Eur Arch Otorhinolaryngol (2014) 271:1609–1619
DOI 10.1007/s00405-013-2730-7

LARYNGOLOGY

The value of the Acoustic Voice Quality Index as a measure of dysphonia severity in subjects speaking different languages

Youri Maryn · Marc De Bodt · Ben Barsties · Nelson Roy

Received: 29 April 2013 / Accepted: 23 September 2013 / Published online: 26 October 2013

Abstract The Acoustic Voice Quality Index (AVQI) is a relatively new clinical method to quantify dysphonia severity. Since it partially relies on continuous speech, its performance may vary with voice-related phonetic differences and thus across languages. The present investigation therefore assessed the AVQI's performance in English, Dutch, German, and French. Fifty subjects were recorded reading sentences in the four languages, as well as producing a sustained vowel. These recordings were later edited to calculate the AVQI. The samples were also perceptually rated on overall dysphonia severity by three experienced voice clinicians. The AVQI's cross-linguistic concurrent validity and diagnostic precision were assessed. The results support earlier data, and confirm good cross-linguistic validity and diagnostic accuracy. Although no statistical differences were observed between languages, the AVQI performed better in English and German and less well in French. These results validate the AVQI as a potentially robust and objective dysphonia severity measure across languages.

Keywords Dysphonia · Clinical assessment · Acoustic Voice Quality Index · Different languages

Introduction

Acoustic measures of dysphonia severity are commonly employed in voice clinics because they have the potential to automated analysis algorithms, to yield non-invasive data, and to quantify dysphonia changes and outcomes following intervention. Such metrics of dysphonia severity are traditionally derived from sustained vowel productions and not from continuous (or connected, or running) speech. The relatively steady state (i.e., time and frequency invariance) of phonation associated with sustained vowels make them attractive for acoustic analysis as compared to the rapid and frequent glottal and supraglottal changes observed during continuous speech. Sustained vowels normally do not contain voiceless phonemes, fast voice onsets and terminations, and prosodic fluctuations in F_0 and amplitude. Furthermore, they are not influenced by speech rate and stress, vocal

Y. Maryn (✉)
Department of Speech-Language Pathology and Audiology, Sint-Jan General Hospital, Ruddershove 10, 8000 Brugge, Belgium
e-mail: youri.maryn@azsintjan.be

Y. Maryn
Department of Otorhinolaryngology and Head & Neck Surgery, Sint-Jan General Hospital, Brugge, Belgium

Y. Maryn
Department of Speech Therapy and Audiology, Faculty of Health Care, University College Ghent, Ghent, Belgium

M. De Bodt
Department of Communication Disorders, University Hospital, Antwerp, Belgium

M. De Bodt
Department of Otorhinolaryngology and Head & Neck Surgery, University Hospital, Antwerp, Belgium

B. Barsties
Faculty of Health Care, University of Applied Sciences Utrecht, Utrecht, The Netherlands

N. Roy
Division of Otolaryngology and Head & Neck Surgery, Department of Communication Sciences and Disorders, University of Utah, Salt Lake City, USA

pauses, and especially phonetic context. Finally, sustained vowels are relatively insulated from influences related to language [1–5]. Hence, the sustained vowel is considered the most common "language-independent" voice material used in clinical voice assessment. However, one major limitation of the sustained vowel is that it is an artificial type of phonation and that it lacks 'ecological validity' (i.e., it is insufficiently representative of daily speech and voice use patterns) [4, 5]. Connected speech samples, on the other hand, are highly representative of daily voicing and can be considered more ecologically valid [4, 5]. Furthermore, symptoms of disordered voice quality usually emerge in conversational voice production instead of sustained vowels (except for singing voice) and are typically revealed when patients use continuous speech [6]. However, continuous speech samples can vary substantially in their phonetic and phonatory composition depending on language, dialect, and region. Vocal physiology and phonatory output can, therefore, be hypothesized to be different across languages.

In fact, what appears to be vocally pathological in one language, can be necessary for the purpose of phonological contrast in another language. As noted by Ladefoged [7], one person's voice disorder might be another person's phoneme. For example in Dutch, a breathy voice is considered to be pathological, whereas in Portuguese the same degree of breathiness can be of phonological salience, and provides phonemic contrast for vowels after a stressed syllable. Similarly, pathological degrees of roughness in Dutch can be phonologically distinctive in, for example, Danish syllable stress. In addition to the phonological value of phonatory variations in specific languages, voice can also vary normally across languages, depending on the set of target phonemes, their surrounding phonetic contexts, their position in the prosodic contour of utterances and their duration. Many basic science studies have yielded evidence for this glottal–supraglottal (i.e., phonatory–articulatory) interdependency. For instance, Schulman [8] simultaneously examined lip and jaw movement changes associated with variations in speech intensity. The results revealed meaningful relations between intensity and articulatory movement and timing. Furthermore, Higgins et al. [9] investigated laryngeal behavior during different vowels, and found significant vowel-related differences in subglottal air pressure, electroglottographic cycle width, fundamental frequency and voice onset time. Dromey and Ramig [10] investigated the effects of changing intensity and speech rate during sentence production on respiratory, phonatory and articulatory function. They found that both intensity and speech rate are associated with changes in lip movement, lung volume and fundamental frequency (F_0). Cookman and Verdolini [11] explored the mandibular-laryngeal interaction and the results of their study revealed an increase in laryngeal adduction during increased jaw lowering and increased jaw biting force. Additionally, McClean and Tasko [12] investigated the relation between the velocity of lip, jaw and tongue movements and the two larynx-related measures F_0 and intensity. Their results showed significant correlations between orofacial (and especially mandibular) speed and laryngeal physiology, suggesting neural orofacial-laryngeal coupling during vocalization. Based on these findings, language- or even dialect-dependent phonetic differences can be expected to influence vocal behavior, and possibly the quality of voice. Therefore, voice assessment methods which employ continuous speech, may be vulnerable to certain inter-linguistic variations; and thus the introduction of continuous speech might induce inter-linguistic differences that should be identified and accounted for in voice assessment.

One such method of voice assessment which includes continuous speech in its analysis is the Acoustic Voice Quality Index (AVQI). Maryn et al. [4] developed the AVQI as an objective method to quantify dysphonia severity. This index was the first to combine both continuous speech and sustained vowels, by concatenating recordings of 3 s of [*a*:] with the voiced segments of two sentences read aloud. By combining the two voice materials, the goal was to provide a more ecologically valid index of dysphonia severity. To calculate an AVQI score, a weighted combination of six acoustic time domain [i.e., shimmer local, shimmer local dB and harmonics-to-noise ratio (HNR)], frequency domain (i.e., general slope of the spectrum and tilt of the regression line through the spectrum) and quefrency domain [i.e., smoothed cepstral peak prominences (CPPs)] parameters are modeled in a linear regression formula. Testing of concurrent validity and diagnostic precision revealed equally favorable results in both an initial group of 251 Dutch-speaking subjects [4] and a new group of 39 Dutch-speaking subjects [5]. Both groups were considered to be representative of a clinical population of voice-disordered patients, reflecting different age and gender groups, and different types and degrees of voice quality disruption and vocally induced disability, including nonorganic as well as organic laryngeal pathologies. Furthermore, the studies of Reynolds et al. [13] and Barsties and Maryn [14] showed similar findings for English-speaking children and German-speaking children and adults, respectively. Lastly, examination of sensitivity to change over the course of voice treatment in 33 Dutch-speaking dysphonic patients showed that the AVQI is a promising treatment outcomes measure [5].

To standardize clinical voice assessment and to optimize communication of clinical data in a multilinguistic society such as in Belgium [i.e., with French, German and Dutch as the three leading languages, and with English commonly present in both media (e.g., music, television, internet, etc.) and education], it is important to investigate the influence

of language and phonetic context/structure on the outcome of running speech-based analysis methods. Therefore, this study was designed to examine the impact of language on the AVQI, which was originally constructed on Dutch-spoken continuous speech samples.

Methods

To assess the effects of language on AVQI performance, sustained vowels as well as sentences spoken in Dutch, German, French and English were recorded and concatenated according to the AVQI's procedure. Later, sustained vowel–continuous speech–combinations were rated for overall dysphonia severity (or Grade, "G" from the GRBAS rating scale [15]) by experienced voice clinicians. Two analyses were then completed. In the first analysis, the cross-linguistic criterion-related concurrent validity of the AVQI was examined in terms of proportional relationships between G-ratings and AVQI scores. In the second analysis, the cross-linguistic diagnostic precision of the AVQI was investigated in terms of its ability to correctly differentiate between normophonia and dysphonia in selected languages.

Subjects

Voice samples from 50 Flemish Dutch-speaking subjects, consecutively recruited from the ENT department of the Sint-Jan General Hospital in Bruges, Belgium, were employed in this investigation. Subjects were selected for inclusion when the first author judged them to be sufficiently fluent in the four languages of interest. The first author is a native Dutch speaker who is also fluent in French and German (i.e., the second and third "official" languages of Flanders, Belgium). He is also highly proficient in English with excellent receptive and expressive language skills. While English is not an official language in Belgium, it is ubiquitous in Belgian culture, society, and media: television programs, music, advertising, etc. Furthermore, like French, it is taught throughout middle and high school years. Thus, the first author possesses the requisite expertise in these languages to judge the proficiency/fluency of the speakers. Although they were not native speakers of French, German and English, all subjects also affirmed to be familiar enough with these languages to comfortably and accurately read all the sentences. Thus, the speech recordings were considered to be dialectal variants of the respective languages.

Table 1 lists the laryngological diagnoses included in the sample, using an Olympus ENF-V flexible transnasal chip-on-tip laryngostroboscope (Olympus Corp., Tokyo, Japan). This group of subjects was considered to be adequately representative of a voice clinic population, reflecting different ages, genders, different types and degrees of voice quality, and voice-related disability. They included non-organic as well as organic laryngeal pathologies. As part of routine clinical practice, their voices were recorded (i.e., at the beginning of the standard voice assessment). The group consisted of 29 females and 21 males, ranging in age from 10 to 77 years (mean = 44.9 years, SD = 19.2 years). The methods of data acquisition and analysis are identical to those described previously in Maryn et al. [4, 5].

Table 1 Primary laryngostroboscopic diagnoses of subjects included in the study

Laryngeal status	Absolute number	% Of total
Nodules	13	26
Normal	12	24
Unilateral vocal fold paralysis	9	18
Post head and neck cancer treatment[a]	5	10
Muscle tension dysphonia	4	8
Laryngitis	3	6
Polypoid mucosa	2	4
Presbylarynx	1	2
Leukoplakia	1	2
Total	50	100

[a] The treatment consisted of only radiation in two subjects, only surgery in one subject, and radiation plus surgery in two subjects

Voice recordings

All voice samples were recorded using an AKG C420 head-mounted condenser microphone (AKG Acoustics, München, Germany), digitized at a sampling rate of 44.1 kHz and a resolution of 16 bits using the "Computerized Speech Lab model 4500´´ (Kay Elemetrics Corp., Lincoln Park, NJ, USA), and saved in WAV-format. Further editing, selecting, segmenting, etc., of the recordings were completed using the program "Praat" (Institute of Phonetic Sciences, University of Amsterdam, The Netherlands). First, subjects were asked to sustain the vowel [*a*:] at comfortable pitch and loudness. A time window was applied during recording to include only a mid-vowel portion of 3 s. Second, they were asked to read aloud six text samples representing the four languages: two Dutch sentences, two English sentences, one German sentence, and one French sentence. The six text samples are attached in "Appendix" together their phonetic transcriptions.

The frequency of occurrence of the phonemes in these six samples is listed in Table 2. With $N = 12$, the most frequent phoneme was [ə] in Du2, followed by [*n*] with $N = 9$ in Du2 and Ge5. The frequencies in this contingency table were examined to determine the possible influence of differences in phonetic content across the six sequences.

Table 2 Contingency table (6 rows × 46 columns) with the frequency of the 46 phones across the six sets of connected speech

Phone	*p*	*a*	*ɛ*	*n*	*m*	*r*	*l*	*u*	*s*	*t*	*aˑ*	*ɔ*	*ə*	*i*	*j*	.	*z*	*w*	*χ*	*d*	*ɛʲ*	*oˑ*	*ɪ*
Du1	4	5	1	5	1	2	1	1	3	5	1	3	4	1	1	1	1	1	1	1	1	0	0
Du2	0	2	2	9	0	7	0	0	4	6	2	1	12	0	0	0	1	5	0	6	1	2	2
En3	0	1	2	5	0	0	0	1	2	3	0	3	2	1	1	0	0	5	0	1	0	0	3
En4	2	0	3	5	2	6	2	0	3	3	0	1	3	0	1	1	1	2	0	1	0	0	2
Ge5	0	0	1	9	0	7	1	1	1	7	0	3	7	1	0	1	2	2	1	2	0	0	3
Fr6	2	3	2	0	0	2	7	0	4	2	0	1	2	2	0	1	1	0	0	1	1	0	0

Phone	*v*	*o*	*E*	*k*	*b*	*y*	*ð*	*ɔˑ*	*θ*	*ɑˑ*	*ɑˑ*	*g*	*ŋ*	*ʃ*	*f*	*aʲ*	*eʲ*	*æ*	*ə̃*	*ɔˑ*	*ɔʏ*	*iːˑ*	*v*
Du1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Du2	2	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
En3	0	0	1	0	0	0	3	1	1	2	1	2	2	1	1	0	0	0	0	0	0	0	0
En4	0	1	1	2	2	0	1	1	1	0	0	0	0	0	1	3	3	1	0	0	0	0	0
Ge5	1	1	3	1	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	2
Fr6	0	0	2	2	1	3	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	1	0

Dysphonia severity ratings

For the auditory-perceptual judgments of dysphonia severity, each of the six text segments, a pause of 2 s and the central vowel segment were concatenated into one single sound wave. An example of the resulting waveform can be found in Fig. 1.

Three experienced speech-language pathologists, all Dutch, were asked to rate all concatenated voice samples. Except for the first author (who collected all recordings), all raters were blinded regarding the identity, diagnosis and disposition of the 50 subjects (i.e., normal, pre-treatment, post-treatment, etc.,). For each language, all concatenated voice samples were presented in random order and judged on overall severity of dysphonia (i.e., Grade or G) with a four-point equal-appearing interval scale (i.e., "normal" voice quality, 0; slight dysphonia, 1; moderate dysphonia, 2; and severe dysphonia, 3), as suggested by the Japan Society of Logopedics and Phoniatrics [15]. The listening environment and procedures were comparable to those described in the previous studies [4, 5]. Furthermore, as recommended by Chan and Yiu [16], an attempt was made to establish an external standard, to putatively increase the reliability of listener ratings. Thus, 12 samples were selected from the database from previous studies, i.e., three samples per level of G. These samples were selected based upon prior unanimous agreement across raters regarding the degree of dysphonia, thus these samples were considered to be highly representative of a specific level of G. All listener ratings were conducted within a single session.

Five voice samples per language (i.e., 10 % of all samples) were randomly selected and repeated a second time at the end of the perceptual experiment to assess intra-rater reliability.

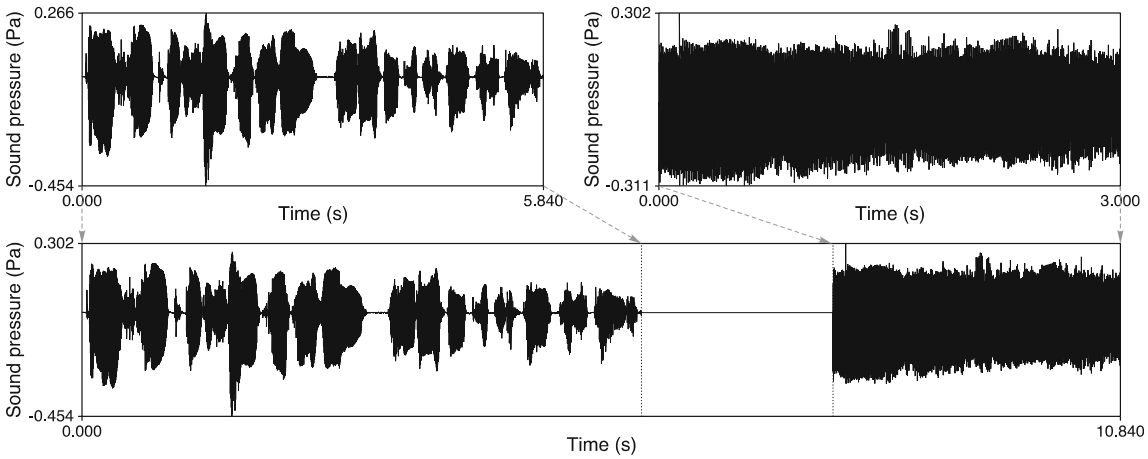

Fig. 1 Oscillograms of the sound recordings that were used in the auditory-perceptual evaluations of this study: (*upper left*) connected speech with the two sentences (*upper right*) 3 s of the sustained vowel [*a*.], and (*lower*) concatenation of these two sound files separated by 2 s of silence

Acoustic measures

Since certain acoustic measures employed in this study are only valid for voiced segments of the continuous speech samples, an algorithm for detection, segmentation, and concatenation of these voiced segments were used. This algorithm was originally based on Parsa and Jamieson [2] and customized in Praat by Maryn et al. [4]. An example of the resulting waveform is depicted in Fig. 2.

Objective measurement of overall voice quality consisted of determining the six acoustic parameters for calculating the AVQI: smoothed CPPs with the computer program "SpeechTool" (James Hillenbrand, Western Michigan University, Kalamazoo, MI, USA) and HNR, shimmer local (SL), shimmer local dB (SLdB), general Slope of the spectrum (i.e., slope) and tilt of the regression line through the spectrum (i.e., tilt) with Praat. The smoothed CPP is the distance between the first rahmonic's peak and the point with equal quefrency on the regression line through the smoothed cepstrum, as illustrated in Fig. 3. The HNR is the base-10-logarithm of the ratio between the periodic energy and the noise energy, multiplied by 10. The shimmer local is the absolute mean difference between the amplitudes of successive periods, divided by the average amplitude. The SLdB is the base-10-logarithm of the difference between the amplitudes of successive periods, multiplied by 20. The general Slope is the difference between the energy in 0–1,000 Hz and the energy in 1,000–10,000 Hz of the long-term average spectrum. The Tilt is the difference between the energy in 0–1,000 Hz and the

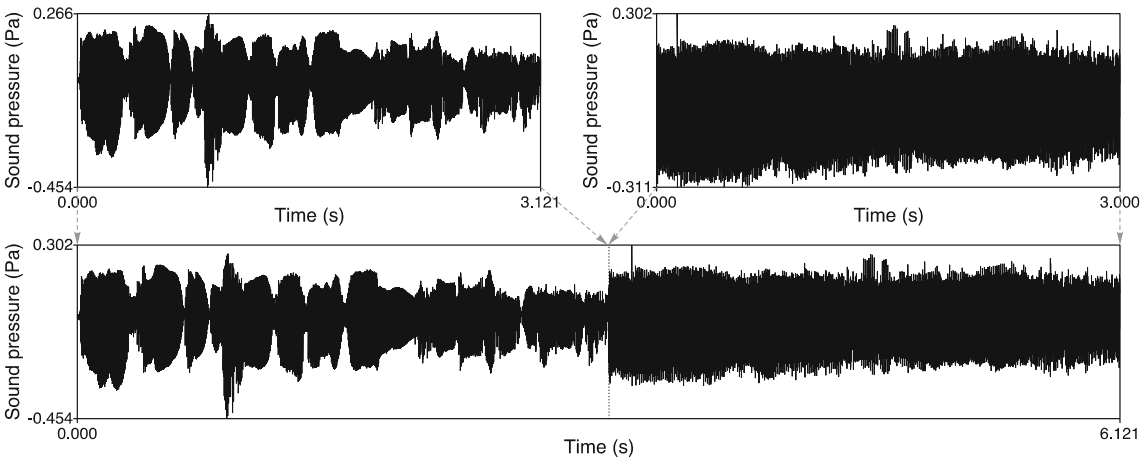

Fig. 2 Oscillograms of the voice samples used for the acoustic method in this study: (*upper left*) the concatenated voiced segments of the two read sentences (*upper right*) 3 s of the sustained vowel [*a*.], and (*lower*) concatenation of these two sound files

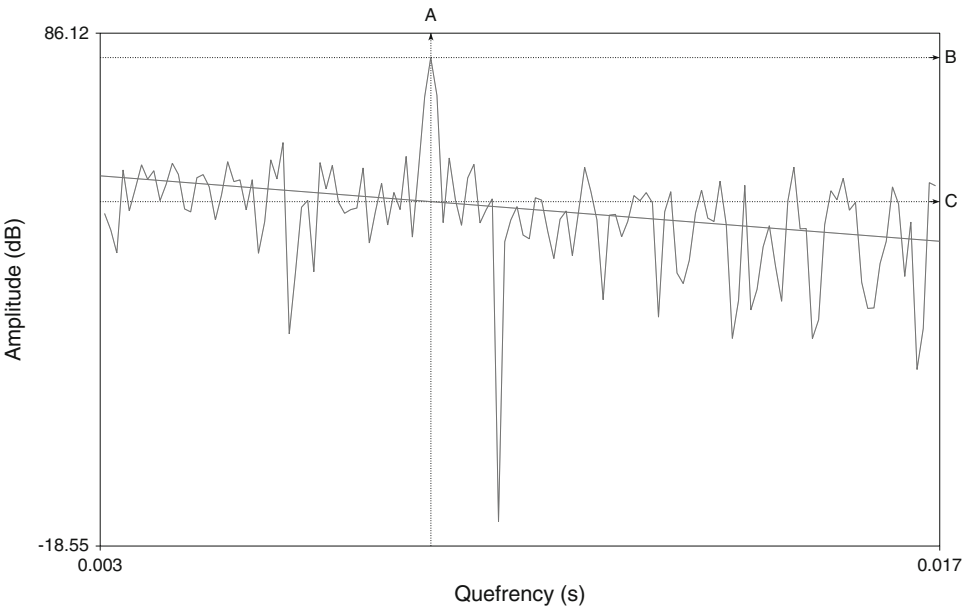

Fig. 3 An example of a smoothed cepstrum to illustrate how the smoothed cepstral peak prominence (CPPs) is determined. At the quefrency of the top of the first rahmonic or cepstral peak (i.e., point A, quefrency = 0.008 s), a difference is calculated between the amplitude of the smoothed cepstrum (i.e., point B, amplitude = 81.12 dB) and the amplitude of the tilt line through the smoothed cepstrum (i.e., point C, amplitude = 51.73 dB). The subtraction of B−C results in a CPPs = 29.39 dB

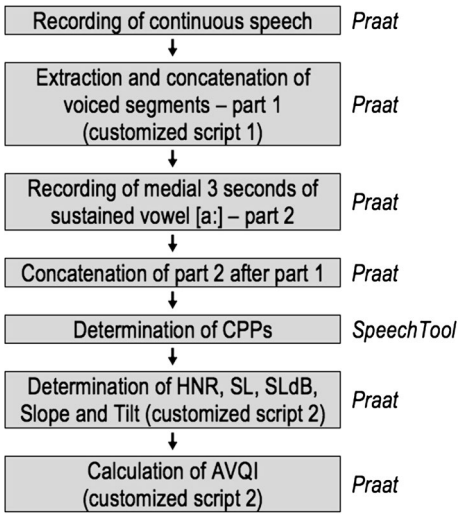

Fig. 4 Flowchart illustrating the procedure to determine the AVQI (on the *right*: the computer programs in which the steps are carried out)

energy in 1,000–10,000 Hz of the trendline through the long-term average spectrum.

The method used to compute the six acoustic measures was identical to the method employed in the two prior studies, and the customized Praat script for automated voice analysis and AVQI calculation is provided in Maryn et al. [4]. A flowchart of the consecutive steps to determine the AVQI is presented in Fig. 4. Subsequently, the AVQI was calculated according to the regression formula:

$$\begin{aligned}\text{AVQI} = 2.571 \times [3.295 &- (0.111 \times \text{CPPs}) \\ &- (0.073 \times \text{HNR}) - (0.213 \times \text{SL}) \\ &+ (2.789 \times \text{SLdB}) - (0.032 \times \text{Slope}) \\ &+ (0.077 \times \text{Tilt})].\end{aligned}$$

Statistical analyses

All statistical analyses were completed using SPSS for Windows version 12.0 (SPSS Inc., Chicago, IL, USA), except when stated otherwise. First, the inter-rater reliability of the three raters was assessed for each language using the single-measures intraclass correlation coefficient (ICC). This ICC is a standard reliability index that measures the degree of consistency among ratings from individual listeners. Like other reliability coefficients, the ICC ranges from 0.00 (i.e., total absence of reliability) to 1.00 (i.e., perfect reliability). Although there are no standard values for the interpretation of the ICC, a general guideline suggests that values above 0.75 indicate good reliability, and values below 0.75 correspond with poor to moderate reliability [17]. Second, the intra-rater reliability of the three evaluators for the 30 repeated voice samples was also estimated with the single-measure ICC. Third, differences related to the phonetic composition of the six connected speech samples was examined with a general Chi square (χ^2) homogeneity test for a 6 × 46 table as computed on all of the frequency data in the contingency table (i.e., Table 2) [18]. Furthermore, 15 χ^2 homogeneity statistics for 2 × 46 contingency tables were calculated across all pairs of phonetic sequences (i.e., Du1–Du2, Du1–En3, Ge5–Fr6, etc.,).

To quantify the cross-linguistic criterion-related concurrent validity of the AVQI (i.e., how well the AVQI correlates with the listener-based estimates of dysphonia severity across six different speech segments for four different languages), the Spearman's rank-order correlation coefficient (r_s) and the coefficient of determination (r_s^2) between AVQI and mean G (G_{mean}, as averaged over the three raters) were computed for the six running speech segments/four languages. In addition to these correlations, a partial Spearman's rank-order correlation coefficient ($r_{s\text{-partial}}$) was calculated with G_{mean} and AVQI as criterion and predictor variable, respectively, and with language as a nuisance variable. The $r_{s\text{-partial}}$ more purely measures the degree of association between G_{mean} and AVQI across the six different utterances, after removing any linear interdependence between G_{mean} and language or between AVQI and language [18]. To calculate the $r_{s\text{-partial}}$ between G_{mean} and AVQI after eliminating the influence of language from the analysis, the following formula was applied:

$$r_{s-\text{partial}(G_{\text{mean}},\text{AVQI})} = \frac{r_{s(G_{\text{mean}},\text{AVQI})} - r_{s(G_{\text{mean}},\text{language})} r_{s(\text{AVQI},\text{language})}}{\sqrt{\left[1 - r^2_{s(G_{\text{mean}},\text{language})}\right]\left[1 - r^2_{\text{AVQI,language}}\right]}}.$$

Furthermore, to assess the equality between these six r_s values (i.e., is the AVQI equally valid across the different groups of recordings?), the significance of the difference between two G_{mean}-AVQI-based r_s values was determined with the VassarStats software (website for statistical computation, Richard Lowry, Vassar College, Poughkeepsie, NY, USA). With the Fisher's z_r transformation, a z value is computed to estimate the significance of the difference (i.e., the two-tailed p value) between two correlation coefficients [18]: two r_s values are considered statistically significantly different when $p \leq 0.05$, and vice versa.

To evaluate the cross-linguistic diagnostic accuracy of the AVQI, several estimates were calculated. Diagnostic precision of a clinical metric is usually evaluated by its sensitivity and specificity. However, depending on the AVQI's cut-off score/threshold chosen to define a positive result, its sensitivity and specificity can vary. This trade-off between sensitivity and specificity can be graphically by generating the receiver operating characteristic (ROC) curve. The AVQI's ROC curve is created by plotting a point that, per

AVQI cut-off score, represents the true positive rate (i.e., sensitivity) on the ordinate and the false positive rate (i.e., 1−specificity) on the abscissa. As in the previous studies, a voice was considered normophonic only when all three judges rated it as normal (i.e., $G_{mean} = 0.0$). On the other hand, a voice was considered dysphonic as soon as one judge evaluated it as slightly dysphonic or $G1$ (i.e., $0.3 \leq G_{mean} \leq 3$). The ability of the AVQI to differentiate between normal and dysphonic voices was represented by the "area under ROC curve" (i.e., A_{ROC}). An $A_{ROC} = 1.0$ indicates perfect distinguishing between normal and dysphonic voices. An $A_{ROC} = 0.5$ corresponds with chance-level diagnostic accuracy [17]. To provide additional evidence regarding the value of a diagnostic measure and to help reduce problems with sensitivity/specificity related to the base-rate differences in the samples (i.e., the uneven percentages of 20 % normophonia, and 80 % dysphonia in the 300 voice samples), likelihood ratios should also be calculated [19]. The "likelihood ratio for a positive result" (i.e., LR^+) yields information regarding how the odds of the disease increase when the test is positive. It is calculated by $LR^+ = \frac{\text{sensitivity}}{1-\text{specificity}}$ and gives information regarding the likelihood that an individual is dysphonic when testing positive. The "likelihood ratio for a negative result" (i.e., LR^-) is an estimate that helps to determine if an individual does not have a particular disorder when they test negative on the diagnostic test. It is calculated by $LR^- = \frac{1-\text{sensitivity}}{\text{specificity}}$ and gives information regarding the likelihood that an individual has a normal vocal quality when testing negative. As a general guideline, the diagnostic value of a measure is considered to be high when $LR^+ \geq 10$ and $LR^- \leq 0.1$. Because the LR statistics consider sensitivity and specificity simultaneously, they are less vulnerable to sample size characteristics and base-rate differences in the sample between vocally normal and voice-disordered subjects [19]. Based on the ROC curves in the present study, the diagnostic statistics LR^+ and LR^- were calculated using the following AVQI cut-off points: 3.19 for Du1, 3.66 for Du2, 3.29 for En3, 3.25 for En4, 3.07 for Ge5, and 3.05 for Fr6.

Results

Reliability of listener auditory-perceptual ratings

The average single-measures ICC of 0.791 (with the average 95 % confidence interval from 0.691 to 0.867) exceeds the ICC threshold of 0.75, and confirms acceptable inter-rater reliability between the three raters. The mean single-measures ICC of 0.847 (with the mean 95 % confidence interval from 0.712 to 0.923) also exceeds the ICC criterion of 0.750, and indicates acceptable intra-rater reliability. Thus, for the purpose of this study both inter- and intra-rater reliability were considered acceptable.

Homogeneity across phonetic sequences

A significant omnibus $\chi^2 = 314.88$, $p = 0.000$ revealed significant differences in phonetic content across the six speech sequences. To identify where these differences resided, a detailed post hoc χ^2 analysis comparing all 15 pairs of samples was computed. The results of these analyses are displayed in Table 3. A significant difference between phonetic sequences was found for only 4 of the 15 pairs: Du2–En3, Du2–Fr6, En3–Fr6, and Ge5–Fr6. For all the other pairs, the composition of phonetic content (as measures by the proportion of phonemes) was not significantly different.

AVQI's cross-linguistic criterion-related concurrent validity

Table 4 lists the descriptive data for AVQI in the six batches of voice samples. Although the ranges between minimal and maximal AVQI scores vary somewhat between batches, it is obvious from Table 4 that the means, interquartile ranges and standard deviations are quite similar.

Table 3 Matrix showing Chi square (χ^2) values and their corresponding p values as statistics of homogeneity/difference in occurrence of phones across the six sets of phonetic sequence

Voice samples	Du1	Du2	En3	En4	Ge5
Du2	$\chi^2 = 34.80$				
	$p = 0.144$				
En3	$\chi^2 = 38.02$	$\chi^2 = 47.34$			
	$p = 0.180$	$p = 0.030$			
En4	$\chi^2 = 34.51$	$\chi^2 = 42.59$	$\chi^2 = 40.49$		
	$p = 0.349$	$p = 0.100$	$p = 0.206$		
Ge5	$\chi^2 = 33.19$	$\chi^2 = 30.39$	$\chi^2 = 41.67$	$\chi^2 = 32.93$	
	$p = 0.229$	$p = 0.345$	$p = 0.118$	$p = 0.421$	
Fr6	$\chi^2 = 34.42$	$\chi^2 = 54.95$	$\chi^2 = 54.53$	$\chi^2 = 42.61$	$\chi^2 = 53.64$
	$p = 0.264$	$p = 0.003$	$p = 0.019$	$p = 0.176$	$p = 0.010$

Table 4 Descriptive data of the AVQI values in the six sets of voice samples

Statistics	Du1	Du2	En3	En4	Ge5	Fr6
Minimum	1.16	1.79	1.88	1.71	1.90	1.68
Maximum	10.75	11.41	9.69	10.07	10.63	10.44
Range	9.59	9.62	7.81	8.36	8.73	8.76
First quartile	3.30	3.55	3.27	3.42	3.48	3.64
Third quartile	5.31	5.85	5.61	5.47	5.82	5.55
Interquartile range	2.01	2.30	2.34	2.05	2.34	1.91
Mean	4.63	4.97	4.64	4.73	4.81	4.93
Standard deviation	2.18	2.06	2.03	2.03	2.10	2.00

The statistics of AVQI's criterion-related concurrent validity across continuous speech fragments/languages are summarized in Table 5. With $r_s = 0.781$, the poorest correlation between G_{mean} and AVQI was found for the French Fr6 samples, indicating that 61.0 % (i.e., $r_s^2 = 0.610$) of the variance in G_{mean} was accounted for by AVQI. On the other hand, the English En3 samples yielded a $r_s = 0.868$—the strongest correlation—signifying that 75.3 % of G_{mean}'s variation is explained by AVQI (i.e., $r_s^2 = 0.753$). The correlation coefficients of the other four sets of samples were positioned between these two extremes. Across all six sets of phonetic sequences, an average $r_{s\ mean} = 0.829$ was found (i.e., $r_{s\ mean}^2 = 0.687$). A very similar $r_{s-partial(G_{mean},AVQI)} = 0.826$ was yielded when all samples were entered in a single batch. All these correlation coefficients designate good to very good cross-language criterion-related concurrent validity of AVQI.

Furthermore, to estimate whether the AVQI is significantly more valid in one or more specific languages, statistical tests comparing the differences in r_s between all sample pairs were computed, as represented in the matrix in Table 6. None of the paired comparisons of the correlations were statistically significant, not even between the highest r_s and the lowest r_s of the En3 and the Fr6 samples (i.e., $p = 0.180$), respectively. This implies that all r_s values can be considered as equally strong.

Table 5 Correlation coefficients (r_s) and coefficients of determination (r_s^2) between the G_{mean} scores and the AVQI data across the six different voice recordings

Voice samples	r_s	r_s^2
Du1 ($N = 50$)	0.808	0.653
Du2 ($N = 50$)	0.809	0.654
En3 ($N = 50$)	0.868	0.753
En4 ($N = 50$)	0.849	0.721
Ge5 ($N = 50$)	0.858	0.736
Fr6 ($N = 50$)	0.781	0.610
Partial ($N = 300$)	0.826	0.682
Average	0.829	0.688

AVQI's cross-linguistic diagnostic accuracy

To assess the AVQI's potential to accurately differentiate between normophonia and dysphonia, its ROCs were determined per set of running speech samples with G_{mean} as the gold standard variable (i.e., $G_{mean} = 0$ indicating normal voice quality, and $G_{mean} > 0$ indicating disordered voice quality). The A_{ROC} and other diagnosis outcomes data are summarized in Table 7.

First, with $A_{ROC\ mean} = 0.921$, the A_{ROC} statistics revealed adequate discrimination across the six sets of between vocally normal and pathological recordings (with statistical significance at $p = 0.000$, non-parametric distribution). The highest diagnostic accuracy was yielded for the Ge5 samples (i.e., $A_{ROC} = 0.958$). The least, yet still satisfactory, diagnostic adequacy was found for the Fr6 samples (i.e., $A_{ROC} = 0.869$).

Second, every ROC curve was visually inspected to specify the optimal AVQI cut-off score. Table 7 lists the separate AVQI thresholds, together with their diagnostic strength in terms of sensitivity and specificity. On average, this inspection revealed an AVQI = 3.25—producing

Table 6 Matrix exhibiting the inter-correlation differences between the six sets of voice samples

Voice samples	Du1	Du2	En3	En4	Ge5
Du2	0.992				
En3	0.322	0.332			
En4	0.522	0.535	0.726		
Ge5	0.424	0.435	0.849	0.873	
Fr6	0.719	0.711	0.180	0.322	0.250

The numbers contained within each cell represent p values (two-tailed)

Table 7 Statistics illustrating the AVQI's ability to differentiate normophonia/dysphonia across the six batches of voice recordings

Voice samples	A_{ROC}	AVQI threshold	Sensitivity	Specificity	LR^+	LR^-
Du1 ($N = 50$)	0.893	3.19	0.92	0.73	3.41	0.11
Du2 ($N = 50$)	0.894	3.66	0.85	0.80	4.25	0.19
En3 ($N = 50$)	0.953	3.29	0.90	0.90	9.00	0.11
En4 ($N = 50$)	0.956	3.25	0.95	0.82	5.28	0.06
Ge5 ($N = 50$)	0.958	3.05	0.98	0.75	3.92	0.03
Fr6 ($N = 50$)	0.869	3.07	0.97	0.70	3.23	0.04
Average	0.921	3.25	0.93	0.78	4.85	0.09

A_{ROC} area under the receiver-operating curve, LR^+ likelihood ratio for a positive result, LR^- likelihood ratio for a negative result

excellent sensitivity = 0.93 and reasonable specificity = 0.78—to be the most appropriate demarcation point across the sample sets. The lowest AVQI cut-off score was obtained for Ge5 (i.e., AVQI = 3.05, with sensitivity = 0.98 and specificity = 0.75), and the highest threshold point was found for Du2 (i.e., AVQI = 3.66, with sensitivity = 0.85 and specificity = 0.80).

Finally, because of obvious base-rate differences within the sample (i.e., the disproportionate number of dysphonic samples as compared to normophonic subjects; 80 vs. 20 %, respectively), likelihood ratios were also calculated. The higher the LR^+, the more confident the clinician can be that a person with a higher AVQI-score is voice-disordered/dysphonic. A $LR^+ \geq 10$ indicates that a positive AVQI-score (i.e., above the language-specific AVQI threshold score) is very likely to have come from a dysphonic person. The LR^+ data never reached threshold of $LR^+ \geq 10$. Except for the $LR^+ = 9.00$ of En3, LR^+ ranged from 3.23 to 5.28. This resulted in an intermediate $LR^+_{mean} = 4.85$ and implies that elevated AVQI scores do not always correspond with dysphonia. The lower the LR^-, the more confident the clinician can be that a person with a low AVQI score (i.e., below the language-specific AVQI cut-off point) is normophonic. A $LR^- \leq 0.10$ indicates that a low AVQI score is very likely to have come from a person without dysphonia. Consequently, the $LR^-_{mean} = 0.09$ in this study indicates that the lower AVQI scores sufficiently correspond with normophonia. This was especially the case for the En4, Ge5 and Fr6 samples (i.e., with very reasonable $LR^- = 0.06$, $LR^- = 0.03$ and $LR^- = 0.04$, respectively). For the Du1 and En3 samples, LR^- data of 0.11 were found. To summarize, across the A_{ROC}, sensitivity, specificity, LR^+ and LR^- statistics, the results indicate the En3 samples tended to yield the best diagnostic precision.

Discussion

Different phonetic contexts and speech behaviors may contribute to variance in vocal physiology, and thus phonatory output. Over the last decade, there has been a growing interest in acoustically quantifying dysphonia severity in continuous speech [4–6, 20, 21]. However, one should be aware of the potential for variability in acoustic data based on the influence of phonetic differences across running speech fragments, languages, dialects, etc. The AVQI [4], for example, is a method that concatenates continuous speech and sustained vowels to acoustically estimate dysphonia severity. Consequently, the AVQI may be more susceptible to speech-induced phonatory variability than other methods relying solely on a sustained vowel. The AVQI's validity has been previously validated using two specific Dutch sentences. However, it is unclear how the AVQI performs (i.e., in terms of validity and accuracy) in languages aside from Dutch. Therefore, cross-linguistic validation of the AVQI is necessary to establish its generalizability.

Belgium, for example, is a multilingual country with Dutch-, French- and German-speaking citizens. Furthermore, Belgians are commonly confronted with English in international and educational contexts and via media. For the AVQI to be administered in languages other than the original Dutch sentences (i.e., "Papa en Marloes staan op het station. Ze wachten op de trein."), it is important to investigate its cross-linguistic robustness. The present study was designed to answer the following question: how is the AVQI's concurrent and diagnostic validity in Dutch, German, French and English?

The results of this study indicate that the performance of the AVQI is relatively insulated from inter-language phonetic differences. The correlation coefficients between perceived and estimated dysphonia severity ranged from 0.781 to 0.868 and were equally strong, regardless the language. Furthermore, all A_{ROC}-data ranged between 0.869 and 0.958, indicating generally strong and equivalent discriminatory performance across languages (i.e., to distinguish between normophonia and dysphonia). These data are comparable with and sometimes even exceeded the data from the initial study of Maryn et al. [4], i.e., $r_s = 0.781$ and $A_{ROC} = 0.895$, and the external cross-validation of Maryn et al. [5], i.e., $r_s = 0.796$ and $A_{ROC} = 0.920$. The AVQI can be considered a cross-linguistically robust measure of dysphonia severity.

Caveats, limitations and future research

Several caveats regarding the findings are worth noting. First, the 50 samples were obtained from the same subjects in all languages. This resulted in a sustained vowel that was essentially identical across the six sets of sentences (and concatenated samples). Given the prominent role, the sustained vowel plays in the AVQI, this might have positively influenced (i.e., inflated) the correlational and diagnostic statistics. Furthermore, this means that the raters would have been confronted with six times the same 50 sustained vowels, and thus that their perceptual ratings could be influenced by a familiarity effect. On the other hand, this familiarity effect was theoretically even more present in the acoustic algorithms that treated a sustained vowel equally every time it was repeated. With the sustained vowel kept invariant, the small differences in AVQI between and within languages presumably resulted from variations in the connected speech recordings. However with this study, we exactly aimed at assessing the impact of differences between connected speech samples in different languages. To rule out the familiarity effect, voice recordings in future

research employing the AVQI's or another method's cross-linguistic robustness ideally are derived from separate groups per language/set of running speech fragments.

Second, the degree in which listeners are familiar with a specific language and their native language experience impact their sensitivity to acoustic cues of voice quality and their perceptual strategy [22]. In the design of future studies of cross-linguistic characteristics of acoustic dysphonia severity measures, therefore, it will be crucial to take this factor into consideration. However, in all languages of the present study, voice quality variations are neither allophonic nor phonemic. This means that, for example, breathiness and/or roughness are not a normal part of the Dutch, French, German and English phonology and prosody; and that voice quality only varies at a post-lexical level or changes only with pathology. The three Dutch listeners, being similarly acquainted with these languages, all with acceptable intra-rater reliability, and all presumed to apply the same perceptual strategy, were thus equally sensitive to the acoustic voice quality cues across the languages. In this regard, Hartelius et al. [23] investigated whether perceptual ratings of 33 speech dimensions of dysarthric Australian and Swedish speakers (with multiple sclerosis) depended on whether speakers and listeners/judges speak the same or different language. The results showed high inter-rater reliability irrespective of the listener's knowledge of the speaker's language, and thus indicated that the perceptual impression of many of these dimensions is language independent. Ghio et al. [24] compared perceptual dysphonia ratings across French and Italian listeners and speakers, and they also concluded that overall voice quality perception is not language dependent.

Finally, the lowest r_s- and A_{ROC}-values were found for the French voice signals. Although these values are not significantly different from the other values, it is interesting because French was the only Roman language in this study and thus linguistically most separated from the language in which the AVQI originated (i.e., Dutch). The other languages were all German and shared a common origin. It would be interesting to study the cross-linguistic validity of the AVQI in other Roman languages (e.g., Spanish and Portuguese) or totally different languages (e.g., Arabic or Chinese) in future research. Furthermore, the four languages in the present study are closely related in terms of articulatory and prosodic physiology, and non-modal voice quality is considered abnormal/pathological in all of them. Frequency counts of the phonemes contained in the samples, revealed significant differences in only 4 of the 15 pairs of phonetic sequences (i.e., 11 of the 15 pairs of strings were phonetically similar). Small differences across these languages could thus have been anticipated. It would, therefore, be interesting in the future to examine the AVQI's cross-linguistic validity in the assessment of less related languages (e.g., Japanese and Chinese), so-called non-modal tonal languages (e.g., Gujarati, Jalapa Mazatec, or Vietnamese) or Khoisan/click languages (e.g., Hadza, Sandawe, !Xóõ or Jul'Hoan), and to investigate the relationship between the AVQI's outcome and the dysphonia severity as perceived by native speakers of such languages.

Conclusion

Based on the current data, the favorable results from earlier reports on the AVQI's performance were corroborated, confirming the AVQI as a promising measure of dysphonia severity in multiple languages (i.e., English, German, Dutch, and French).

Acknowledgments The authors thank Dr. Gwen Van Nuffelen (Department of Communication Disorders, University Hospital of Antwerp, Belgium) for her contributions in the perceptual rating of the many concatenated voice samples.

Declarations The study followed the principles of the Declaration of Helsinki. The authors report no declarations of interest.

Appendix: Number of syllables and phonetic transcriptions of the six sentences that were analyzed in this study

Dutch sample number 1—Du1—17 syllables

"Papa en Marloes staan op het station. Ze wachten op de trein."

[papaɛnmarlustaˈnɔpətstasiɹɔn.zəwaχtənɔpdətrɛʲn]

Dutch sample number 2—Du2—22 syllables

"De noorderwind en de zon waren erover aan het redetwisten wie de sterkste was van hun beiden."

[dənoˈrdərwɪntɛndəzɔnwaˈrənəroˈvəraˈnətredət wɪstənwidəstɛrkstəwasvanynbɛʲdən]

English sample number 3—En3—18 syllables

"The north wind and the sun were arguing 1 day which of them was stronger."

[ðənɔ˞ θwɪntɛnðəsynwə˞ɑ˞ guɪŋwandejwɪtʃɔfðɛ mwɔstrɔŋgə˞]

English sample number 4—En4—19 syllables

"When light strikes raindrops in the air, they act like a prism and form a rainbow."

[wɛnlaˈtstraˈksreˈndrɔpsɪnðejær.θeˈɛktlaˈkə prɪzəmɛntfɔ˞ məreˈnbow]

German sample number 5—Ge5—22 syllables

"Einst stritten sich Nordwind und Sonne, wer von Ihnen beiden woll der stärkere wäre."

[aˈnstʃtrɪtənzɪχnɔrtwɪntuntzɔnə.wervɔninən baˈdənvolderʃtɛrkərəverə]

French sample number 6—Fr6—21 syllables

"La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort."

[labiːzeləsɔlɛˈsədispytɛ.ʃakɑ̃asyrɔ̃ˈkiletɛləplyfɔˑr].

References

1. Askenfelt AG, Hammarberg B (1986) Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures. J Speech Hear Res 29:50–64
2. Parsa V, Jamieson DG (2001) Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. J Speech Lang Hear Res 44:327–339
3. Zraick RI, Wendel K, Smith-Olinde L (2005) The effect of speaking task on perceptual judgment of the severity of dysphonic voice. J Voice 19:574–581
4. Maryn Y, Corthals P, Van Cauwenberge P, Roy N, De Bodt M (2010) Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. J Voice 24:540–555
5. Maryn Y, De Bodt M, Roy N (2010) The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders. J Commun Disord 43:161–174
6. Yiu E, Worrall L, Longland J, Mitchell C (2000) Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding. Clin Linguist Phon 14:295–305
7. Ladefoged P (1983) The linguistic use of different phonation types. In: Bless D, Abbs J (eds) Vocal fold physiology: contemporary research and clinical issues. College-Hill Press, San Diego, pp 351–360
8. Schulman R (1989) Articulatory dynamics of loud and normal speech. J Acoust Soc Am 85:295–312
9. Higgins MB, Netsell R, Schulte L (1998) Vowel-related differences in laryngeal articulatory and phonatory function. J Speech Lang Hear Res 41:712–724
10. Dromey C, Ramig LO (1998) Intentional changes in sound pressure level and rate: their impact on measures of respiration, phonation, and articulation. J Speech Lang Hear Res 41:1003–1018
11. Cookman S, Verdolini K (1999) Interrelation of mandibular laryngeal functions. J Voice 13:11–24
12. McClean MD, Tasko SM (2002) Association of orofacial with laryngeal and respiratory motor output during speech. Exp Brain Res 146:481–489
13. Reynolds V, Buckland A, Bailey J, Lipscombe J, Nathan E, Vijayasekaran S, Kelly R, Maryn Y, French N (2012) Objective assessment of pediatric voice disorders with the Acoustic Voice Quality Index. J Voice 26:672.e1–672.e7
14. Barsties B, Maryn Y (2012) Der Acoustic Voice Quality Index in Deutsch: ein Messverfahren zur allgemeinen Stimmqualität [The Acoustic Voice Quality Index: toward expanded measurement of dysphonia severity in German subjects]. HNO 60:715–720
15. Hirano M (1981) Psycho-acoustic evaluation of voice. In: Arnold GE, Winckel F, Wyke BD (eds) Disorders of human communication 5, clinical examination of voice. Springer, Vienna, pp 81–84
16. Chan KM, Yiu EM (2002) The effect of anchors and training on the reliability of perceptual voice evaluation. J Speech Lang Hear Res 45:111–126
17. Portney LG, Watkins MP (2000) Foundations of clinical research, applications to practice, 2nd edn. Prentice-Hall, Upper Saddle River
18. Sheskin DJ (1997) Handbook of parametric and nonparametric statistical procedures. CRC Press LLC, Boca Raton
19. Dollaghan CA (2007) The handbook for evidence-based practice in communication disorders. MD Brookes, Baltimore
20. Awan SN, Roy N, Dromey C (2009) Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model. Clin Linguist Phon 23:825–841
21. Lowell SY, Colton RH, Kelley RT, Hahn YC (2011) Spectral- and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker. J Voice 25:e223–e232
22. Kreiman J, Gerratt BR, Khan SD (2010) Effects of native language on perception of voice quality. J Phon 38:588–593
23. Hartelius L, Theodoros D, Cahill L, Lillvik M (2003) Comparability of perceptual analysis of speech characteristics in Australian and Swedish speakers with multiple sclerosis. Folia Phoniatr Logop 55:177–188
24. Ghio A, Weisz F, Baracca G, Cantarella G, Robert D, Woisard V, Fussi F, Giovanni A (2011) Is the perception of voice quality language-dependant? A comparison of French and Italian listeners and dysphonic speakers. In: Proceedings of interspeech 2011, pp 525–528