Smartphones Offer New Opportunities in Clinical Voice Research

*C. Manfredi, †J. Lebacq, ‡G. Cantarella, §J. Schoentgen, *S. Orlandi, *A. Bandini, and ¶P.H. DeJonckere, *Firenze, Italy, †¶Brussels, Belgium, ‡Milano, Italy, and §Belgium

Summary: Smartphone technology provides new opportunities for recording standardized voice samples of patients and sending the files by e-mail to the voice laboratory. This drastically improves the collection of baseline data, as 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 this experiment, two smartphones (a very inexpensive one and a high-level one) were tested and compared with direct microphone recordings in a soundproof room. The voice stimuli consisted in synthesized deviant voice samples (median of fundamental frequency: 120 and 200 Hz) with three levels of jitter and three levels of added noise. All voice samples were analyzed using *PRAAT* software. The results show high correlations between jitter, shimmer, and noise-to-harmonics ratio measured on the recordings *via* both smartphones, the microphone, and measured directly on the sound files from the synthesizer. Smartphones thus appear adequate for reliable recording and digitizing of pathologic voices.
Key Words: smartphone–dysphonia–baseline design–voice assessment–synthetic voices.

INTRODUCTION

In recent years, the use of smartphones for clinical applications has gained increasing scientific interest thanks to advancements in digital technology, making these portable devices suitable for recording acoustic signals and transmitting the digitized audio files *via* e-mail. In the field of voice, digital technology enables a decisive improvement in audio quality compared with telephone transmission. For example, Uloza et al[1] explored the potential role of smartphone (Samsung Galaxy Note 3, Samsung, Daegu, South Korea) recordings in screening for laryngeal diseases and for subsequent referral of selected individuals for medical examination and visualization of the larynx, thus improving early diagnosis of laryngeal diseases. Guidi et al[2] showed that the quality of audio acquisitions from Samsung I9300 Galaxy S III smartphones was adequate for the investigation of fundamental frequency (Fo) features of running speech (using an *ad hoc* Android application) in subjects with bipolar mood disorders. Also, Lin et al[3] found that iPhone (Apple A1303, Apple Inc., Cupertino, CA, USA) recordings are suitable for acoustic measurements of voice quality.

If the validity of voice analyses achieved on smartphone recordings (including the most inexpensive ones) sent *via* e-mail is confirmed in dysphonic patients (with a wide range of deviances), new opportunities are opened for clinical voice research: repeated measurements over time become possible without multiple visits to the voice laboratory. The patient can record his/her own voice (according to a standardized protocol) with his/her smartphone and send the audio file to the voice laboratory by e-mail. Obviously, the basic requirement is the suitability of smartphones for recording and digitizing pathologic voices (that are mainly characterized by period perturbations and noise) without significant distortion. Thus, the investigation about the reliability of acoustic voice parameters obtained using smartphone microphones is of particular clinical interest.[1]

Repeated measurements are the basis of single-subject and baseline designs. Multiple baseline designs are widely recognized in many areas of research as easily implemented, highly sensitive, and internally valid. Many areas of research in which randomized-group designs with blinding are disqualified by practical or ethical considerations are easily investigated using at least one of the variants of the multiple-baseline design.[4] This applies in particular to clinical research in the field of voice pathology, all the more so as it has been shown that short-time variability of acoustic parameters of voice quality is far from negligible.[5] Therefore, measuring the dependent variable at single time points (pre- and posttreatment measures) may provide a biased estimate.[6] In a baseline design, voice quality parameters of a single subject are measured repeatedly and plotted as a function of time to establish a baseline (Figure 1). At a given time, a treatment is initiated while measurements are continued. If a change occurs, the posttreatment measurements determine a new baseline, the level of which will differ from the previous baseline. The term "multiple" refers to the fact that each patient included in the study has a different starting baseline, depending on the degree of severity of his/her dysphonia. Figure 1 illustrates this approach: in patient 1, after a set of eight consecutive measurements (showing spontaneous random variations) of a given parameter (eg, jitter % in a sustained /a:/), the treatment (eg, phonosurgery) occurs (fat arrow). An efficient treatment is expected to show a downward shift of the baseline (less jitter). Patient 2, included in the study 1 month later than patient 1, presents with a less deviant voice than patient 1. After five pretreatment measurements, the treatment is given, and again an improvement is noticed in the posttreatment baseline. In the case of a treatment of long duration (eg, voice therapy or antireflux medication), one may expect—in case of success—a downward slope of the regression line determined by the

Accepted for publication December 30, 2015.
From the *Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy; †Neurosciences Institute, University of Louvain, B-1200 Brussels, Belgium; ‡Otolaryngology Department, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy; §Department of Signals, Images and Acoustics, Faculty of Applied Sciences, Université Libre de Bruxelles, Belgium; and the ¶Neurosciences, University of Leuven & Federal Institute for Occupational Diseases, Brussels, Belgium.
Address correspondence and reprint requests to Philippe H. DeJonckere, Neurosciences, University of Leuven & Federal Institute for Occupational Diseases, Brussels, Belgium. E-mail: philippe.dejonckere@med.kuleuven.be
Journal of Voice, Vol. 31, No. 1, pp. 111.e1–111.e7
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2015.12.020

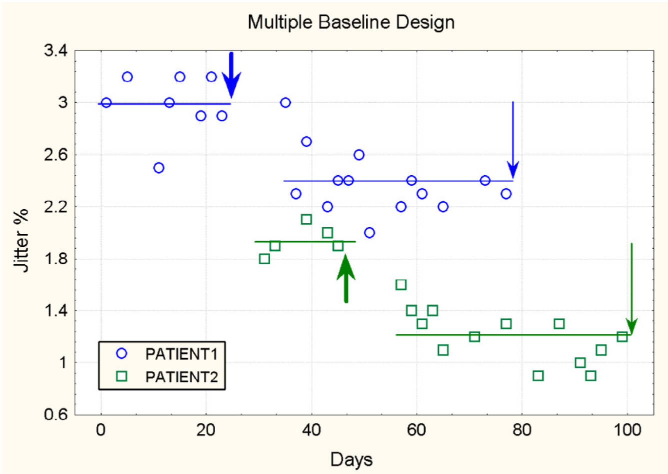

FIGURE 1. Schematic graph with data of two patients consecutively included in a multiple baseline design. In patient 1, eight measurements of a given voice quality parameter (here jitter %) precede the treatment (eg, phonosurgery; *fat arrow*) and define the "pre baseline." After treatment, 14 additional measurements define the "post baseline", which is shifted downward (*tiny arrow*), indicating a reduction of the jitter %. Notice that some "pre" measurement points indicate lower jitter values than some "post" measurement points. Patient 2 is included in the study 1 month after patient 1 and has a less deviant voice than patient 1. Here, treatment occurs after five measurements, and there are 13 "post" measurements.

posttreatment measurement points. Adequate statistical approaches are available to deal with such data.[7]

In this study, two smartphones (a very basic, inexpensive one and a high-level one) are tested and compared with direct microphone recordings in a soundproof room. The comparison concerns the main basic acoustical parameters of clinical interest: cycle-to-cycle perturbation (jitter and shimmer) and noise.[5] Of prime importance is also the voice material. In this experiment, we used synthesized deviant voices that have the advantage of an exact calibration of period perturbation parameters as well as of noise. Such samples have been used in checking the adequacy of voice analysis programs[8–12] and are used here to test the reliability of the two smartphones.

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.[13,14] The polynomial coefficients are obtained by constant, linear, and invertible transforms of the Fourier series coefficients of the Klatt template cycle that is asymmetric and skewed to the right.[14] This waveform is in fact typical for the glottal area cycle, allowing a maximal glottal area of 0.2 cm^2. The discrete phase of the harmonic excitation functions changes from iteration to iteration with a step defined by the inverse of the sampling frequency. The sampling frequency is set at 200 kHz to simulate voices, the frequency modulation of which is of the order of 1% of the Fo, 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, and/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 noise 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 noise-to-harmonics ratio (NHR) of the clean volume velocity signal at the glottis relative to the noise. The synthesizer also generates varying levels of shimmer that automatically increases when jitter increases. Indeed, jitter and shimmer are physiologically linked with each other. Once the glottal area has been obtained, the flow rate is simulated taking into account a model of the glottal impedance and tract load.[15] 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.[16] The bandwidths of the vocal tract formants are calculated *via* the formant frequencies.[17] The radiation at the lips is also taken into account *via* a high-pass filter. The signal is then normalized, dithered, quantized, converted into ".wav" format, and stored on the computer hard disk.

Synthetic voices

The synthesized deviant voice samples consist of sustained /a:/ utterances at a median Fo = 120 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 additive noise are considered, with a flow rate to aspiration noise of respectively 97.5 dB, 23.8 dB, and 17.6 dB. They perceptually correspond to usual dysphonic patients' voices, from slightly to severely deviant, rough, as well as breathy. Shimmer increases from about 7% up to 23% with increasing jitter.

Smartphones

The devices were selected at the extremes of the commercial price range. The more expensive one is an HTC One (named hereafter Smart1), the basic, inexpensive one is a Wiko model CINK SLIM2 (named Smart2). Relevant specifications are given in Table 1. Price ratio is 1/15.

Microphone

The microphone is a Sennheiser model MD421U (Wedemark, Germany) (frequency response 30–17.000 Hz) commonly used in the voice laboratory to make recordings of voice patients.

Amplifier and loudspeaker

A Bowers & Wilkins model CM1 (Worthing, UK) loudspeaker was used. Its frequency response is flat ± 1.5 dB between 50 Hz and 20 kHz. It was driven by a Yamaha UK YHT-380 amplifier (Hamamatsu, Japan). The frequency response of the amplifier is