Maximal Ambient Noise Levels and Type of Voice Material Required for Valid Use of Smartphones in Clinical Voice Research

*Jean Lebacq, †Jean Schoentgen, ‡Giovanna Cantarella, §Franz Thomas Bruss, ¶Claudia Manfredi, and #Philippe DeJonckere, *†§#*Brussels, Belgium, and* ‡*Milano and* ¶*Firenze, Italy*

Summary: Purpose. Smartphone technology provides new opportunities for recording standardized voice samples of patients and transmitting the audio files to the voice laboratory. This drastically improves the achievement of baseline designs, used in research on efficiency of voice treatments. However, the basic requirement is the suitability of smartphones for recording and digitizing pathologic voices (mainly characterized by period perturbations and noise) without significant distortion. In a previous article, this was tested using realistic synthesized deviant voice samples (/a:/) with three precisely known levels of jitter and of noise in all combinations. High correlations were found between jitter and noise to harmonics ratio measured in (1) recordings via smartphones, (2) direct microphone recordings, and (3) sound files generated by the synthesizer. In the present work, similar experiments were performed (1) in the presence of increasing levels of ambient noise and (2) using synthetic deviant voice samples (/a:/) as well as synthetic voice material simulating a deviant short voiced utterance (/aiuaiuaiu/).

Results. Ambient noise levels up to 50 dB_A are acceptable. However, signal processing occurs in some smartphones, and this significantly affects estimates of jitter and noise to harmonics ratio when formant changes are introduced in analogy with running speech. The conclusion is that voice material must provisionally be limited to a sustained /a/.

Key Words: Smartphone–Dysphonia–Recording–Noise–Acoustics.

INTRODUCTION

In recent years, the use of smartphones and web-based systems for clinical applications has gained increasing scientific interest, thanks to developments in digital technology, making these devices suitable for recording acoustic signals and transmitting the digitized audio files.[1,2] Specifically, as far as voice is concerned, digital technology enables a decisive improvement in audio quality compared with telephone transmission. Smartphones are pocket-sized highly mobile computers; they contain the required interfaces for easy voice recording at home or on site. Transmission can be web-based and is no longer restricted by bandwidth limitations of the telephonic pathway.

For general information about potential use of smartphones in pathologic voice research, particularly to help in carrying out single-case designs and multiple baseline designs, the reader is referred to our previous paper.[1]

In a first experiment, we demonstrated the reliability of smartphones with regard to quality of recordings over a wide range of degrees of deviance (perturbation and additive noise) and in the male and female ranges of fundamental frequency (F0) values. The comparison was carried out using realistic synthesized voice signals (sustained /a:/ altered by three levels of jitter and three levels of noise, the two basic acoustic voice quality parameters), which guarantee exact knowledge of reference values for voice quality parameters. The absence of significant distortion by the smartphone (during recording or data processing) is the basic requirement for the use of such devices in the transmission of audio signals from the patient to the voice laboratory for analysis of deviant voice quality. Furthermore, it was assumed that all types of smartphones were likely to be adequate, and we selected two smartphones at the extremes of the commercially available price range. However, our experiments were conducted in a laboratory setting, that is, in a soundproof booth. It was mentioned that for clinical purposes, the sound pressure level of the ambient noise should be controlled while recording voice samples. This is made possible by current smartphone technology (sound measurement applications or "apps"), although so far only some "apps" are really accurate.[3] As regards noise, very recently, Maryn et al[4] found that chains of dysphonic sustained vowels and continuous speech recorded by means of mobile communication devices (two tablet computers and three smartphones) were significantly impacted by ambient noise. They concluded that, due to combined differences in hardware, software, and ambient sound conditions, acoustic voice quality measures may differ between recording systems.

For a voice laboratory setting with direct recording, Deliyski et al[5] found that a level of noise in the acoustic environment of <46 dB was to be recommended, and <58 dB was acceptable. However, practical limits of tolerable ambient noise intensity values for the application considered here—that is, a patient recording his or her voice at home or at the workplace for sending to the voice clinic—are not accurately known. Furthermore, in our first paper,[1] only sustained /a:/ was used as voice material. It is thus worthwhile to assess the extent to which smartphones possibly distort synthesized samples comparable with natural voice

Accepted for publication February 24, 2017.

From the *Neurosciences Institute, University of Louvain, Brussels, Belgium; †B.E.A.M.S. Department, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium; ‡Otolaryngology Department, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy; §Department of Mathematics, Université Libre de Bruxelles, Brussels, Belgium; ¶Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy; and the #Neurosciences, University of Leuven (KULeuven) and FEDRIS (Federal Agency for Occupational Risks), Brussels, Belgium.

Address correspondence and reprint requests to Philippe DeJonckere, Neurosciences, University of Leuven and FEDRIS (Federal Agency for Occupational Risks), Brussels, Belgium. E-mail: philippe.dejonckere@kuleuven.be

Journal of Voice, Vol. 31, No. 5, pp. 550–556
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2017.02.017

productions, for which different reference values for F0 perturbation and noise to harmonics (N:H) ratio are exactly known.

In the first part of this work, the same synthetic voice signals as those used previously (sustained /a:/) were recorded simultaneously by two smartphones in the presence of stepwise increasing levels of ambient noise. The smartphone audio files were sent by e-mail and analyzed using *Praat*. The results of jitter % and N:H ratio could then be compared with those of the direct recording (microphone without added ambient noise) and with those of the direct analysis of the original signal (from computer to computer). In the second part of the present experiments, the sustained /a:/ was replaced by relatively realistic synthetic test signals consisting of a sequence of three repetitions of the fragment /aiu/ with intonation and formant changes, simulating a short voiced utterance, as frequently used in clinical practice.[6,7] Again, three levels of jitter and three levels of noise were introduced in the signals, and the same comparisons were made: the results of jitter % and N:H ratio in the audio files of the smartphones were compared with those of the direct recording (microphone without added ambient noise) and with those of the direct analysis of the original signal (from computer to computer).

MATERIALS AND METHODS

Synthesizer

The synthesizer uses a model of the glottal area based on a polynomial distortion function that transforms two excitatory harmonic functions into the desired waveform.[8,9] The polynomial coefficients are obtained by constant, linear, and invertible transforms of the Fourier series coefficients of Klatt's template cycle that is asymmetric and skewed to the right.[9] This waveform is in fact typical for the glottal area cycle, allowing a maximal glottal area of 0.2 cm^2. The discrete phase increment of the harmonic excitation functions evolves proportionally to the instantaneous vocal frequency F0. The sampling frequency is set at 200 kHz to simulate voices, the frequency modulation of which is of the order of 1% of the fundamental frequency F0, thus requiring high temporal resolution. The harmonic excitation functions are low-pass filtered and down-sampled to 50 kHz before their transformation by the distortion function. To simulate voice perturbations as jitter, phase, or amplitude fluctuations, disturbances of the harmonic excitation functions are introduced. Specifically, jitter is simulated with a model based on low-pass filtered white noise of adjustable size. The noisy signal is obtained by adding pulsatile or aspiration noise to the clean flow rate. Pulsatile noise simulates additive noise due to turbulent airflow in the vicinity of the glottis and its size evolves proportionally to the glottal volume velocity. It is obtained by low-pass filtering white Gaussian noise, the samples of which are multiplied by the clean glottal volume velocity. Low-pass filtering is performed with linear second-order filters. Additive noise is measured as the N:H ratio of the clean volume velocity signal at the glottis relative to the noise. The synthesizer also generates varying levels of shimmer via modulation distortion by the vocal tract transfer function, which automatically increases when jitter increases. Indeed, jitter and shimmer are acoustically linked to each other. Once the glottal area has been obtained, the glottal flow rate is calculated numerically via the interactive voice source model proposed by Rothenberg, which takes into account the glottal impedance and tract load.[10] Each formant is modeled with a second-order bandpass filter. The vocal tract transfer function is obtained by cascading several second-order filters, including the nasal and tracheal formants, the frequencies and bandwidths of which are fixed.[11] The bandwidths of the vocal tract formants are calculated via the formant frequencies.[12] The first three formant frequency values have been equal to 640 Hz, 1212 Hz, and 2254 Hz for [a], 230 Hz, 2000 Hz, and 3000 Hz for [i], as well as 298 Hz, 730 Hz, and 2172 Hz for [u].The radiation at the lips is simulated via a high-pass filter. The signals are then normalized, dithered, quantized, converted into ".wav" format, and stored on the computer hard disk.

Synthetic voices

The synthesized deviant voice samples consisted of sustained /a:/ samples at a median F0 of 120 Hz and 200 Hz, of 2 seconds of duration, with a slight falling and rising intonation, and with three levels of jitter: 0.9%, 2.8%, and 4.5%. For each level of jitter, three levels of added noise were considered: the lowest level corresponding to a volume velocity to noise ratio at the glottis equal to 17 dB, the intermediate level equal to 23 dB, and the highest level equal to 90 dB. These true levels correspond to numeric N:H ratios obtained via *Praat* equal to 0.2, 0.6, and 0.8, respectively. Perceptually, they correspond to common dysphonic patients' voices, from slightly to severely deviant, rough as well as breathy.

In the second part of the present work, the sustained /a:/ was replaced by realistic synthetic test signals consisting of a sequence of three repetitions of the fragment /aiu/ with slight intonation and formant changes from one stimulus to the next, simulating a short voiced utterance of 4.4 seconds.[13] The jitter and noise levels were similar to those for the /a:/. An example of the spectrogram of an /aiu/ utterance obtained with the *Praat* program is given in Figure 1 (F0 median: 120 Hz; jitter: 3.35%; N:H ratio: 0.28).

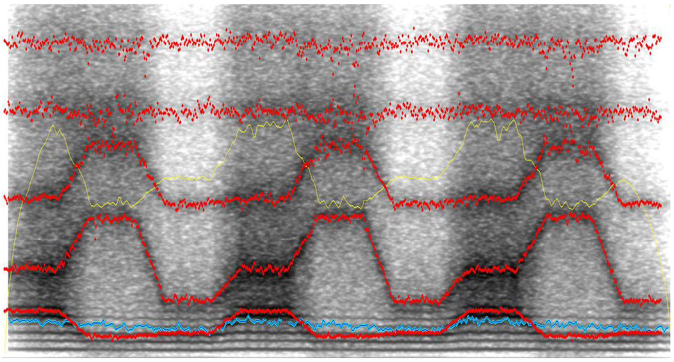

FIGURE 1. Example of spectrogram (0–5 KHz) of a synthetic voice sample (3×/aiu/). Duration: 4.4 seconds; average F0: 120 Hz; jitter %: 3.35%; noise to harmonics ratio: 0.28. *Blue dots*: F0 (total scale = 0–500 Hz linear; slight intonation). *Red dots*: formant locations (total scale = 0–5000 Hz linear). *Yellow line*: intensity (total scale 50–100 dB linear). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)