameters and prompt types, but results were by and large comparable. Given the highly dynamic smartphone market it might be useful to develop a set of reference recordings that can be used to calibrate mobile phones for clinical use.

Results suggest that acoustic monitoring of voice with smartphones could become a central component of a voice care strategy that engages users and supports self-management, if acoustic parameters used for acoustic assessment are selected and interpreted with care.

We would also discourage the exclusive use of statistical tests like t-tests for reliability assessments of parameter measurements and interpreting non-significance as an indicator of good reliability. Significance tests of mean differences across devices will only provide evidence for systematic errors, which are rather a matter of validity than reliability. Reliability assessments need to focus on the random error and large random errors are a serious problem for any measurement, even if there is no change of device between recordings.

7 REFERENCES

AWAN, S.N., ROY, N., ZHANG, D. & COHEN, S.M., 2015. Validation of the cepstral spectral index of dysphonia (CSID) as a screening tool for voice disorders: development of clinical cutoff scores. Journal of voice, 30, 130-144.

ALTMAN, D.G. & BLAND, J.M. (1983) Measurement in Medicine: the Analysis of Method Comparison Studies. The Statistician 32: 307–317.

LEHNERT, B. 2015). BlandAltmanLeh: Plots (Slightly Extended) Bland-Altman Plots. R package version 0.3.1. https://CRAN.R-project.org/package=BlandAltmanLeh

BLAND, J.M. & ALTMAN, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet. 327(8476), 307–310.

BOERSMA, P. & WEENINK, D., 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/

BROWN, A., & DOCHERTY, G. J. 1995. Phonetic variation in dysarthric speech as a function of sampling task. European Journal of Disorders of Communication*, 30*(1), 17-35.

CREER, T.L., HOLROYD, K.A., 1997. Self-management, in: Baum, A. (Ed.), Cambridge Handbook of Psychology, Health and Medicine. (Cambridge, UK: Cambridge University Press), pp. 255–261.

DA COSTA, V., PRADA, E., ROBERTS, A., COHEN, S., 2012. Voice disorders in primary school teachers and barriers to care. Journal of Voice, 26, 69–76.

DALY, L.E. & BOURKE, G.J. (2000) Interpretation and Use of Medical Statistics. Oxford: Blackwell Science Ltd.

DEJONCKERE, P.H., 2001. Occupational Voice: Care and Cure. (Amsterdam: Kugler Publications).

GREEN, E.C., MURPHY, E., 2014. Health belief model. In W.C. Cockerham, R. Dingwall, S. R. Quah (eds) The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society. (Chichester: Wiley Blackwell), pp.766-769.

GRILLO, E.U., BROSIOUS, J.N., SORRELL, S.L. & ANAND, S., 2016. Influence of smartphones and software on acoustic voice measures. International journal of telerehabilitation. 8, 9-14.

HAZARDS (2004) *Work hoarse*. [Online]. Available from: http://www.hazards.org/voiceloss/workhoarse.htm [Accessed: 20–1 2018].

HAZLETT, D.E., DUFFY, O.M., MOORHEAD, S.A., 2011. Review of the impact of voice training on the vocal quality of professional voice users: implications for vocal health and recommendations for further research. Journal of Voice, 25, 181–191.

HIRANO, M., & BLESS, D. M. (1993). Videostroboscopic examination of the larynx. San Diego, Calif: Singular Pub. Group

ILOMÄKI, I., LAUKKANEN, A.-M., LEPPÄNEN, K., VILKMAN, E., 2008. Effects of voice training and voice hygiene education on acoustic and perceptual speech parameters and self-reported vocal well-being in female teachers. Logopedics Phoniatrics Vocology, 33, 83–92.

KayPENTAX. 2008. Operations Manual: Multi- dimensional Voice Program (MDVP) Model 5105. Lincoln Park, NJ: KayPENTAX.

KOVAČIĆ, G., 2005. Voice education in teacher training: an investigation into the knowledge about the voice and voice care in teacher‐training students. Journal of Education for Teaching: International research and pedagogy, 31, 87–97

LEBACQ, J., SCHOENTGEN, J., CANTARELLA, G., BRUSS, F.T., ET AL., 2017. Maximal ambient noise levels and type of voice material required for valid use of smartphones in clinical voice research. Journal of Voice. 31, 550–556.

LEE, J., KOH, D. & ONG, C., 1989. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Computers in biology and medicine. 19, 61–70.

LIN, E., HORNIBROOK, J. & ORMOND, T., 2012. Evaluating iPhone recordings for acoustic voice assessment. Folia phoniatrica et logopaedica, 64, 122–130.

LYBERG ÅHLANDER, V., HOLM, L., KASTBERG, T., HAAKE, M., BRÄNNSTRÖM, J., SAHLÉN, B., 2015. Are children with stronger cognitive capacity more or less disturbed by classroom noise and dysphonic teachers? International Journal of Speech-Language Pathology, 17, 577-588.

MANFREDI, C., LEBACQ, J., CANTARELLA, G., SCHOENTGEN, J., ET AL., 2016. Smartphones offer new opportunities in clinical voice research. *Journal of voice,* 31, 111.e1-111.e7.

MARTIN, S., DARNLEY, L., 2004. The Teaching Voice. 2nd Ed. (London: Whurr Publishers).

MARTINS, R., PEREIRA, E., HIDALGO, C., TAVARES, E., 2014. Voice disorders in teachers. a review. Journal of Voice 28: 716–724.

MARYN, Y. & WEENINK, D., 2015. Objective dysphonia measures in the program praat: smoothed cepstral peak prominence and acoustic voice quality index. Journal of voice, 29, 35–43.

MARYN, Y., YSENBAERT, F., ZAROWSKI, A. & VANSPAUWEN, R., 2017. Mobile communication devices, ambient noise, and acoustic voice measures. Journal of Voice. 31, 248e11-248e23.

MATHIESON, L., 2001. Greene and Mathieson's The Voice and its Disorders. (London: Whurr Publishers).

MORTON, V., & WATSON, D.R., 2001 The impact of impaired voice quality on children's ability to process spoken language. Logopedics Phoniatrics Vocology, 26, 17-25.

ROGERSON, J. and DODD, B., 2005 Is there an effect of dysphonic teachers' voices on children's processing of spoken language? Journal of Voice 19:47-60.

ROSENSTOCK, I. M., 1990. The health belief model: explaining health behavior through expectancies. In K. GLANZ, F. M. LEWIS, & B. K. RIMER (eds.), The Jossey-Bass health series. Health behavior and health education: Theory, research, and practice. (San Francisco, CA, US: Jossey-Bass), pp. 39-62.

SAWESI, S., RASHRASH, M., PHALAKORNKULE, K., CARPENTER, J.S., ET AL., 2016. The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. JMIR medical informatics. 4, e1

SCHAEFFLER, F., BECK, J., 2017. Monitoring Voice Condition Using Smartphones, in: Proceedings of the 10th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA). Presented at the MAVEBA, Firenze University Press, Florence, Italy.

SCHUNK, D.H., 1982. progress self-monitoring: effects on children’s self-efficacy and achievement. The Journal of Experimental Education. 51, 89–93.

SAUDER, C., BRETL, M. and EADIE, T., 2017. Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and Analysis of Dysphonia in Speech and Voice (ADSV). *Journal of Voice*, *31*(5), pp.557-566.

STRECHER, V.J., ROSENSTOCK, I.M., 1997. The Health Belief Model, in: Baum, A. (Ed.), Cambridge Handbook of Psychology, Health and Medicine. Cambridge University Press, pp. 255–261.

ULOZA, V., PADERVINSKIS, E., VEGIENE, A., PRIBUISIENE, R., ET AL., 2015. Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening. European Archives of Oto-Rhino-Laryngology, 272, 3391-3399.

VAN HOUTTE, E., CLAEYS, S., WUYTS, F., LIERDE, K.V., 2011. The impact of voice disorders among teachers: vocal complaints, treatment-seeking behavior, knowledge of vocal care, and voice-related absenteeism. Journal of Voice, 25, 570–575.

VERDOLINI, K. and RAMIG, L.O., 2001. Occupational risks for voice problems. *Logopedics Phoniatrics Vocology*, *26*(1), 37-46.

VERDOLINI, K., RAMIG, L.O., 2001. Review: occupational risks for voice problems. Logopedics Phoniatrics Vocology, 26, 37–46.

VOGEL, A.P., ROSEN, K.M., MORGAN, A.T. & REILLY, S., 2014. Comparability of modern recording devices for speech analysis: smartphone, landline, laptop, and hard disc recorder. Folia phoniatrica et logopaedica, 66, 244–250.

WEBB, T.L., JOSEPH, J., YARDLEY, L., MICHIE, S., 2010. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. Journal of Medical Internet Research. 12.

ZAMBON, F., MORETI, F., BEHLAU, M., 2014. Coping strategies in teachers with vocal complaint. Journal of Voice 28: 341–348.