The Effect of Anchors and Training on the Reliability of Voice Quality Ratings for Different Types of Speech Stimuli

*Lilia Brinca, †Ana Paula Batista, *Ana Inês Tavares, *Patrícia N. Pinto, and ‡Lara Araújo, *†*Faro, and* ‡*Marco de Canaveses, Portugal*

Summary: Objective. The main objective of the present study was to investigate if the type of voice stimuli—sustained vowel, oral reading, and connected speech—results in good intrarater and interrater agreement/reliability.

Study design. A short-term panel study was performed.

Methods. Voice samples from 30 native European Portuguese speakers were used in the present study. The speech materials used were (1) the sustained vowel /a/, (2) oral reading of the European Portuguese version of "The Story of Arthur the Rat," and (3) connected speech. After an extensive training with textual and auditory anchors, the judges were asked to rate the severity of dysphonic voice stimuli using the phonation dimensions G, R, and B from the GRBAS scale. The voice samples were judged 6 months and 1 year after the training.

Results. Intrarater agreement and reliability were generally very good for all the phonation dimensions and voice stimuli. The highest interrater reliability was obtained using the oral reading stimulus, particularly for phonation dimensions grade (G) and breathiness (B). Roughness (R) was the voice quality that was the most difficult to evaluate, leading to interrater unreliability in all voice quality ratings.

Conclusions. Extensive training using textual and auditory anchors and the use of anchors during the voice evaluations appear to be good methods for auditory-perceptual evaluation of dysphonic voices. The best results of interrater reliability were obtained when the oral reading stimulus was used. Breathiness appears to be a voice quality that is easier to evaluate than roughness.

Key Words: Auditory anchors–Vocal quality rating–Oral reading stimulus–Agreement–Reliability.

INTRODUCTION

Voice quality is the sensation that the voice acoustic signal evokes in the listener. Considering that the goal of speech is communication, the central role of voice quality perception is not surprising. Auditory-perceptual evaluation is a highly valued clinical tool used in voice diagnosis, assessment, and treatment.[1,2] Moreover, perceptual measures are frequently used as a standard against which acoustic measures are validated or compared.[3–5] However, a review of the literature indicates that because the perceptual voice evaluation is a subjective process, both intrarater and interrater reliability fluctuate greatly from study to study. Kreiman et al[1] proposed a theoretical framework that attributes variability in ratings to the variable internal standards acquired by individuals and subsequently stored in memory. These internal standards are unstable because they can be affected by such factors as the individual's memory and the acoustic context in which the voice signals are evaluated.[1,6,7] The use of external standards has been suggested by some authors[1,6] to overcome the effect of the variability of internal standards. External references could replace the internal representations and should, therefore, lead to relatively more reliable evaluation. Several studies showed that using synthetic and/or natural pathologic voice anchors increases listener agreement/reliability.[6,8–13] The use of anchors may even improve agreement across listeners with different backgrounds.[14] To improve listener agreement/reliability, training and textual anchors have been added to the auditory anchors.[10–12] Two levels of training can be considered[1]: (1) orientation, in which listeners are provided with definitions of scale terms, sample or anchor stimuli, and/or a limited number of practice trials and (2) extensive training, in which listeners are provided with definitions of scale terms, anchor stimuli, and a detailed training program. Chan and Yiu[11] used an extensive training format wherein each participant has been subjected to a pretraining rating test, as a baseline measurement, followed by a training session and a posttraining rating test.

The main objective of the present study was to investigate what type of voice stimuli (sustained vowel, oral reading, and connected speech) results in the highest values of intrarater and interrater agreement/reliability. Another purpose was to investigate if this agreement/reliability is maintained when tested 1 year after the extensive training was finished.

METHODS

Participants

The present study was based on data selected from an archival database of users who attended our voice clinic. Female voices were specifically selected because, similar to many multidisciplinary voice clinics, most of our patients seeking help for voice difficulties are women. Participants included 27 native

Accepted for publication January 27, 2015.

From the *Laboratory of Audiology and Speech and Language Therapy, School of Health, University of Algarve, Faro, Portugal; †Service of Physical Medicine and Rehabilitation, Algarve Central Hospital, Faro, Portugal; and the ‡Clinic Center "Crescer a Norte, Centro de Apoio Psicopedagógico Lda.", Marco de Canaveses, Portugal.

Address correspondence and reprint requests to Lilia Brinca, Laboratory of Audiology and Speech and Language Therapy, School of Health, University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal. E-mail: lbrinca@ualg.pt

Journal of Voice, Vol. 29, No. 6, pp. 776.e7-776.e14
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2015.01.007

European Portuguese speakers diagnosed with a variety of laryngeal pathologies by one of two experienced otolaryngologists and three individuals with no laryngeal pathology (Table 1). The mean age of all participants was 44.4 years (age range, 20–72 years). Diagnosis of laryngeal pathology was based on a video image system (Xion Medical, Berlin, Germany), that allowed the amplification and recording of the images using a specific software (*Divas*; Xion Medical). To obtain the images, two procedures were used, the videolaryngoscope with a rigid optic fiber or the nasolaryngoscope with a flexible optic fiber. The latter procedure was used whenever the participant presented impediments caused by the rigid laryngoscope. A continuous light source was used for laryngoscopic examination.

Informed consent was obtained from all the subjects included in the study.

Speech material

The speech materials used for the auditory-perceptual measurements were (1) the sustained vowel /a/, (2) oral reading of the European Portuguese version of "The Story of Arthur the Rat,"[15] and (3) connected speech, a verbal response to the question "What have you been doing today?"

Voice samples were collected in a soundproof booth with the participants seated on a comfortable chair. An omnidirectional microphone was used (DPA 4006-TL, P48; Harman International Industries Ltd., Hertfordshire, United Kingdom), pointed to the speaker's mouth at a distance of approximately 30 cm. Outside the soundproof booth, the microphone was connected to a computer through a mixer (Soundcraft Compact 4, Harman International Industries Ltd.) and a USB audio interface (Edirol UA-1EX; DPA Microphones A/S, Denmark). All vocal recordings were digitized at a sampling rate of 44.1 kHz and a resolution of 16 bits and saved in WAV format, using *Audacity 2.0.2*, a General Public License program published in 2000 by SourceForge.net.

TABLE 1.
Age (Years) and Clinical Diagnosis of Participants

Subject	Age	Diagnosis	Subject	Age	Diagnosis
1	46	LG	16	41	LG
2	72	LG + E	17	29	VFE + GER
3	52	LG	18	58	LG
4	43	VFN	19	34	LG
5	48	VFC	20	48	LG
6	34	LG	21	58	VFP
7	51	LG	22	61	E
8	43	LG + GER	23	46	VFP
9	61	LG	24	23	N
10	42	N	25	60	LG
11	55	LG	26	20	VFN
12	29	VFN + E	27	22	VFN
13	41	LG	28	31	VFN
14	50	N	29	42	LG
15	41	LG	30	52	E

Abbreviations: LG, longitudinal gap; E, generalized edema; VFE, vocal folds edema; GER, gastroesophageal reflux; VFN, vocal folds nodules; VFC, vocal fold cyst; VFP, ventricular fold phonation; N, normal structure and function of the vocal folds.

Experimental design

Judges. Three female speech and language therapists, all with over 3 years of extensive experience evaluating and treating vocal pathologies, participated as judges in the present study. All were trained during their speech and language therapy course to evaluate disordered voices perceptually and to rate their order of severity using the GRBAS system. All judges passed hearing screening tests at 20 dB for the octave frequencies of 250–8000 Hz. A speech and language therapist, with more than 15 years of clinical experience in judging dysphonia severity, acted as "coach." She selected the anchors and the speech samples to train the judges, on the basis of her individual's own internal standards.

Rating scale. Judges were asked to rate the severity of dysphonic voice stimuli using the GRBAS scale. The GRBAS scale of the Japan Society of Logopedics and Phoniatrics was introduced and gained international recognition with the publication of Hirano's "Clinical Examination of Voice."[16] The scale comprises five parameters, and each parameter represents a dimension of phonation: grade, roughness, breathiness, asthenicity, and strain. The GRBAS scale uses a four-point Likert scale of 0 (normal) to 3 (extreme) for all five parameters (ie, it is an ordinal scale). Three voice quality parameters were examined in the present study—G (grade), which relates to the overall voice quality, integrating all deviant components[17]; R (roughness), perceptually defined as "an impression of irregular glottis pulses"[17]; and B (breathiness), perceptually defined as "an auditive impression of turbulent air leakage through an insufficient glottic closure, including short aphonic moments."[17] The rationale for including only G, R, and B from GRBAS scale is based on a basic protocol for functional assessment of voice pathology issued by the Committee on Phoniatrics of the European Laryngological Society in 2001.[17] According to the committee, the parameters R and B have shown sufficient reliability during current clinical use; in contrast to this, the behavioral parameters A (asthenicity) and S (strain) are currently found to be less reliable and the committee suggested that they could be omitted from the basic protocol. A simplified scale, GRB, was proposed to be used during present clinical perceptual evaluation.

Procedures. Speech samples were presented free field via two loudspeakers (DYN-BM15A Dynaudio Acoustics A/S, Skanderborg, Denmark) in a listening environment free of ambient noise.

(1)Each judge first took a "pretraining session" as a baseline rating test. The "coach" presented the definitions of voice quality parameters G, R, and B to be used in the present study. Moreover, a total of 10 female voice samples were presented during the pretraining session using two examples of nondysphonic voices and two examples of each of the two dysphonic voice qualities—roughness