The Suitability of iPhone Recordings for the Acoustic Measures of Speech and Voice Quality

Emily Lin, PhD.* and Jeremy Hornibrook, BSc, BMedSc, MBChB, FRACS**

*Department of Communication Disorders, University of Canterbury, Christchurch, New Zealand
**Department of Otolaryngology, Christchurch Hospital, Christchurch, New Zealand

Abstract

This study examined the quality of iPhone recordings for acoustic measurements of speech and voice quality. A selection of acoustic measures were extracted from voice samples recorded using the "voice memo" application in an iPhone and compared with those derived from signals directly digitized (DD) in a laptop via a 12-bit A/D converter. Participants were 11 healthy adults, including six females and five males, aged between 27 to 67 years (Mean = 41.8 years, SD = 16.7). The participant was asked to read the first six sentences of the "rainbow passage". In addition, two participants were asked to produce sustained vowels (/i/, /a/, and /u/) and a sentence ("We saw two cars") ten times. The simultaneously recorded iPhone and DD signals were analysed to derive 10 acoustic measures, including spectral tilt for the whole sentence and fundamental frequency (F0), percent jitter, percent shimmer, signal-to-noise ratio, amplitude of the first harmonic relative to that of the second harmonic, singing power ratio, and frequencies of the first and second formants (F1 and F2), and vowel space area for the vowel segment. A series of Pearson's correlation procedures revealed that measures from iPhone and DD signals were highly correlated. Findings of the vowel effect on the experimental measures obtained from iPhone signals were consistent with those from DD signals. However, the mean normalized absolute differences between measures from iPhone and DD signals are optimal (i.e., lower than 20%) only for F0, F1, and F2. These findings suggest that iPhone recordings are as adequate as other types of high quality digital recordings for acoustic measurements of voice quality but most voice measures from different digital recording systems are not directly comparable.

Introduction

The increasingly greater accessibility of multimedia-enabled mobile phones with advanced computing capability and connectivity necessitates an investigation on the suitability of auditory signals recorded via these portable devices for acoustic measurements of speech and voice. The Apple iPhone, for example, is a handheld wireless multifunctional phone with the capacity of recording and playing audio-visual signals and transmitting them via emailing and internet access. Since its first release in 2007, iPhone has been gaining much popularity amongst the public as well as positive reviews from medical professionals attesting its usefulness in processing medically related data (Luo, 2008). The recent technological advancement most relevant to voice clinicians is the increase of audio sampling rate from moderately low in earlier models (e.g., 8,000 Hz in first-generation iPhone) to relatively high in more recent models (e.g., 48,000 Hz in iPhone 3G, 3GS, and up). These latest iPhone models share many of the quality characteristics and capabilities of a portable non-compression voice recorder such as a minidisc recorder, which has been evaluated and considered suitable for voice perturbation analysis (Winholtz & Titze, 1998). With multifunctional capacity, open linkage to third-party applications, and high-quality voice recording, devices such as iPhone have a great potential for enhancing voice management not only by improving the efficiency and flexibility in voice recording for acoustic measurements of speech and voice quality but also by facilitating the application of an acoustic tracking or biofeedback device for voice training.

Methods

Participants and Participant's Task A total of 11 healthy adults, including six females and five males, were recruited as subjects. Participants aged between 27 to 67 years (Mean = 41.8 years, SD = 16.7). Four participants were native and seven were non-native English speakers. All participants were asked to read the first six sentences in the "rainbow passage" (Fairbanks, 1960), one sentence at a time. Additionally, two of the participants, Participants 10 and 11, were asked to read the sentence "We saw two cars" 10 times and sustain each of the isolated vowels, /i, a, u/, 10 times. Participant 10 was a 63-year-old female native speaker of American English and Participant 11 was a 32-year-old male non-native English speaker. For Participants 10 and 11, the order of the 30 sustained vowel productions (3 vowels X 10 trials) was randomized, with three sustained vowel productions followed by one sentence.

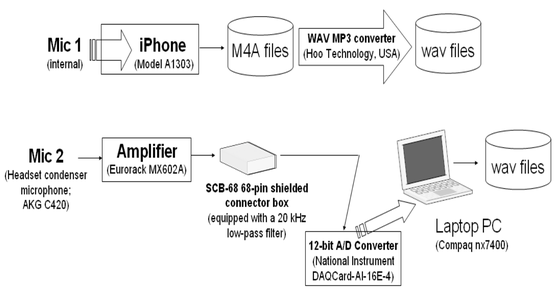
Procedure Each participant was seated in a sound-treated room, which was monitored to ensure that the ambient noise level did not exceed 30 dBA. The simultaneously recorded signals (iPhone vs. directly digitized) were saved in separate digital audio files.

Experimental Measures

I. Sentence-based:
- Spectral tilt (ST): amplitude difference between the highest spectral peak between 0 and 1 kHz and that between 1 and 5 kHz;
steeper ST = vocal hypofunction (Löfqvist, 1987; Mendoza, Munoz, & Valencia Naranjo, 1996)

II. Vowel-based (50-ms mid portions of the selected vowel embedded in the sentences)
1. Fundamental frequency (F0): affected by mass and stiffness, e.g.,
   -Edema (smokers): decreased F0 (Sorensen & Horii, 1982)
   -Voice patients have difficulties maintaining a constant pitch (Kotby, Titze, Saleh, & Berry, 1993)
   -Speaking F0 changes after treatment of functional voice (Roy & Taskco, 1994)
2. Perturbation measures (related to voice quality):
   -Percent jitter (%Jit): cycle-to-cycle frequency variation
   (Eskenazi, Childers, & Hicks, 1990; Dejonckere, Remacle, Fresnel-Ebaz, Woisard, Crevier-Buchman, & Millet, 1996; Wolfe & Martin, 1997; Bhuta, Patrick, & Garnett, 2004)
   -Percent shimmer (%Shim): cycle-to-cycle amplitude variation
   (Dejonckere et al., 1996; Wolfe & Martin, 1997; Bhuta et al., 2004)
   -Signal-to-noise ratio (SNR): energy ratio between periodic and aperiodic components
   more hoarse = higher %Jit (+) higher %Shim (+) lower SNR
   (Wolfe & Martin, 1997; Brockmann, Storck, Carding, & Drinnan, 2008)
3. Frequencies of Formants One and Two (F1 &F2): affected by tongue placement or vocal tract constriction.
   (Bradlow, Toretta, & Pisoni, 1996; Roy, Nissen, Dromey, & Sapir, 2009; Turner, Tjaden, & Weismer, 1995; Weismer, Jeng, Laures, Kent, & Kent, 2001)
4. Prominence of the first harmonic (H1-H2): amplitude difference between the first two harmonics
   greater H1-H2 = more breathy or thinner voice
   (Klatt & Klatt, 1990; Hillenbrand, Cleveland, & Erickson, 1994; de Krom, 1995; Hillenbrand & Houde, 1996; Stone, Cleveland, Sundberg, & Prokop, 2003)
5. Singing power ratio (SPR): the amplitude difference between the highest spectral peak between 0 to 2 kHz and that between 2 and 4 kHz;
   lower SPR = greater voice projection power (e.g., Omori, Kacker, Carroll, Riley, & Blaugrund, 1996)

Instrumentation **Two digital recording systems were employed, including iPhone (internal microphone placed at 13 cm away from mouth) and a direct digitization device (microphone at 5 cm). The acoustic signals directly digitized onto a laptop PC via a 12-bit A/D converter were saved as "WAV" files using a locally developed algorithm written in MATLAB 12 (The Mathworks, Inc.) installed in the laptop. The sampling rate was set at 44.1 kHz. The Adobe Audition 3.0 (Adobe, USA) was used for intensity normalization for all signal files. The TF32 acoustic analysis software (Milenkovic, 1987) was used to play back and process all normalized signals to extract the experimental measures.**



Results

Measures from iPhone and directly digitized (DD) signals were highly correlated for F1 (r = 0.98, n =33), F2 (r = 0.98, n = 33), F0 (r = 0.96, n = 33), %Shim (r = 0.81, n = 33), vowel space area (r = 0.94, n = 11), and SNR (r = 0.81, n = 33) and moderately high for H1-H2 (r = 0.77, n = 33), %Jit (r = 0.77, n = 33), SPR (r = 0.74, n = 33), and ST (r = 0.61, n = 66).

The descriptive statistics of the "normalized absolute difference [NAD = |(iPhone measure - DD measure) / DD measure| X 100] measure for each of the experimental measures are summarized in Table 2 for the three separate data sets (one for the "rainbow passage" production by all participants and two for the sustained vowel and sentence productions by Participants 10 and 11). As shown in Table 2, the mean inter-recorder NAD is consistently low (i.e., lower than 20%) for F0, F1, and F2, suggesting that these measures are least susceptible to the recorder effect and more comparable than other acoustic measures.

	N	Mean	SD	Min	Max
All Participants: "rainbow passage"					
F0	33	**3.13**	8.66	0.08	50.68
%Jit	33	42.94	74.74	0.00	396.23
%Shim	33	122.08	178.47	1.01	899.39
SNR	33	28.04	14.61	1.32	67.32
H1-H2	33	165.61	201.50	0.78	730.77
SPR	33	75.81	65.22	0.00	254.29
F1	33	**6.61**	6.52	0.62	27.11
F2	33	**5.35**	5.19	0.00	22.00
VSA	11	**19.88**	12.04	7.29	41.83
ST	66	52.34	20.64	2.79	93.01
Participant 10 (Female): sustained vowels and "We saw two cars."					
F0	60	**1.91**	0.99	0.89	6.22
%Jit	60	**14.03**	17.76	0.00	105.33
%Shim	60	158.67	71.87	11.98	360.47
SNR	60	26.61	11.01	5.00	50.23
H1-H2	60	92.55	88.64	8.24	492.31
SPR	60	43.95	20.69	0.00	85.71
F1	60	**10.85**	9.74	0.25	56.83
F2	60	**5.90**	5.82	0.23	31.47
Participant 11 (Male): sustained vowels and "We saw two cars."					
F0	60	**2.19**	6.40	0.53	50.93
%Jit	60	**14.18**	11.56	0.00	45.24
%Shim	60	40.23	26.83	1.78	135.37
SNR	60	**9.76**	8.33	0.00	50.30
H1-H2	60	302.43	529.40	44.74	2400.00
SPR	60	46.15	36.84	0.37	211.29
F1	60	**9.28**	8.54	0.37	34.98
F2	60	**6.73**	9.19	0.00	48.40

Table 2. The normalized absolute difference (NAD) between measures from signals simultaneously recorded with two recording systems (iPhone and direct digitization), including NAD for fundamental frequency (F0), percent jitter (%Jit), percent shimmer (%Shim), signal-to-noise ratio (SNR), dominance of Harmonic one (H1-H2), singing power ratio (SPR), frequencies of the first two formants (F1 and F2), vowel space area (VSA), and spectral tilt (ST). Mean NAD lower than 20% is boldfaced.

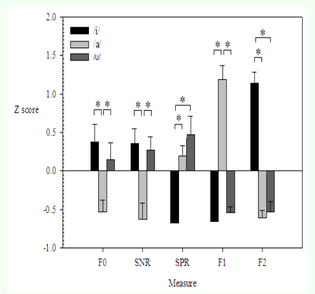


Figure 5. Means and standard errors of the standardized scores (z scores) of fundamental frequency (F0), signal-to-noise ratio (SNR), singing power ratio (SPR), and frequencies of the first two formants (F1 and F2) for three vowels (*/i/, /a/,* and */u/*) embedded in one of the "rainbow passage" sentence, with 22 tokens (11 participants X 2 recorders) in each vowel type. The asterisk ("*") indicates a significant inter-recorder difference.

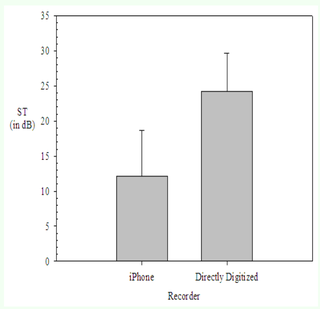


Figure 6. Means and standard deviations of spectral tilt (ST) measures for two types of recording (iPhone vs. direct digitization) of the "rainbow passage", with 66 tokens (11 participants X 6 sentences) in each recorder type. The asterisk ("*") indicates a significant inter-recorder difference.

Conclusion

In summary, the iPhone recording method was found to be compatible with the direct digitization method for acquiring voice samples for acoustic measurements of speech and voice quality. In particular, F0, F1, and F2 were found to yield minimal inter-recorder variations. The %Jit was found to be less susceptible to the recorder effect than %Shim. Spectral measures involving measurements at the low frequency band show greater inter-recorder variations. Although the inter-recorder reliabilities are generally high, noise introduced by the circuitry of the recording systems, including the difference in the sensitivity of the microphone used, may have resulted in the high absolute inter-recorder difference for some of the experimental measures, suggesting that most acoustic measures of speech and voice quality should be obtained from the same recording system for meaningful comparisons. In other words, a direct comparison between measures from different digital recording systems for voice evaluation is not indicated for most acoustic measures.

	Recorder Effect		Vowel Effect		Recorder by Vowel	
	F(1, 10)	p	F(2, 20)	p	F(2, 20)	p
F0	3.638	0.086	25.416	< 0.001*	0.792	0.466
%Jit	0.982	0.345	0.957	0.401	0.231	0.796
%Shim	23.237	< 0.001*	2.019	0.159	3.862	0.038
SNR	47.975	< 0.001*	7.110	0.005*	0.652	0.535
H1-H2	30.948	< 0.001*	1.102	0.351	4.058	0.033
SPR	51.682	< 0.001*	12.499	< 0.001*	5.056	0.017
F1	0.037	0.851	62.446	< 0.001*	0.433	0.655
F2	0.787	0.396	77.449	< 0.001*	1.454	0.257

* p = < 0.005

Table 1. Results from a series of two-way (recorder by vowel) RM ANOVAs for eight vowel-based measures obtained from one "rainbow passage" sentence (i.e., /i/ from "these", /a/ from "arch", and /u/ "two") produced by all 11 participants.

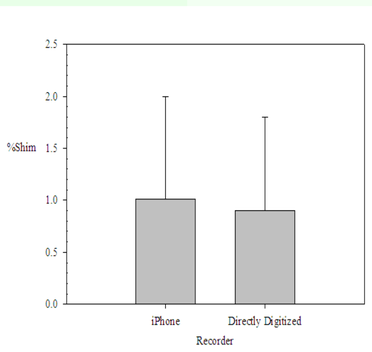


Figure 1. Means and standard deviations of percent shimmer (%Shim) measures for two types of recording (iPhone vs. direct digitization) of vowels embedded in one of the "rainbow passage" sentences, with 33 tokens (11 participants X 3 vowels) in each recorder type.

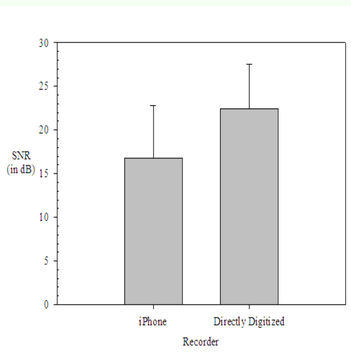


Figure 2. Means and standard deviations of signal-to-noise ratio (SNR) measures for two types of recording (iPhone vs. direct digitization) of vowels embedded in one of the "rainbow passage" sentences, with 33 tokens (11 participants X vowel in each recorder type.

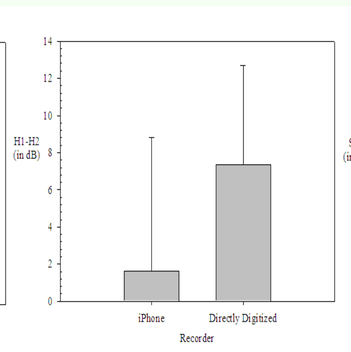


Figure 3. Means and standard deviations of H1 dominance (H1-H2) measures for two types of recording (iPhone vs. direct digitization) of vowels embedded in one of the "rainbow passage" sentences, with 33 tokens (11 participants X 3 vowels) in each recorder type.

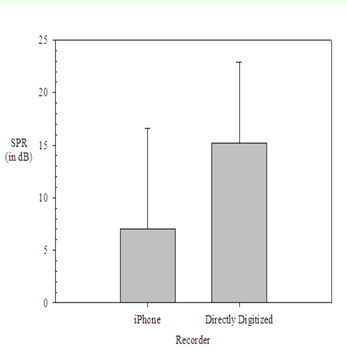


Figure 4. Means and standard deviations of singing power ratio (SPR) measures for two types of recording (iPhone vs. direct digitization) of vowels embedded in one of the "rainbow passage" sentences, with 33 tokens (11 participants X 3 vowels) in each recorder type.

References

Bhuta, T., Patrick, L., & Garnett, J. D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice, 18,* 299-304.
Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech I: global and fine grained acoustic-phonetic talker characteristics. *Speech Communication, 20,* 255-272.
Brockmann, M., Storck, C., Carding, P. N., & Drinnan, M. J. (2008). Voice loudness and gender effect on jitter and shimmer in healthy adults. *Journal of Speech, Language, and Hearing Research*, *51*, 1152-1160.
de Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research, 38,* 794-811.
Dejonckere, P. H., Remacle, M., Fresnel-Ebaz, E., Woisard, V., Crevier-Buchman, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. Rev Laryngol Otol Rhinol (Bord), 117, 219-224.
Eskenazi, L., Childers, D. G., & Hicks, D. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33,* 298-306.
Fairbanks, G. (1960). Voice and Articulation Drillbook. New York: Harper & Row.
Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy voice quality. *Journal of Speech and Hearing Research, 37,* 769-778.
Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *Journal of Speech and Hearing Research, 39,* 311-321.
Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87,* 820-854.
Kotby, M. N., Titze, I. R., Saleh, M. M., & Berry, D. A. (1993). Fundamental frequency stability in functional dysphonia. *Acta Otolaryngologica, 113,* 4439-444.
Löfqvist, A., & Mandersson, B. (1987). Long-time average spectrum of speech and voice signals. *Folia Phoniatrica, 21,* 221-229.
Luo, J. S. (2008). Medical Applications for the iPhone. *Primary Psychiatry, 16,* 14-16.
Mendoza, E., Munoz, J., Valencia Naranjo, N. (1996). The long-term average spectrum as a measure of voice stability. *Folia Phoniatrica et* Logopaedica, *48,* 57–64.
Omori, K., Kacker, A., Carroll, L. M., Reiley, W. D., & Blaugrund, S. M. (1996). Singing power ratio: quantitative evaluation of singing voice quality. *Journal of Voice, 10,* 228-235.
Roy, N., Nissen, S. L., Dromey, C., & Sapir, S. (2009). Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlagyngeal therapy. *Journal of Communication Disorders, 42,* 124-135.
Roy, N., & Tasko, S. M. (1994). Speaking fundamental frequency (SFF) changes following successful management of functional dysphonia. Journal of Speech-Language Pathology and Audiology, 18, 115-120.
Sorensen, D., & Horii, Y. (1982). Cigarette smoking and voice fundamental frequency. *Journal of Communication Disorders, 15,* 135-144.
Stone, Jr., R. E. (ED), Cleveland, T. F., Sundberg, P. J., & Prokop, J. (2003). Aerodynamic and acoustical measures of speech, operative, and Broadway vocal styles in a professional female singer. *Journal of Voice, 17,* 283-297.
Turner, G. S., Tjaden, K., & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amytrophic lateral sclerosis. *Journal of Speech and Hearing Research, 38,* 1001-1013.
Weismer, G., Jeng, J-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica, 53,* 1-18.
Winholtz, W. S., & Titze, I. R. (1998). Suitability of Minidisc (MD) Recordings for Voice Perturbation Analysis. *Journal of Voice, 12,* 138-142.
Wolfe, V. & Martin, D. (1997). Acoustic correlates of dysphonia: type and severity. *Journal of Communication Disorders, 30,* 403-416.