A DATABASE AND DIGITAL SIGNAL PROCESSING FRAMEWORK FOR THE PERCEPTUAL ANALYSIS OF VOICE QUALITY

R. M. Bermúdez de Alvear[1], J. Corral[2], L. J. Tardón[2], A. M. Barbancho[2], E. Fernández Contreras[1], S. Rando Márquez[1], A. G. Martínez-Arquero[3], I. Barbancho[1]
Universidad de Málaga, Andalucía Tech,
[1]Departmento de Radiología y Medicina Física, Oftalmología y Otorrinolaringología
[2]Departmento de Ingeniería de Comunicaciones
[3] Hospital Regional de Málaga Carlos Haya, Unidad de Gestión Clínica Otorrinolaringología
bermudez@uma.es, jcorral@ic.uma.es, lorenzo@ic.uma.es, abp@ic.uma.es, ginesmart@gmail.com, ibp@ic.uma.es

Abstract

Introduction. Clinical assessment of dysphonia relies on perceptual as much as instrumental methods of analysis [1]. The perceptual auditory analysis is potentially subject to several internal and external sources of bias [2]. Furthermore acoustic analyses which have been used to objectively characterize pathological voices are likely to be affected by confusion variables such as the signal processing or the hardware and software specifications [3]. For these reasons the poor correlation between perceptual ratings and acoustic measures remains to be a controversial matter [4]. The availability of annotated databases of voice samples is therefore of main importance for clinical and research purposes. Databases to perform digital processing of the vocal signal are usually built from English speaking subjects' sustained vowels [5]. However phonemes vary from one language to another and to the best of our knowledge there are no annotated databases with Spanish sustained vowels from healthy or dysphonic voices. This work shows our first steps to fill in this gap. For the aim of aiding clinicians and researchers in the perceptual assessment of voice quality a two-fold objective was attained. On the one hand a database of healthy and disordered Spanish voices was developed; on the other an automatic analysis scheme was accomplished on the basis of signal processing algorithms and supervised learning machine techniques. **Material and methods.** A preliminary annotated database was created with 119 recordings of the sustained Spanish /a/; they were perceptually labeled by three experienced experts in vocal quality analysis. It is freely available under Links in the ATIC website (www.atic.uma.es). Voice signals were recorded using a headset condenser cardioid microphone (AKG C-544 L) positioned at 5 cm from the speaker's mouth commissure. Speakers were instructed to sustain the Spanish vowel /a/ for 4 seconds. The microphone was connected to a digital recorder Edirol R-09HR. Voice signals were digitized at 16 bits with 44100 Hz sampling rate. Afterwards the initial and last 0.5 second segments were cut and the 3 sec. mid portion was selected for acoustic analysis. Sennheiser HD219 headphones were used by judges to perceptually evaluate voice samples. To label these recordings raters used the Grade-Roughness-Breathiness (GRB) perceptual scale which is a modified version of the original Hirano's GRBAS scale, posteriorly modified by Dejonckere et al., [6]. In order to improve intra- and inter-raters' agreement two types of modifications were introduced in the rating procedure, i.e. the 0-3 points scale resolution was increased by adding subintervals to the standard 0-3 intervals, and judges were provided with a written protocol with explicit definitions about the subintervals boundaries. By this way judges could compensate for the potential instability that might occur in their internal representations due to the perceptual context influence [7]. Raters' perceptual evaluations were simultaneously performed by means of connecting the Sennheiser HD219 headphones to a multi-channel headphone preamp Behringer HA4700 Powerplay Pro-XL. The Yin algorithm [8] was selected as initial front-end to identify voiced frames and extract their fundamental frequency. For the digital processing of voice signals some conventional acoustic parameters [6] were selected. To complete the analysis the Mel-Frequency Cepstral Coefficients (MFCC) were further calculated because they are based on the auditory model and they are thus closer to the auditory system response than conventional features. **Results.** In the perceptual evaluation excellent intra-raters agreement and very good inter-raters agreement were achieved. During the supervised machine learning stage some conventional features were found to attain unexpected low performance in the classification scheme selected. Mel Frequency Cepstral Coefficients were promising for assorting samples with normal or quasi-normal voice quality. **Discussion and conclusions.** Despite it is still small and unbalanced the present annotated data base of voice samples can provide a basis for the development of other databases and automatic classification tools. Other authors [9, 10, 11] also found that modeling the auditory non-linear response during signal processing can help develop objective measures that better correspond with perceptual data. However highly disordered voices classification remains to be a challenge for this set of features since they cannot be correctly assorted by either conventional variables or the auditory model based

measures. Current results warrant further research in order to find out the usability of other types of voice samples and features for the automatic classification schemes. Different digital processing steps could be used to improve the classifiers performance. Additionally other types of classifiers could be taken into account in future studies.
Acknowledgment. This work was funded by the Spanish Ministerio de Economía y Competitividad, Project No. TIN2013-47276-C6-2-R has been done in the Campus de Excelencia Internacional Andalucía Tech, Universidad de Málaga.

References

[1] Carding PN, Wilson JA, MacKenzie K, Deary IJ. Measuring voice outcomes: state of the science review. The Journal of Laryngology and Otology 2009;123,8:823-829.

[2] Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. Folia Phoniatrica et Logopaedica 2009;61,1:49-56.

[3] Maryn et al. Meta-analysis on acoustic voice quality measures. J Acoust Soc Am 2009; 126, 5: 2619-2634.

[4] Vaz Freitas et al. Correlation Between Acoustic and Audio-Perceptual Measures. J Voice 2015;29,3:390.e1

[5] "Multi-Dimensional Voice Program (MDVP) Model 5105. Software Instruction Manual", Kay PENTAX, A Division of PENTAX Medical Company, 2 Bridgewater Lane, Lincoln Park, NJ 07035-1488 USA, November 2007.

[6] Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, Van De Heyning P, Remacle M, Woisard V. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Comm. on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol 2001;258:77–82.

[7] Kreiman et al. Voice Quality Perception. J Speech Hear Res 1993;36:21-4

[8] De Cheveigné A, Kawahara H. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Amer. 202; 111,4:1917.

[9] Shrivastav et al. Measuring breathiness. J Acoust Soc Am 2003;114,4:2217-2224.

[10] Saenz-Lechon et al. Automatic Assessment of voice quality according to the GRBAS scale. Eng Med Biol Soc Ann 2006;1:2478-2481.

[11] Fredouille et al. Back-and-forth methodology for objective voice quality assessment: from/to expert knowledge to/from automatic classification of dysphonia. EURASIP J Appl Si Pr 2009.