Integrating Voice Evaluation: Correlation Between Acoustic and Audio-Perceptual Measures

*,†,‡Susana Vaz Freitas, §Pedro Melo Pestana, ||Vítor Almeida, and ||Aníbal Ferreira, *†‡§||*Porto, Portugal*

Summary: Objectives/Hypothesis. This article aims to establish correlations between acoustic and audio-perceptual measures using the GRBAS scale with respect to four different voice analysis software programs.

Study Design. Exploratory, transversal.

Methods. A total of 90 voice records were collected and analyzed with the *Dr. Speech* (Tiger Electronics, Seattle, WA), *Multidimensional Voice Program* (Kay Elemetrics, NJ, USA), *PRAAT* (University of Amsterdam, The Netherlands), and *Voice Studio* (Seegnal, Oporto, Portugal) software programs. The acoustic measures were correlated to the audio-perceptual parameters of the GRBAS and rated by 10 experts.

Results. The predictive value of the acoustic measurements related to the audio-perceptual parameters exhibited magnitudes ranging from weak ($R^2_a = 0.17$) to moderate ($R^2_a = 0.71$). The parameter exhibiting the highest correlation magnitude is B (Breathiness), whereas the weaker correlation magnitudes were found to be for A (Asthenia) and S (Strain). The acoustic measures with stronger predictive values were local Shimmer, harmonics-to-noise ratio, APQ5 shimmer, and PPQ5 jitter, with different magnitudes for each one of the studied software programs.

Conclusions. Some acoustic measures are pointed as significant predictors of GRBAS parameters, but they differ among software programs. B (Breathiness) was the parameter exhibiting the highest correlation magnitude.

Key Words: Voice–Assessment–Audio-perceptual–Acoustic–Correlation.

INTRODUCTION

The audio-perceptual evaluation is considered to be the gold standard by some researchers, especially those who use it in clinical routine and see it as the preferred option in relation to others,[1–5] although consistency within and between the judges is from mild to moderate.[6–9] Owing to its auditory component, this phenomenon essentially depends on the training,[10,11] type of stimulus,[12–14] task instruction,[15] and judge experience.[4,16–18] The audio-perceptual evaluation is a quick and noninvasive assessment, easy to use, and does not need any form of electronic equipment.[19] Speech pathologist is able to inform others, including the patient, about the results of the evaluation and the voice therapy, in an accurate and meaningful way,[20,21] as well as support the therapy goals and follow therapy efficiency.[8,17,21,22]

In the past two decades, the nonlinear dynamics of the Acoustic Theory of Speech Production has introduced a new perspective into the analysis of systems, which evolve over time and are sensitive to initial conditions. The voice quality objective measures[5,23–25] are likely to be affected by confusion variables as: signal recording conditions, hardware and software specifications, the protocol of data capture and analysis, individual variability (acoustic and aerodynamic), as well as the severity and type of vocal disturbance.[1,3,13,26–40]

Acoustic method allows the integration of data with the audio-perceptual assessment and the physiology of voice production because it can specify the process of voice production through indirect estimation of vibratory patterns of the vocal folds, of the supraglottic vocal tract format, and of its modifications.[24,41] Different acoustic parameters are used—periodicity, amplitude, duration, and spectral composition[42,43]—to characterize the physical aspects of voice in several domains (time, frequency, and intensity), as well as complex measures that mingle these domains.[23,24,44–46]

From the first studies of voice analysis in the 1990s, various voice analysis tools (based on nonlinear dynamics) have been applied.[47–50]

Signal processing is an enabling technology that provides a large number of functionalities at the level of software and hardware. In the area of voice quality evaluation, it also constitutes the principle that underlies the operation of equipment and software of great interest,[51,52] for example, facilitating the task of analysis and classification of a voice (disturbed or not) throughout the various phases of the therapeutic process.[15,38,53]

The spoken voice involves, probably, the most elaborate system of human communication, so it is quite understandable the difficulty or even impossibility to use a single method to assess comprehensively and accurately the voice quality or its deterioration.[1,39,43]

In this context, it should be noticed that performing a multifactorial analysis plays a very important role allowing a broad, appropriate, and effective knowledge about the laryngeal function and voice quality. We point out that a certain type of assessment cannot replace another of a different nature—they all complement each other and are constructive in the therapeutic process.[36,38,54–60]

The correlation between the results of acoustic and audio-perceptual assessment is neither obvious nor direct, being still a topic of research and debate, which confirms the classical

Accepted for publication August 12, 2014.

From the *Serviço de Otorrinolaringologia, Centro Hospitalar do Porto, Largo Prof. Abel Salazar, Porto, Portugal; †Speech Therapy Department—Faculty of Health Sciences, Universidade Fernando Pessoa, Porto, Portugal; ‡Biomedical Engineering Department—Faculty of Engineering, Universidade do Porto, Porto, Portugal; §Unidade de Cuidados Continuados de Longa Duração e Manutenção da Gelfa, Vila Praia de Âncora, Portugal; and the ||Electric and Computer Engineering Department—Faculty of Engineering, Universidade do Porto, Rua Doutor Roberto Frias, Porto, Portugal.

Address correspondence and reprint requests to Susana Vaz Freitas, Serviço de Otorrinolaringologia, Centro Hospitalar do Porto, Largo Prof. Abel Salazar, 4099-001 Porto, Portugal. E-mail: svazfreitas@gmail.com

Journal of Voice, Vol. 29, No. 3, pp. 390.e1-390.e7
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.08.007

difficulty that exists in expressing the human auditory acuity, using mathematical models.[6,39,52,56,61–63]

Several researchers investigated the relationship between isolated acoustic measures and audio-perceptual evaluation. Their results were largely inconclusive because they revealed no statistically significant correlations,[3,4,36,61,64–69] which indicates that the analysis of isolated parameters does not describe the vocal behavior and should be avoided.[18,23,63,70–73] It should be noticed that acoustic analysis applied to extremely disturbed voice signals, specifically to the ones not included in type I suggested by Titze[32] is controversial. Titze[32] published a document based on the "Workshop on Acoustic Voice Analyses" in which he suggests the existence of three types of voice signals. In the present study the authors analyze voices that belong to types I and II. This allows the measure of fundamental frequency (F0), jitter, shimmer, and harmonic-to-noise ratio (HNR). It depends on the quality of the recording and the pre-processor method, with selection of the most stable parts of the signal—this was attained to develop the voices database used in this study.[32,74–77]

This study differs from others as it is one of the few that investigated directly the relation between a set of acoustic measures (from four common acoustic programs) and the audio-perceptual correlations.

The aim of this article is to present and discuss the magnitude of the correlations that were found between acoustic measures and audio-perceptual parameters of the GRBAS scale, as provided by the four types of software programs of voice signal analysis.

METHODS

Voice samples

A total of 90 voices have been selected from a database, after obtaining the authorization of the Ethics Committee of the Centro Hospitalar do Porto, in Portugal. Among these 90 voices, 20 were considered normal and 70 presented some degree of disturbance (in different degrees of severity, from mild to severe) as considered by the Speech and Language Pathologist who followed the cases.

The gender distribution is as follows: 28% of male voices ($n = 25$) and 72% of female voices ($n = 65$). This distribution corresponds to the representation of each of these genders in the database. All cases are adults, older than 18 years.

Voice signal recordings

All of the voice recordings followed the same protocol. The selected sampling frequency was 44100 Hz, with a resolution of 16 bits per sample, using a desktop microphone Philips SBC ME 400 (Amsterdam, The Netherlands), unidirectional (cardioid), in a room with a noise level less than 40-dB sound pressure level, although not acoustically treated. The distance from the microphone to the mouth was fixed at 10 cm, and the patient was asked—after illustration—to produce a sustained and comfortable vowel (a),[15,66,75,78–80] for at least 5 seconds,[15,80,81] in two attempts. The last one, with the speaker standing up, was the one used for the study. These recordings are part of the routine vocal assessment in the ENT appointments of the Centro Hospitalar do Porto, and were collected with the *Dr. Speech* software, version 4.0 (Tiger Electronics). The signal resulting from the second second of each sample was analyzed,[40,66,78,80,81] segmented, and the most stable part of the segment was used.

Acoustic analysis

The 90 segmented voice samples have been evaluated and classified according to the following acoustic parameters, namely Mean F0, standard deviation (SD) F0, local jitter, PPQ5 shimmer, local shimmer, APQ5 shimmer, HNR, as provided by four acoustic analysis software programs (open source [*PRAAT*]) and commercialized [*Dr. Speech* from Tiger Electronics, *Voice Studio* from Seegnal, and *Multidimensional Voice Program* (MDVP) from Kay Elemetrics]).

Because the samples were collected using a sampling frequency of 44 100 Hz, this was maintained for the software *Dr. Speech* and *MDVP*, and it was adjusted to 22 050 Hz to facilitate the analysis by the software *PRAAT* and *Voice Studio*. The final version of the samples implied pre-processing using the sampling frequency set as mentioned previously, for the most stable parts of each sample.

Audio-perceptual scale

The audio-perceptual scale used was the GRBAS.[82] The audio-perceptual evaluation—of a database consisting of 100 voices and including 10% repetitions—was presented on a Web page, in a random order, heard and rated by a panel of 10 judges. The Web page was presented as a form for each voice that had to be assessed, with the help of a system of preallocation of the audio file to be heard and a set of five interactive and adjustable cursors, one for each GRBAS parameter. These cursors used a visual analog scale, depicted as a 10-cm ruler (the more on the right, the stronger the voice disturbance), so that the evaluation of each parameter is closer to the analog procedure. A written definition of each parameter of the scale was provided to judges and was considered as an "anchor" or written clue to the assessment.

The final response was automatically stored in an SQL database, specifically built for this purpose. The results were obtained between February and May 2011.

The results of the intra- and inter-judges consistency were analyzed and discussed in a recent article,[83] which used the same set of data.

Statistical analysis

The statistical analysis used the computer program *Statistical Package for the Social Sciences*—IBM SPSS for Windows, version 19.0 (SPSS Inc., Chicago, IL).

We used the multiple regression analysis, with stepwise variable selection to obtain a parsimonious linear model that allows predicting the GRBAS parameters (dependent variables) as a function of acoustic measurements (independent variables), for each one of the four software programs under study. The matrix of linear regression output helped to determine the signal prediction (positive and negative) and its quality (quality <20% = trivial; between 21% and