The Acoustic Breathiness Index (ABI): A Multivariate Acoustic Model for Breathiness

***,†Ben Barsties v. Latoszek, *,‡,§Youri Maryn, ¶,**,††Ellen Gerrits, and *,‡‡,¶¶Marc De Bodt,** *‡‡‡*Antwerp and* §¶¶*Ghent, Belgium, and* †*Nijmegen and* ¶**††*Utrecht, The Netherlands*

Summary: Objective. The evaluation of voice quality is a major component of voice assessment. The aim of the present study was to develop a new multivariate acoustic model for the evaluation of breathiness.

Method. Concatenated voice samples of continuous speech and the sustained vowel [a:] from 970 subjects with dysphonia and 88 vocally healthy subjects were perceptually judged for breathiness severity. Acoustic analyses were conducted on the same concatenated voice samples after removal of the non-voiced segments of the continuous speech sample. The development of an acoustic model for breathiness was based on stepwise multiple linear regression analysis. Concurrent validity, diagnostic accuracy, and cross validation were statistically verified on the basis of the Spearman rank-order correlation coefficient (r_s), several estimates of the receiver operating characteristics plus the likelihood ratio, and iterated internal cross correlations.

Results. Ratings of breathiness from four experts with moderate reliability were used. Stepwise multiple regression analysis yielded a nine-variable acoustic model for the multiparametric measurement of breathiness (Acoustic Breathiness Index [ABI]). A strong correlation was found between ABI and auditory-perceptual rating ($r_s = 0.840$, $P = 0.000$). The cross correlations confirmed a comparably high degree of association. Additionally, the receiver operating characteristics and likelihood ratio results showed the best diagnostic outcome at a threshold of ABI = 3.44 with a sensitivity of 82.4% and a specificity of 92.9%.

Conclusions. This study developed a new acoustic multivariate correlate for the evaluation of breathiness in voice. The ABI model showed valid and robust results and is therefore proposed as a new acoustic index for the evaluation of breathiness.

Key Words: Voice assessment–Voice quality–Breathiness–Acoustic measurement–Auditory-perceptual judgment.

INTRODUCTION

In laryngology and vocology, a main part in voice assessments is the evaluation of vocal quality. Breathiness is one of the major subtypes of vocal quality that refers to abnormal voice quality. Some vocal pathologies are dominantly characterized by breathiness like nodules with medium or large size,[1] acute laryngitis,[1] paralysis or paresis of the recurrent laryngeal nerve,[1,2] and vocal fold bowing associated with presbyphonia.[2] Breathy voices are characterized by turbulent noise during phonation with excessively high frequency resulting from air leakage during glottal closure.[3] The concept of breathiness is auditory-perceptually based, which is the response of the brain to specific acoustic features in the voice. In general, in the evaluation of voice quality, voice clinicians use standardized and quantified auditory-perceptual rating scales like the Grade, Roughness, Breathiness, Asthenia and Strain (GRBAS) scale[4] or Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V).[5] However, the evaluation of voice quality is subjective, which induces notable intra-rater and inter-rater variability by the listeners. A recent overview[6] presented many factors that influence the reliability and accuracy of the perceptual evaluation, which can be categorized in three groups: listener, stimulus, and scale.[6] Despite these limitations, the perceptual evaluation remains the candidate for gold-standard assessment.[6,7] First, voice quality is a perceptual phenomenon by nature and it is related to the response of the brain to specific acoustic features that are mainly associated with periodicity (ie, prominence of fundamental frequency) in the voice signal.[8] This implies that factors attenuating the degree or dominance of vocal periodicity contribute to the perception of increased dysphonia. Second, it is a simple and efficient method in daily clinical practice to document the presence, degree, and progression of any type of abnormal voice quality.[8] Notwithstanding, to improve the validity and reliability in abnormal voice quality judgments, various kinds of instrumental methods have been developed to quantify abnormal voice quality. Among these methods, acoustic measurements have received attention because they are the most used diagnostic instruments to identify voice disorders in research,[9] they use noninvasive technology,[10] they are affordable and easy to use,[10] and they have relatively low costs.[11] As such, acoustic measurements have the potential to offer an objective adjunct to existing perceptual assessments. However, acoustic measurements are traditionally applied to sustained vowels and less frequently to continuous speech.[6] Using acoustic methods to judge voice quality on sustained vowels alone might exceed a limitation in the evaluation of voice quality because the judgment on sustained vowels

Accepted for publication November 22, 2016.

From the *Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium; †Institute of Health Studies, HAN University of Applied Sciences, Nijmegen, The Netherlands; ‡European Institute for ORL, Sint-Augustinus Hospital, Antwerp, Belgium; §Faculty of Education, Health & Social Work, University College Ghent, Ghent, Belgium; ¶Faculty of Health Care, HU University of Applied Sciences Utrecht, Utrecht, The Netherlands; **Faculty of Humanities, University of Utrecht, Utrecht, The Netherlands; ††Department of Otolaryngology, University Medical Center Utrecht, Utrecht, The Netherlands; ‡‡Department of Otorhinolaryngology and Head & Neck Surgery, Antwerp University Hospital, Antwerp, Belgium; and the ¶¶Faculty of Medicine & Health Sciences, University of Ghent, Ghent, Belgium.

Address correspondence and reprint requests to Ben Barsties v. Latoszek, Faculty of Medicine and Health Sciences, University of Antwerp, Universiteitsplein 1, 2610 WILRIJK, Antwerp, Belgium. E-mail: ben.barsties@t-online.de

Journal of Voice, Vol. 31, No. 4, pp. 511.e11–511.e27
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.11.017

does not necessarily correspond with continuous speech.[6] Thus, both speech tasks lower the ecological validity of the judgment of voice quality. Additionally, in past studies, single acoustic measures revealed poor reliability or poor documentation of improvement in voice quality,[12,13] plus many acoustic parameters had low or poor correlation with auditory-perceptual judgment.[3,14] Therefore, a milestone has recently been reached in acoustic measurements by developing two multivariate acoustic models for the evaluation of overall voice quality: the Acoustic Voice Quality Index (AVQI)[15] and the Cepstral Spectral Index of Dysphonia.[16,17] These models have proven to adequately quantify continuous speech and sustained phonation in several studies.[6] It was shown in all studies that these two acoustic models are valid and robust tools.[6]

Meta-analysis identified only a minority (ie, 12 out of 96, or 12.5%) of the acoustic measures to be moderate to strong predictors of auditorily perceived breathiness.[3] The measures that correlated strongest with perceptual ratings of breathiness were cepstral peak prominence (CPP) and smoothed cepstral peak prominence (CPPs),[18,19] glottal-to-noise excitation (GNE) ratio,[20] and high-frequency noise (Hfno).[21] In general, these four acoustic measures revealed the highest outcome to adequately objectify breathiness for a larger group of voice samples.[3] Furthermore, CPP and CPPs were investigated in many studies and were found to be the best predictors in continuous speech and sustained vowels.[3] To summarize, these parameters had normalization features that achieved independence to frequency and sound pressure level and explain the robustness and high concurrent validity in the evaluation of breathiness.[3] However, in the past, several attempts were undertaken to create multivariate models for breathiness to improve their validity and predictive power, which are limited in single acoustic markers.[22–34] Table 1 lists relevant methodological features and outcomes of only the multivariate acoustic models that have been investigated on concurrent validity and classification accuracy in measuring auditory-perceptual judgment of breathiness.

The validity outcomes of Table 1 demonstrate that the use of a multivariate acoustic model in the evaluation of breathiness reached moderate to high results in accuracy (45%–77%) and concurrent validity (r = 0.67–0.92), but the studies listed in the table mostly included small numbers of subjects. Furthermore, all of these models contain only one speech task, whereas it can be beneficial for both perceptual ratings and instrumental analysis to be based on both speech types to be considered ecologically valid.[15] Analogous to the methods of the AVQI, it was assumed that improved acoustic prediction of breathiness may be derived from combining both sustained vowels and continuous speech. To proceed with the example of AVQI, many studies already investigated its concurrent validity[15,35–44] as well as diagnostic precision,[15,35–41,43,44] and they all concluded it to be robust and ecologically valid in objectifying overall voice quality.

The aim of the present study is to define an acoustic multivariate model for the quantification of breathiness that complements the auditory-perceptual assessment of breathiness. This study considered concatenated voice samples, a large dataset, a stricter selection of the rater panel, and various validity investigations.

METHOD

Subjects

All subjects were recruited from the interdisciplinary otolaryngology and speech-language pathology assessment of the otolaryngology caseload of the Sint-Jan General Hospital in Bruges, Belgium, in the period from October 2002 to February 2014. Every voice patient who visited the otolaryngology consultation hour was included for the study. The group consisted of 970 participants with dysphonia and 88 vocally healthy subjects. The dysphonia group presented various organic and nonorganic etiologies and various degrees in dysphonia severity. Laryngological diagnoses were made with a flexible transnasal chip-on-tip laryngostroboscope (Olympus ENF-V, Olympus Medical Systems Europa GmbH, Hamburg, Germany).

Tables 2–4 summarize further details of the dysphonia group. This includes the variety of voice disorders, clinical assessments (ie, voice range profile,[45] speaking fundamental frequency,[46] jitter%,[47] electroglottal-closed-quotient,[48] maximum phonation time,[49,50] vital capacity,[51] phonation quotient,[52] Dysphonia Severity Index,[53] AVQI version 03.01,[42] and Voice Handicap Index[54,55]), and personal details such as age, gender, and occupation. This group of subjects represents a clinical population of nonorganic and organic laryngeal pathologies. Considerations include different ages, gender groups, and different types and degrees of voice quality disruption and vocally induced disability.

Additionally, this study included vocally healthy subjects without any reported voice complaints or history of voice, speech, or hearing problems or disorders. There were 55 women with a mean age of 35.95 years ± 16.18 years and 33 men with a mean age of 34.06 years ± 18.50 years. They were free of any voice abnormality as judged by three experts using the GRBAS scale.[4] The assessment of these vocally normal subjects was limited to the recording of voice samples.

This research was approved by the ethical committee according to International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use-Good Clinical Practice guidelines (ECNR 15/33/343).

Voice sample

All recordings were conducted in a soundproof booth. To verify the level of environmental noise of the voice recordings *post hoc*, guidelines for interpreting signal-to-noise ratio (SNR) by Deliyski et al[56,57] were used. All voice samples were consistent with the recommended SNR norm for acceptable circumstances of acoustic recordings and analysis. The results showed a mean SNR of 38.56 dB and standard deviation of 3.78 dB.

The voice samples were recorded using an AKG C420 head-mounted condenser microphone (AKG Acoustics, Munich, Germany) with a mouth-to-microphone distance of about 10 cm and 45° azimuthal angle. The recordings were digitized with a sampling rate of 44.1 kHz and 16 bits of resolution using the *Computerized Speech Lab* model 4500 (Kay Pentax, Lincoln Park, NJ).

Every voice sample from each participant contained the concatenation of the first 34 syllables (ie, "Papa en Marloes staan op het station. Ze wachten op de trein. Eerst hebben ze een kaartje gekocht. Er stond een hele lange rij") of a Dutch phonetically

TABLE 1.
Multivariate Acoustic Models in the Evaluation of Breathiness Summarized in Their Methodology Features and Outcome

Source	Number of Subjects	Speech Task	Multivariate Statistical Method	Included Objective Measures in the Model	Scale of the Auditory-Perceptual Judgment of Breathiness	Outcome	
						Concurrent Validity	Classification Accuracy (%)
Hammarberg et al[22]	17	A short story of 92 words	Stepwise multiple regression analysis	Differences between the amplitudes of 0–2 kHz and 2–5 kHz, and between 2–5 kHz and 5–8 kHz; mean fundamental frequency	EAI-5 points	Multiple correlation: r = 0.69	47%
Hammarberg et al[23]	16	A short story of 92 words	Stepwise multiple regression analysis	Long-term average spectrum between 4 and 6 kHz; Long-term average spectrum between 5 and 10 kHz; level of the fundamental	EAI-5 points	Multiple correlation: r = 0.83	69%
Wolfe et al[27]	102	Sustained vowel /a/ and /i/	Stepwise multiple regression analysis	Shimmer in percent	EAI-7 points	Multiple correlation: r = 0.67	45%
Stráník et al[34]	593	Phonetically balanced text of 34 words	MP5 decision tree	F_0 estimated in the cepstral domain; glottal-to-noise excitation ratio; high- to mid- or low-frequency energy; ratio of number of samples in voiced parts to the number of all samples in record	EAI-4 points	Pearson correlation: r = 0.92	77%

Abbreviations: EAI, equal-appearing interval; F_0, fundamental frequency.

TABLE 2.
List of Demographic Characteristics and Type of Voice Disorders From 970 Subjects With Dysphonia

Variable	Results
Gender	
Male	353 (36.4%)
Female	617 (63.6%)
Age in years (mean ± SD)	42.40 ± 21.13
Voice disorder	
Functional dysphonia	231 (23.81%)
Nodules	201 (20.72%)
Paralysis or paresis	132 (13.61%)
Polypoid mucosa (edema)	73 (7.53%)
Cyst	35 (3.61%)
Reflux laryngitis	28 (2.89%)
Polyp	28 (2,89%)
Presbylarynx	26 (2.68%)
Tumor	23 (2.37%)
Chronic laryngitis	22 (2.27%)
Post phonosurgery	17 (1.75%)
Thyroidectomy	15 (1.55%)
Sulcus vocalis	14 (1.40%)
Trauma	12 (1.24%)
Ventricular hypertrophy	12 (1.24%)
Acute laryngitis	11 (1.13%)
Leukoplakia	10 (1.03%)
Other benign voice disorders	40 (4.12%)
Other voice disorders related to surgery	20 (2.06%)
Other neurologic voice disorders	20 (2.06%)
Profession	
Pensioner	197
Pupil	135
Teacher	110
Student: future professional voice user	69
Child care worker	40
Secretary or administrative work	39
Salesman	34
Student: non–future professional voice user	30
Manager	28
Nurse	27
Housewife	17
Nonworker	14
Social worker	12
Cleaning woman	11
Waiter	10
Consulting	10
Others without further clustering or a low number of a specific profession	140
No data available in the medical file	47

balanced text ("Papa en Marloes")[58,59] and a sustained phonation of 3 seconds without voice onset and voice offset from the vowel [a:]. For both recordings, the participants used comfortable pitch and loudness, and voice samples were saved in WAV format. The application of these selected durations of the two speech types was found as highly ecologically valid because of equal proportion of continuous speech and sustained vowel segments in acoustic analyses.[39] Furthermore, this length of concatenated voice samples showed a consistent rating of voice quality in perceived judgment estimating the presence and degree of severity of a voice.[8]

Auditory-perceptual judgment

For the auditory-perceptual judgment of breathiness, an expert panel of 12 native Dutch speech-language therapists rated the breathiness severity. The panel consisted of nine women and three men specialized in voice disorders with a professional experience in auditory-perceptual judgment ranging from 4 to 41 years (mean = 22.3 years, and standard deviation = 11.4 years). Each listener rated the breathiness severity of each concatenated voice sample with one judgment for the whole sample (ie, one single wave sound; see Figure 1). They used Hirano's Breathiness (B) from the GRBAS scale,[4] which represents the degree of the extent of air leakage through the glottis. As recommended by Wuyts et al,[60] the judges used the ordinal four-point equal-appearing interval scale (ie, 0 = normal voice or absence of breathiness, 1 = slightly breathy, 2 = moderately breathy, and 3 = severely breathy). All voice samples were presented in a quiet room with a low ambient noise level lower than 40 dB_A, measured with a calibrated PCE-322A sound level meter (PCE Inst., Meschede, Germany). They were presented to each listener individually at a comfortable loudness level through an external soundcard from Creative Soundblaster x-fi 5.1. USB (Creative Technology LTD, Singapore) and a Beyerdynamic DT 770 PRO 80Ω headphone (Beyerdynamic GmbH & Co. KG, Heilbronn, Germany). Every listener was allowed to repeat each voice sample as often as necessary before making a final decision.

All voice samples were judged randomly, totaling four to five sessions. Every rating session contained about 250 voice samples with a duration of about 2 hours. Furthermore, all judges were blinded regarding the identity, diagnosis, and disposition of the voice samples. To assess intra-rater reliability, 104 voice samples, approximately 10% of the 1058 voice samples, were selected randomly. These voice samples were repeated a second time at the end of the perceptual judgment without informing the listeners that stimuli were repeated.

Internal factors such as fatigue, attention, and low concentration may contaminate auditory ratings[61] and were therefore controlled by using a short break after every 25th rating. Furthermore, as recommended by Chan and Yiu,[62] anchor voices were used to putatively increase the reliability of listener ratings by judging voice quality. Thus, six samples of concatenated continuous speech and sustained phonation were selected from the database from previous investigations. The selection criteria of the anchor voices were based on prior unanimous agreement across judges adhering to the three severity degrees of slightly, moderately, and severely hoarse. Two sets of samples with

TABLE 3.
List of Clinical Assessments in the Dysphonia Group From 970 Subjects

Clinical Assessment Category	Parameters	Mean	SD	Range	No Data Available*
	Sex-independent parameters				
Acoustics: voice range profile	Frequency range (in semitones)	27.76	7.64	2–62	43
	Intensity range (in dB)	47.00	10.98	1–70	43
Acoustics: periodicity	Jitter%	2.358	2.210	0.028–24.445	69
Multiparametric indices	Dysphonia Severity Index	0.19	4.54	–30.25 to 10.64	133
	Acoustic Voice Quality Index version 03.01	3.89	2.01	–0.51 to 11.34	0
Subjective (self-)evaluation by patient	Voice Handicap Index—Total	42.42	22.61	0–108	188
	Voice Handicap Index—Functional	11.36	8.24	0–40	188
	Voice Handicap Index—Physical	19.23	8.08	0–40	188
	Voice Handicap Index—Emotional	11.79	8.87	0–38	188
Aerodynamics	Phonation quotient (mL/s)	269.98	155.43	68.50–1475.00	56
Electroglottography	Electroglottal-closed-quotient (in %)	43.83	8.56	26.26–182.63	528
	Sex-dependent parameters				
Acoustics: speech range profile	Speaking fundamental frequency (in Hz)				
	Male	154.40	49.84	78.46–300.21	79
	Female	196.67	48.00	71.8–984.23	105
Acoustics: voice range profile	Highest frequency (in Hz)				
	Male	545.73	238.92	69.30–2352.30	24
	Female	698.99	230.25	146.83–1567.98	15
	Lowest frequency (in Hz)				
	Male	110.23	52.43	58.26–554.37	26
	Female	141.01	57.83	55.00–886.00	16
	Maximum intensity (in dB)				
	Male	103.91	9.46	65–127	25
	Female	103.32	8.40	64–120	16
	Minimum intensity (in dB)				
	Male	58.11	6.22	41–79	24
	Female	56.47	5.60	44–85	15
Aerodynamics	Vital capacity				
	Male	3705.36	1286.71	900–6850	36
	Female	3077.01	837.27	800–5600	21
	Maximum phonation time				
	Male	15.26	8.47	1.47–49.80	20
	Female	14.24	6.25	1.13–47.36	10

* No data were available because of no sufficient periodicity (ie, voice range profile, Jitter%, Dysphonia Severity Index, electroglottography), unworkable assessments in children's voices like Dysphonia Severity Index and Voice Handicap Index, and practical reasons of consultation hour.

continuously increasing hoarseness (ie, three samples per set) were played for the listeners as anchors (ie, one for breathiness, one for roughness). Each listener heard these two sets at the beginning and after the break of every 25th rating.

Acoustic measures

All acoustic analyses were applied to only an appendage of voiced segments of continuous speech (voiced segment extraction and concatenation was done with the *Praat* script of Maryn et al[15]) with a 3-second mid-[a:] segment (see Figure 2). These chained sound files were analyzed with the freely available and downloadable software package "*Praat*" version 5.3.57 (Paul Boersma and David Weenink; Institute of Phonetic Sciences, University of Amsterdam, The Netherlands: http://www.praat.org/). Based on the meta-analysis regarding the correlation between acoustic measurement and auditory-perceptions of breathiness,[3] a list with detailed information of 28 acoustic voice markers (Table 4) was retained. These markers (1) hold promise in objectively quantifying auditorily perceived breathiness, and (2) could be determined or calculated in *Praat*. A custom-made *Praat* script designed by one of the authors (Y.M.) was used to automatically calculate these 28 acoustic measures and to store their numerical output.

Statistical analysis

All statistical analyses were completed using *SPSS Statistics* 22.0 (IBM Corp., Armonk, NY), except when otherwise stated. First, the intra-rater reliability of the 12 raters was assessed using the Cohen kappa coefficient (C*k*) analyzed with the R-Studio v3.0.1 software package (R Core Team, Vienna, Austria). This statistic is a chance-corrected index of the agreement between the ratings of two judges or between two ratings, yielding values

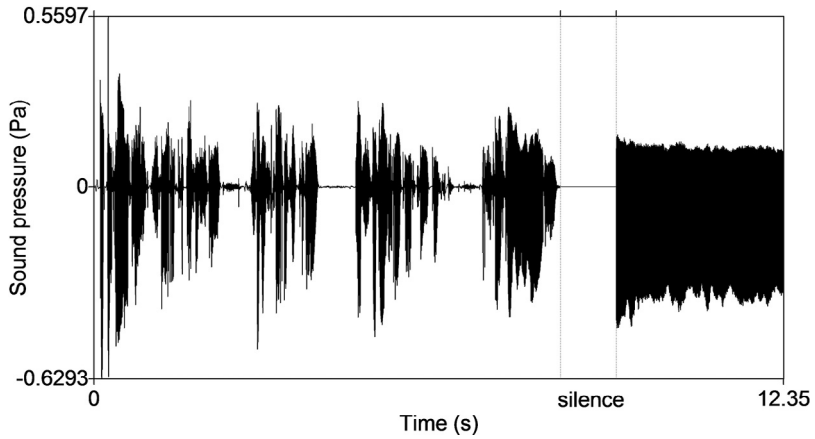

FIGURE 1. Oscillogram of a concatenated voice sample (derived from subject 979), as used in the auditory-perceptual evaluation of this study. It is divided in three areas. The left portion reflects the first 34 syllables of the "Papa en Marloes" text. The right area reflects the middle 3 seconds of a sustained /a/. Both samples were separated by 1 second of silence (area in the middle marked with dashed lines).

of C*k* = 1 for perfect agreement and C*k* = 0 when agreement is no better than that by chance.[64] To assess the agreement of inter-rater reliability among the 12 judges, we computed the kappa coefficient according to Fleiss,[65] who extended the C*k* for more than two judges. The Fleiss kappa (F*k*) was determined by also using the software package of *R-Studio* v3.0.1. Both C*k* and F*k* reached an acceptable reliability level at minimally moderate agreement from $k \geq 0.41$.[66] Furthermore, significant changes (ie, considered statistically significant at $P \leq 0.01$) in all kappa values were tested using bootstrapping with 10,000 replications based on a script by Van Belle.[67] To establish a group of raters with a homogeneous and high level of reliability, the following criteria were followed (next to their long-standing experience in clinical rating voice quality as speech-language therapists), which are identical to the method described by Barsties and Maryn[41]: (1) no significant differences were found in the intra-rater C*k* results between all pairs of raters; (2) each rater reached intra-rater reliability with a level of $Ck \geq 0.41$[66]; (3) all remaining raters with representative and comparably high intra-rater reliability were analyzed to find a homogenous rater group with inter-rater reliability of $Fk \geq 0.41$.[66] If the F*k* result is significantly better by excluding a rater, the rater with the highest significant value has to be excluded for the next round. Thus, in each round, we used a backward stepwise method to exclude the rater with the highest significant kappa value in comparison with the F*k* for all tested raters. This procedure was repeated until a minimum kappa value of ≥0.41 was achieved without significantly better F*k* results for one rater of the group who was excluded in comparison with the F*k* for all tested raters.

Second, to strengthen the validity between acoustic measurement and the selected rater panel of auditory-perceptual judgment, further analysis was performed only on voice samples that were labeled as normal voice or absence of breathiness, slightly breathy, moderately breathy, and severely breathy with a minimum of 50% consensus of the perceived judgment. Thus, this methodological step prevents including only clear classified voice samples.

Third, to assess the predictive validity between the single acoustic measures and the B_{mean} (ie, average B-score over the abovementioned selected rater panel with the best reliability) in the concatenated voice samples, the Spearman rank-order

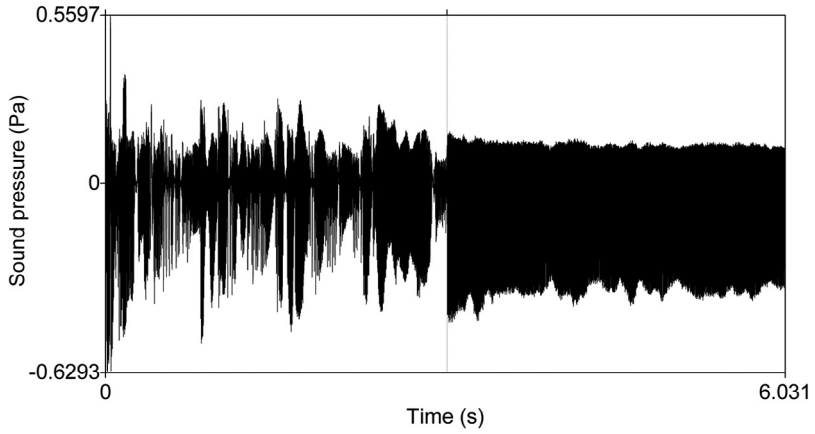

FIGURE 2. Oscillogram of a concatenated voice sample (derived from subject 979), as used for the acoustic analyses of this study. It is divided in two areas separated by a gray line in the figure. The left area reflects the concatenated voiced segments of the first 34 syllables of the "Papa en Marloes" text. The right area reflects the middle 3 seconds of a sustained /a/.

TABLE 4.
List of 28 Acoustic Measures of the Custom-made *Praat* Script

Category	Acoustic Measures	Abbreviation
Fourier and linear prediction coefficients spectra	*Four* parameters of relative level of high-frequency noise between energy: – from 0 to 6 kHz and energy from 6 to 10 kHz – from 0 to 1 kHz and energy from 1 to 10 kHz – from 0 to 2 kHz and energy from 2 to 10 kHz – from 0 to 1 kHz and energy from 1 to 4 kHz	Hfno-6000 Hz Hfno-1000 Hz Hfno-2000 Hz Hfno-4000 Hz
	Harmonics-to-noise ratio from Dejonckere and Lebacq,[63] which analyzes the harmonic emergence of the spectral display by comprising the frequency bandwidth between 500 Hz and 1500 Hz. A cepstrum was performed to determine F_0 and thus to localize the harmonic structure in the long-term average spectrum.	HNR-D
	Differences between the amplitudes of the first and second harmonics in the spectrum. To localize the first harmonic peak, a cepstrum was performed for F_0 determination.	H1-H2
	The smoothed cepstral peak prominence is the distance between the first harmonic's peak and the point with equal quefrency on the regression line through the smoothed cepstrum.	CPPs
	Harmonics-to-noise ratio is the base-10-logarithm of the ratio between the periodic energy and the noise energy, multiplied by 10	HNR
	General slope of the spectrum is defined as the difference between the energy in 0–1000 Hz and the energy in 1000–10,000 Hz of the long-term average spectrum.	Slope
	Tilt of the regression line through the spectrum is the difference between the energy in 0–1000 Hz and the energy in 1000–10,000 Hz of the trendline through the long-term average spectrum.	Tilt
Frequency short-term perturbation measures	The period standard deviation is the variation in the standard deviation of periods in which the length of the sample is important for a valid computation of the standard deviation.	PSD
	To correct the feature of nonlinearity in PSD, a natural logarithm is used.	LNPSD
Frequency short-term perturbation measures	*Three* jitter variations: – Jitter local is the average difference between successive periods, divided by the average period – Jitter of relative average perturbation is the average absolute difference between a period and the average of it and its two neighbors, divided by the average period – Jitter of five-point period perturbation quotient is the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period	Jit, Jit-RAP, Jit-PPQ
Amplitude short-term perturbations measures	*Five* shimmer variations: – Shimmer local is the absolute mean difference between the amplitudes of successive periods, divided by the average amplitude – Shimmer local dB is the base-10-logarithm of the difference between the amplitudes of successive periods, multiplied by 20 – Shimmer of the three-point amplitude perturbation quotient is the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude – Shimmer of the five-point amplitude perturbation quotient is the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbors, divided by the average amplitude – Shimmer of the 11-point amplitude perturbation quotient is the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its 10 closest neighbors, divided by the average amplitude	Shim, Shim-dB, Shim-APQ3, Shim-APQ5, Shim-APQ11
Combines spectral and perturbation features	*Eight* acoustic measures based on the glottal-to-noise-excitation (GNE) ratio:[20] maximum frequency of 4500 Hz and 3500 Hz, the GNE mean with a bandwidth of 1000 Hz and 3000 Hz, the GNE sum with a bandwidth of 1000 Hz and 3000 Hz, and the GNE standard deviation with a bandwidth of 1000 Hz and 3000 Hz	GNEmax-4500 Hz, GNEmax-3500 Hz GNEmean-1000 Hz, GNEmean-3000 Hz GNEsum-1000 Hz, GNEsum-3000 Hz GNEsd-1000 Hz, GNEsd-3000 Hz

correlation coefficient (r_s) and the coefficient of determination (r_s^2) between B_{mean} and the 28 acoustic measures were calculated. Interpretation guidelines for r_s were provided by Frey et al.[68]

Fourth, to construct an acoustic model using the combination of the best acoustic predictors of breathiness, a stepwise multiple linear regression was applied. A multiple regression equation was constructed based on the unstandardized coefficients of the statistical model with a level of significance at $P < 0.05$. Furthermore, the multiple regression model was tested for its normality, linearity, homoscedasticity, and independence.[69] The linearity was inspected for plot regression-standardized residuals and regression-standardized predicted values with the dependent variable of auditory-perceptual judgment of breathiness. Judging the linearity assumption, we need to show a randomized distribution of negative and positive values with no obvious pattern in the plot. To test the independence, the Durbin-Watson statistic was used. This statistic provides a test for significant residual autocorrelation. Values closer to 2.0 mean that residuals are uncorrelated and confirm independence.[70] Values less than 1.0 or greater than 3.0 are definitely a cause for concern in independence.[70] The homoscedasticity is evaluated with the Breusch-Pagan test. Homoscedasticity is not present if the *P* value is less than $P < 0.05$. For testing the normality of the whole distribution, the Anderson-Darling test was used.[71] Normality is not present if the *P* value is less than $P < 0.05$. Additionally, the multicollinearity was analyzed to avoid confounding of high correlations between the acoustic variables. Therefore, we used the variance inflation factor (VIF), which is the reciprocal of tolerance: $1/(1\text{-}r^2)$. It indicates the degree to which the standard errors are inflated because of the levels of multicollinearity. VIF values of 10 or greater were shown as indicative of problematic multicollinearity.[70]

Finally, to simplify the interpretation of the scores of the equation for clinical use, this model was linearly rescaled so that the outcomes of the equation ranged from 0 to 10. Finally, the rescaled model was called Acoustic Breathiness Index (ABI).

Fifth, to investigate the concurrent validity of ABI, the r_s and r_s^2 between B_{mean} and ABI were calculated. The interpretation of the r_s outcome was addressed to the guidelines by Frey et al.[68]

Sixth, to examine the diagnostic accuracy of ABI, several estimates of the receiver operating characteristics (ROC) and likelihood ratio (LR) were evaluated. Diagnostic precision of ABI was evaluated by its sensitivity (ie, correctly identified breathiness which tested positive on ABI) and specificity (ie, correctly identified breathiness when they tested negative on ABI) related to the ROC outcome. The sensitivity and specificity can vary depending on the chosen threshold of ABI to define a positive result. This trade-off between sensitivity and specificity was graphically produced by generating the ROC curve. To create the ROC curve from ABI, a point, per ABI threshold, was plotted, which represented the true-positive rate (ie, sensitivity) on the ordinate and the false-positive rate (ie, 1 − specificity) on the abscissa. A voice sample was considered without breathiness when modal agreement between the selected judges scored a voice sample with a mean breathiness result of $B_{mean} < 0.5$. A breathy voice was considered as $B_{mean} \geq 0.50$. Thus, breathiness ratings ranged from $B_{mean} \geq 0.50$ to ≤3. Furthermore, the ability of ABI to discriminate between normal and breathy voices was represented by the "area under ROC curve" (A_{ROC}). An $A_{ROC} = 1.0$ is found for measures that perfectly distinguish between normal and breathy voices. An $A_{ROC} = 0.5$ corresponds with chance-level diagnostic accuracy.[72] Additionally, to provide further evidence regarding the value of a diagnostic measure and to help reduce problems with sensitivity or specificity related to the base-rate differences in the samples (ie, the uneven percentages of 8% normophonia, and 92% dysphonia in the 1058 voice samples), LRs should also be calculated.[73] The "LR for a positive result" (LR+) yields information regarding how the odds of the disease increase when the test is positive. LR+ provides information regarding the likelihood that an individual is breathy when testing positive. The "LR for a negative result" (LR−) is an estimate that helps to determine if an individual does not have a particular disorder when testing negative on the diagnostic test. LR− provides information regarding the likelihood that an individual has no breathiness when testing negative. As a general guideline, the diagnostic value of a measure is considered to be high when LR+ ≥10 and LR− ≤0.1.[73] Because LR statistics consider sensitivity and specificity simultaneously, they are less vulnerable to sample size characteristics and base-rate differences in the sample between normal and breathy voices.[73]

Seventh, cross-cohort validation was achieved on the number of all selected voice samples. Therefore, a new selected set of these data was used because concurrent validity might differ from the one on which it was initially modeled. This methodological step was necessary to explore whether the initial model loses accounted variance (r_s^2) and concurrent validity when used on groups other than the group on which it was developed. Therefore, r_s scores between B-scores and ABI scores were calculated for 50 randomly selected subgroups of 10, 50, 100, 250, and 400 voice samples. This method of cross-cohort validation is similar to a method described by Maryn et al.[15]

RESULTS

Reliability of auditory-perceptual judgment and final selection of the rater panel

Intra-rater reliability showed no significant differences in C*k* values ($t = 11.509$, $P = 0.403$) between all 12 raters, but four raters did not reach the minimum of the acceptable reliability level (C*k* = 0.32 to 0.39) and had to be excluded. The remaining eight raters had a range of C*k* between 0.43 and 0.54.

Inter-rater reliability was executed on the remaining eight raters who reached an F*k* of 0.31, and a significantly better F*k* result was found if three raters were excluded ($t = 13.9$, $P = 0.000$, to $t = 43.195$, $P = 0.000$). After the fifth round, the F*k* increased sufficiently and this was the first time the minimal rater reliability of F*k* = 0.41 was reached in the group with four raters. Unfortunately, there was still a significantly better F*k* result if one rater was excluded ($t = 6.78$, $P = 0.009$). Avoiding the exclusion of too many raters (ie, in the worst case, only a single rater remains), we decided to deviate from our criteria. Thus, we followed only the criteria of a rater panel with acceptable rater reliability and tolerated a significant improvement in rater reliability by excluding further raters. Finally, all analyses of perceptual B_{mean}

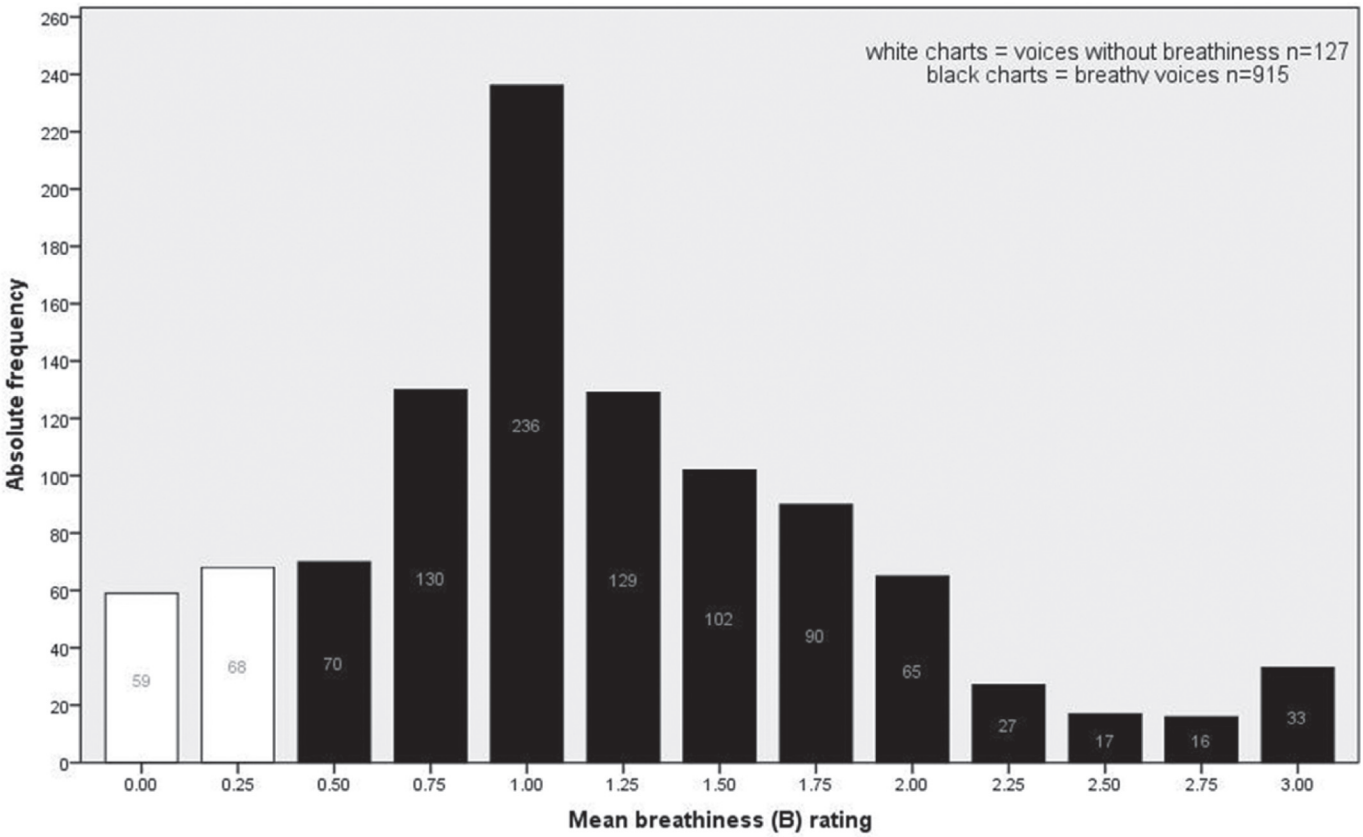

FIGURE 3. Frequency distribution of the mean auditory-perceptual breathiness ratings (average of B-scores of the four identified judges) of the 1042 concatenated voice samples.

ratings were conducted on the panel of the four raters mentioned above. This rater panel judged the breathiness severity of the 1058 voice samples. The breathiness judgment of only 16 voice samples showed a lower than 50% consensus between the four raters. This was necessary to assign each voice sample to a specific breathiness severity. In these 16 cases, the specific assignment of a severity was not feasible and thus, these voice samples were excluded from further analyses.

Figure 3 shows the distribution of the 1042 evaluated voice samples by this panel of the four selected raters. These samples were used for further analyses.

Concurrent validity of single acoustic measures and multivariate model

Table 5 summarizes the descriptive data for the 28 acoustic measures by separating the voice samples between perceptually judged non-breathy voices (n = 127) and breathy voices (n = 915). The correlations (r_s) and coefficients of determination (r_s^2) between B_{mean} and these 28 acoustic measures are shown in Table 6. The concurrent validity of the single acoustic measures revealed a wide range of high absolute r_s-values[68] (eg, CPPs: $r_s = 0.768$ and most of GNE parameters ranging from $r_s = 0.718$ to $r_s = 0.705$) and very low absolute r_s-values (eg, PSD: $r_s = 0.069$, LNPSD: $r_s = 0.069$, H1-H2: $r_s = 0.098$). The strongest correlation was identified for CPPs ($r_s = –0.768$), which explained 59% of the variation of the auditory-perceptual judgment of breathiness. Furthermore, with the exception of PSD and LNPSD, for which low significant correlations were found ($P < 0.05$), all correlations were significant at $P < 0.01$.

Furthermore, the statistics of normality, linearity, homoscedasticity, independence, and multicollinearity showed the following results for the ABI model. The linearity of the model had a randomized distribution of negative and positive values with no obvious pattern in the plot. The independence of the model showed a value of 1.975 in the Durbin-Watson test and it confirmed that the residuals are uncorrelated. There were also no problems in multicollinearity because the VIF was less than 10 in all acoustic parameters. However, the results of the Breusch-Pagan test confirmed heteroscedasticity of the model, because the P level was $P < 0.000$.

Heteroscedasticity exists through more deviation of some ABI results related to the perceived breathiness degree. The prominence breathiness degrees with the highest deviation between ABI scores and perceived breathiness judgment were at the degrees of slightly and severely breathy voices. More disagreement at the perceived degrees of slight breathiness judgment was found,[6] and the higher level of deviation could also be explained for the current data. At severely breathy degrees, acoustic measurements showed numerous variations of consistency because the voice signal is irregular or aperiodic in nature. These findings are especially present in acoustic measures that are based on fundamental frequency.[74] This effect of variation might explain

TABLE 5.
Descriptive Outcomes of the 28 Acoustic Measures Between Non-breathy and Breathy Voice Samples

	Non-breathy Voices (n = 127)				Breathy Voices (n = 915)			
Acoustic Measures	Mean	SD	Maximum	Minimum	Mean	SD	Maximum	Minimum
Hfno-1000 Hz (dB)	1.53	0.88	1.80	1.30	1.60	0.12	1.91	1.11
Hfno-2000 Hz (dB)	1.80	0.13	2.13	1.49	1.83	0.13	2.27	1.29
Hfno-4000 Hz (dB)	1.40	0.07	1.62	1.20	1.47	0.10	1.73	1.02
Hfno-6000 Hz (dB)	2.09	0.15	2.60	1.74	1.96	0.20	2.82	1.15
HNR-D (dB)	25.14	4.43	33.30	14.01	25.78	4.33	34.30	14.40
GNEmean-1000 Hz	0.31	0.03	0.36	0.25	0.26	0.04	0.36	0.10
GNEsum-1000 Hz	809.22	66.10	949.28	641.39	673.74	108.59	945.52	270.73
GNEsd-1000 Hz	0.40	0.03	0.47	0.32	0.33	0.05	0.46	0.13
GNEmax-4500 Hz	0.92	0.04	0.99	0.75	0.81	0.11	0.99	0.33
GNEmean3000 Hz	0.04	0.003	0.05	0.03	0.03	0.006	0.05	0.01
GNEsum3000 Hz	18.10	1.33	20.32	13.78	14.81	2.51	20.31	4.75
GNEsd-3000 Hz	0.18	0.01	0.21	0.14	0.15	0.03	0.21	0.05
GNEmax-3500 Hz	0.88	0.06	0.98	0.69	0.72	0.12	0.97	0.30
H1-H2 (dB)	3.37	2.75	13.26	–4.14	5.19	4.08	16.93	–11.31
PSD (s)	0.001	0.0004	0.002	0.0003	0.0009	0.0006	0.004	0.0002
LNPSD (s)	–6.97	0.38	–5.91	–8.05	–7.18	0.56	–5.46	–8.80
CPPs (dB)	14.38	1.70	18.21	7.47	10.82	2.52	17.63	2.62
Jit (%)	1.72	0.55	4.58	0.80	1.82	0.92	8.90	0.74
Jit-RAP (%)	0.77	0.31	2.61	0.33	0.89	0.53	5.59	0.24
Jit-PPQ (%)	0.87	0.33	2.79	0.43	0.99	0.58	6.19	0.30
Shim (%)	4.38	1.62	13.40	2.14	5.92	3.35	23.38	1.85
Shim-dB (dB)	0.47	0.13	1.23	0.28	0.59	0.28	1.93	0.30
Shim-APQ3 (%)	1.71	0.84	7.47	0.82	2.65	1.73	11.00	0.73
Shim-APQ5 (%)	2.21	0.86	7.23	0.85	3.22	2.02	16.77	0.99
Shim-APQ11 (%)	3.42	1.29	9.33	0.98	4.44	2.50	21.06	1.19
HNR (dB)	18.03	2.56	22.75	7.42	16.80	3.97	25.80	1.22
Slope (dB)	–22.33	3.77	–12.24	–31.83	–25.02	4.89	–5.72	–36.28
Tilt (dB)	–10.75	0.70	–8.92	–12.76	–9.31	1.27	–4.08	–12.44

the higher deviation at severely breathy voices. Both effects of higher deviations are expected and reflect the current state of the art for dysphonia.

Furthermore, the model was analyzed as not normally distributed based on the results of the Anderson-Darling test ($P < 0.0005$). This result can be explained by the heterogenic group of healthy subjects and subjects with voice disorders who have different voice complaints and degrees of dysphonia. Heterogeneity commonly implies a non-normal distribution. The aim of the current study was to include as many as different voice disorders and a group of healthy subjects, assessing a wide range of various degrees in voice quality. The included subjects of the current study were comparable with the results of epidemiological studies of dysphonia under consideration of the range and frequency of dysphonia severity level.[75]

All results of assumptions for multiple regression analysis qualified all expectations to use multiple regression analysis with the current data.

The stepwise multiple regression analysis revealed that a combination of nine acoustic variables best predicted the perceived breathiness judgment. The equation, based on the unstandardized coefficients of the regression, is as follows: ABI = 4.668 – (0,172*CPPs) – (0.193*Jit) – (1.283*GNEmax-4500 Hz) – (0.396*Hfno-6000 Hz) + (0.01*HNR-D) + (0,017*H1-H2) + (1.473*Shim-dB) – (0.088*Shim) – (68.295*PSD). The results of this equation ranged from –0.38 to 3.04. For practically clinical application, the equation was linearly rescaled on a scale with values that range between 0 and 10. The resulting equation is as follows: ABI = (5.0447730915 – [0.172*CPPs] – [0.193*Jit] – [1.283*GNEmax-4500 Hz] – [0.396*Hfno-6000 Hz] + [0.01*HNR-D] + [0.017*H1-H2] + [1.473*Shim-dB] – [0.088*Shim] – [68.295*PSD])*2.9257400394.

The results of the ABI model clearly showed a positive relationship with the auditory-perceptual judgment of breathiness, and thus the higher an ABI score, the more severe the breathiness severity and *vice versa*. The correlation between the result of ABI and the B_{mean}-scores was $r_s = 0.840$ ($P = 0.000$), revealing high concurrent validity.[68] This proportional relationship between B_{mean} and ABI is illustrated in Figure 4. The coefficient of determination was $r^2_s = 0.706$, indicating that 70.6% of the variance in B_{mean} was accounted for by ABI.

Diagnostic accuracy of ABI

To evaluate the diagnostic accuracy of ABI and its ability to distinguish non-breathy voices from breathy voices, a ROC curve was constructed (see Figure 5). The A_{ROC} was 0.948 and revealed high discriminatory power to distinguish non-breathy voices from breathy voices. The ROC curve was also used to

TABLE 6.
Correlation Coefficients (r_s) and Coefficients of Determination (r_s^2) Between the Auditory-Perceptual Judgment of Breathiness and the 28 Acoustic Measures

Acoustic Measure	r_s	r_s^2
PSD	−.069*	.00
LNPSD	−.069*	.00
H1-H2	.098**	.01
Hfno-2000 Hz	−.153**	.02
Jit	.176**	.03
Slope	−.179**	.03
Hfno-1000 Hz	.198**	.04
Hfno-4000 Hz	.235**	.06
HNR-D	−.254**	.06
Jit-RAP	.261**	.07
Jit-PPQ	.273**	.07
HNR	−.412**	.17
Shim-APQ11	.457**	.21
Shim-dB	.489**	.24
Hfno-6000 Hz	−.494**	.24
Shim	.495**	.25
Shim-APQ5	.536**	.29
Shim-APQ3	.552**	.30
Tilt	.628**	.39
GNEmean-1000 Hz	−.683**	.47
GNEsum-1000 Hz	−.683**	.47
GNEsd-1000 Hz	−.691**	.48
GNEmean3000 Hz	−.705**	.50
GNEsum3000 Hz	−.705**	.50
GNEsd-3000 Hz	−.705**	.50
GNEmax-3500 Hz	−.715**	.51
GNEmax-4500 Hz	−.718**	.52
CPPs	−.768**	.59

* Correlation is significant at the 0.05 level (two-tailed).
** Correlation is significant at the 0.01 level (two-tailed).

identify a cutoff score that achieved the best balance between sensitivity and specificity and would provide optimal discrimination between the presence and the absence of breathiness. The ABI threshold of 3.44 was chosen as the optimal cutoff score. First, a very high specificity of 92.9% was reached to classify correctly almost all subjects with non-breathy voices. Furthermore, a high sensitivity of 82.4% was found, which classified correctly the subjects with breathy voices. Second, the likelihood analysis for this ABI threshold revealed the best balanced outcome for discriminatory power in LR+ and LR− statistics. Thus, sufficient LR+ was reached at 11.63, which complied with the recommendation of LR+ ≥10.[73] This indicates that a positive ABI score (ie, ABI > 3.44) is very likely to belong to a person with a breathy voice. The LR− result reached only 0.19 and was slightly above the recommendation of LR− ≤0.1. Generally, the lower the LR−, the more confident the clinician can be that a person with a below-threshold ABI score (ie, <3.44) has an absence of breathiness. An LR− ≤0.10 indicates that a low ABI score is very likely to have come from a person without a breathy voice.[73]

Cross-cohort validation of ABI

The 50 iterated cross-cohort analyses yielded median correlations of 0.835, 0.838, 0.841, 0.843, and 0.843 for randomized subgroups of 10, 50, 100, 250, and 400 voice samples, respectively. All these results were nearly identical to the original correlation for all 1042 voice samples. Figure 6 represents the distribution of these cross-cohort correlations. The median correlation gradually increased by extending the size of the subgroups, and the variation in the correlations decreased considerably (eg, sub10 with a range of $r_s = 0.576$, sub50 with a range of $r_s = 0.246$, sub100 with a range of $r_s = 0.123$, sub250 with a range of $r_s = 0.095$, and sub400 with a range of $r_s = 0.063$). A

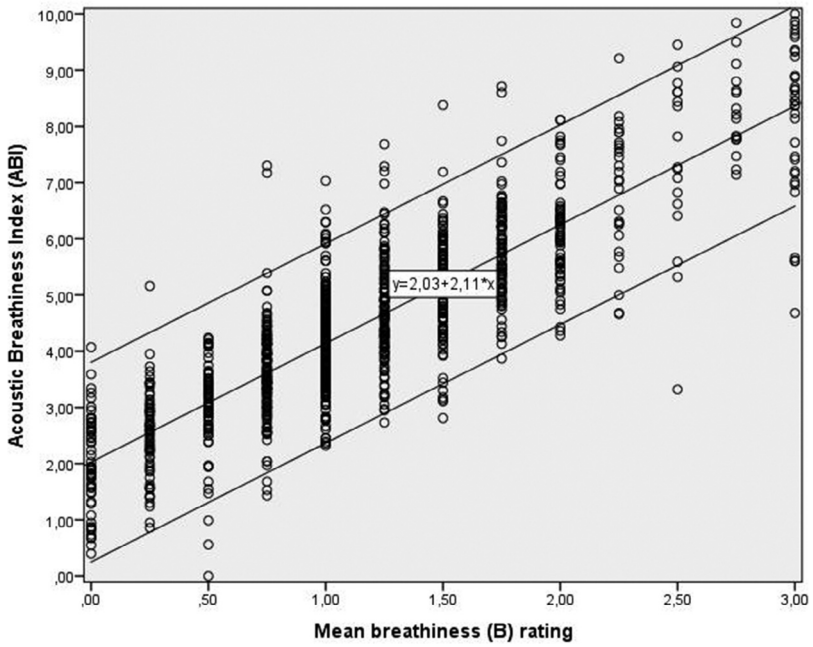

FIGURE 4. Scatterplot and linear regression line illustrating the proportional relationship between Acoustic Breathiness Index and B_{mean} (the two lines above and under the regression fit line delineate the upper and the lower boundaries, respectively, of the 95% prediction interval).

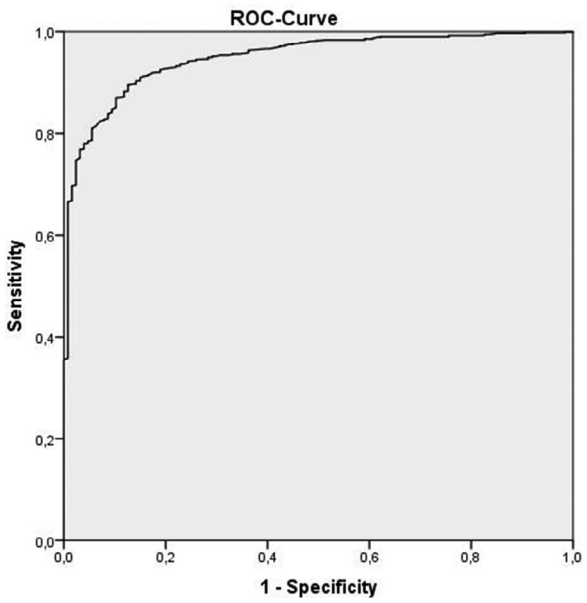

FIGURE 5. ROC curve illustrating the diagnostic accuracy of ABI.

wide variation plus outliers was apparent only in a very small subgroup of 10 selected voice samples (ie, ranged from $r_s = 0.375$ to $r_s = 0.951$). However, the concurrent validity was mainly concentrated at $r_s \geq 0.80$ in all subgroups. These results suggest the stability of ABI across subsets of voice recordings.

DISCUSSION

The aim of this research was to develop an acoustically based quantification of breathiness in voice with research methods analogous to those applied for the development of AVQI.[15,39,41] Therefore, concatenated voice samples of continuous speech and sustained vowel [a:] segments were used. Concatenated voice samples showed perceptually and acoustically high ecological validity in the evaluation of voice quality. In several investigations, the stepwise multiple regression analysis showed the highest concurrent validity in comparison with single acoustic measures by creating a multivariate acoustical model to identify the most robust acoustic predictors in overall voice quality.[15,22,23,30,39] Therefore, this statistical analysis model was used to find the best acoustic predictors in a multivariate model for breathiness as well.

The present study attempted to explore a new acoustic multivariate model for breathiness based on the following. First, a selection of acoustic predictors in breathiness was used.[3] Second, a large number of normophonic and dysphonic voice samples were included. Finally, a strict selection of the rater panel based on the knowledge of several affecting factors[6] disturbing the perceived judgment plus more critical statistical selection criteria in rater reliability were implemented. In this attempt, 28 acoustic measures were used. For clinical and practical reasons, *Praat* freeware was chosen. The most used parameters were spectral or cepstral markers (ie, Hfno, HNR-D, H1-H2, HNR, slope, tilt, and CPPs). Furthermore, frequency perturbation (ie, PSD, LNPSD, and jitter), amplitude perturbation measures (ie, shimmer), and GNE, which combines spectral and perturbation features, were used.

Absolute correlation coefficients between these single acoustic variables and B_{mean}-scores revealed marked correlations for the CPPs and the majority of GNE measures. The findings that these two parameters are the most powerful predictors of breathiness have previously been reported in literature.[3] They are especially designed for the determination of breathiness.[18–20,28]

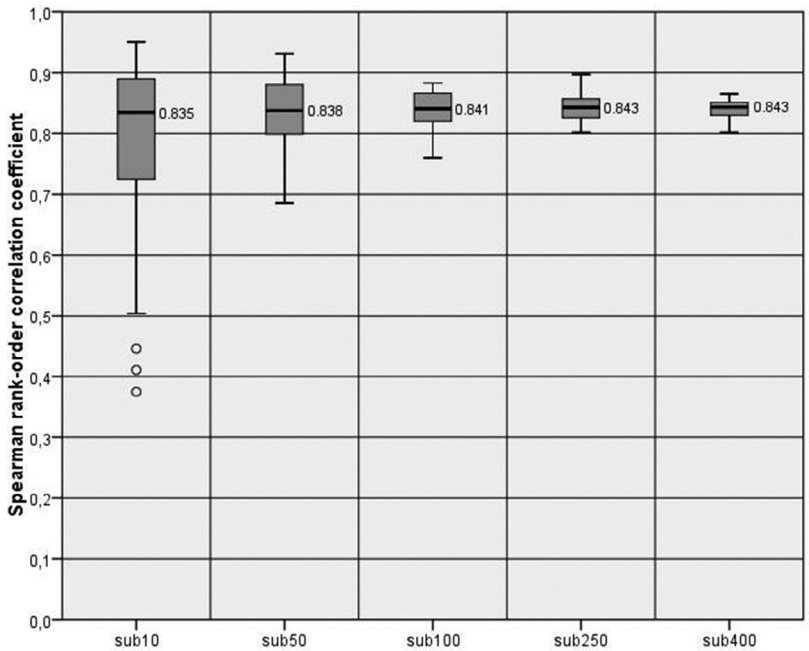

FIGURE 6. Box-and-whisker plots and median illustrating the cross correlations between B and ABI for 50 subgroups of 10, 50, 100, 250, and 400 randomly chosen voice samples of the dataset with 1042 voice samples.

The present development of a weighted algorithm of multiparametric analysis based on stepwise multiple regression created a model for breathiness (ie, ABI) consisting of nine acoustic measures. The two most predictive constituents in this model were CPPs and GNEmax-4500 Hz. These two parameters were also the best predictors as single acoustic measures for the evaluation of breathiness. Although only GNE and CPPs parameters revealed marked absolute correlations between auditory-perceptual judgments (ie, $r_s > 0.70$) in the present study, the ABI model included only one GNE parameter. However, Hfno-6000 Hz, HNR-D, and H1-H2 were also implemented in ABI, which have been found to predict breathiness as well.[3] Although LNPSD revealed the status of an acoustic predictor for breathiness,[3] the non-logarithmic form of PSD was implemented in ABI. Both LNPSD and PSD reached the lowest bivariate correlations with the auditory-perceptual judgment (ie, both $r_s = -0.069$), but PSD was a significant contributing factor in the ABI model. The last three measures that were part of ABI were standard perturbation measures in which two amplitude perturbation measures (ie, Shim and Shim-dB) and one frequency perturbation measure (ie, Jit) were incorporated. Although Shim and Shim-dB showed moderate bivariate correlations in the present study (ie, $r_s = 0.495$ and $r_s = 0.489$, respectively), it is not surprising that shimmer significantly determines the ABI model. First, some shimmer variables revealed high outcomes as one of best predictors in a recent meta-analysis in the evaluation of breathiness.[3] Second, both Shim and Shim-dB were also essential constituents in the AVQI model for the assignment of overall voice quality.

Jit, however, was unexpectedly included in ABI. Only a pitch perturbation quotient variate emerged as one of the best predictors of breathiness in the meta-analysis.[3] Furthermore, the bivariate correlation between Jit and auditory-perceptual judgment was low in this study ($r_s = 0.176$). Nevertheless, the Jit was an important component in the ABI model and has to be included.

To run ABI in *Praat* (see Figure 7), a customized *Praat* script is provided in the supplementary data for clinical and research purposes. In a few seconds, the script automatically analyzed the recorded segments of continuous speech and sustained vowel. At the end of the analysis, one quantified score for the whole voice sample is reported to objectify the presence, the degree, and the progression of breathiness in a sufficiently valid and reliable way.

With an initial value of $r_s = 0.840$ between B_{mean} and ABI, this acoustic model can be considered to relate acceptably and proportionally with perceived breathiness. This outcome outperforms other acoustic models for breathiness reported by Hammarberg et al,[22] Wolfe and Steinfatt,[24] Eskenazi et al,[25] Kreiman et al,[26] Wolfe et al,[27] Bhuta et al,[30] and Eadie and Baylor.[32] Otherwise, comparable or better results in outcome are reported by Hammarberg et al[23] and Stráník et al.[34] These other models, however, analyzed only sustained vowels or continuous speech, whereas ABI was particularly developed to deal with both speech types for voice analysis. Furthermore, the present study included samples of more than 1000 subjects encompassing all degrees and types of dysphonia, which to our knowledge is the largest and therefore most representative multivariate study of acoustic measurement of breathiness.

Next to the concurrent validity, ABI's diagnostic accuracy was also investigated in this study through ROC statistics and LR. In general, ABI correctly identified the presence of breathiness in voice in 94.8% of the cases (ie, $A_{ROC} = 0.948$), and it confirmed a very high accuracy in the evaluation of breathiness. Furthermore, to distinguish between non-breathy voices and breathy voices, the selected cutoff score at ABI = 3.44 showed the best balance between sensitivity (ie, 82.4%) and specificity (ie, 92.9%). This is supported by the LR statistics, which adjust for base-rate differences. A threshold of ABI = 3.44 thus revealed excellent discriminatory accuracy for subjects who test positive (ie, ABI > 3.44).

Unfortunately, diagnostic accuracy statistics were not available for the other acoustic multivariate constructs for breathiness (eg, Table 1). Therefore, a comparison with ABI and these other multivariate constructs for breathiness was not possible on this item.

Limitations and future direction

There are some limitations regarding the present ABI model that not only restrict the validity and power to distinguish nonbreathy voices and breathy voices, but also provide a direction for future research.

First, there are two assumptions of multiple regression analysis (ie, normality and homoscedasticity), which might have limitations in the current ABI model. The majority of assumptions like linearity, statistical independence, and multicollinearity revealed no conflicts using multiple regression analysis for the ABI model. To our knowledge, in the research domain of voice and speech analysis, there was no statistical investigation to test assumptions for multiple regression analysis by creating a model based on linear regression analysis.[15,22,23,27,76] For future directions, it might be useful to consider these aspects when creating a new model for voice and speech analysis.

Second, although the validity of ABI is considerable ($r_s^2 = 0.706$), a variance in breathiness of 29.4% remains not accounted for by ABI. This effect might be explained through the outliers outside of the 95% confidence interval and relative wide range of the 95% confidence interval illustrated in Figure 4. This status implies that there is more overlap in ABI scores between adjacent levels of perceived breathiness. Less variance in ABI per level would increase its discrimination power. In the case of ABI, the auditory-perceptual judgment of breathiness is one factor that may have decreased its ability to accurately measure perceived breathiness. Low inter- and intra-rater reliability ultimately contributes to increased error variance in the regression analysis, leaving less true variance to be explained or accounted for by the acoustic model. The following interventions were chosen to minimize (the influence of) auditory-perceptual noise. First, only speech-language therapists with long-standing experience in the clinical evaluation of voice quality were asked to rate the samples; they all judged the voice samples in their own office at their own pace. Third, anchor voices were used to equalize their internal breathiness severity standard. Fourth, short breaks were included to deal with attention shift and motivational constraints. Fifth, only ratings from judges who in combination showed least variability were used for further statistical analysis and ABI development. However, in other studies

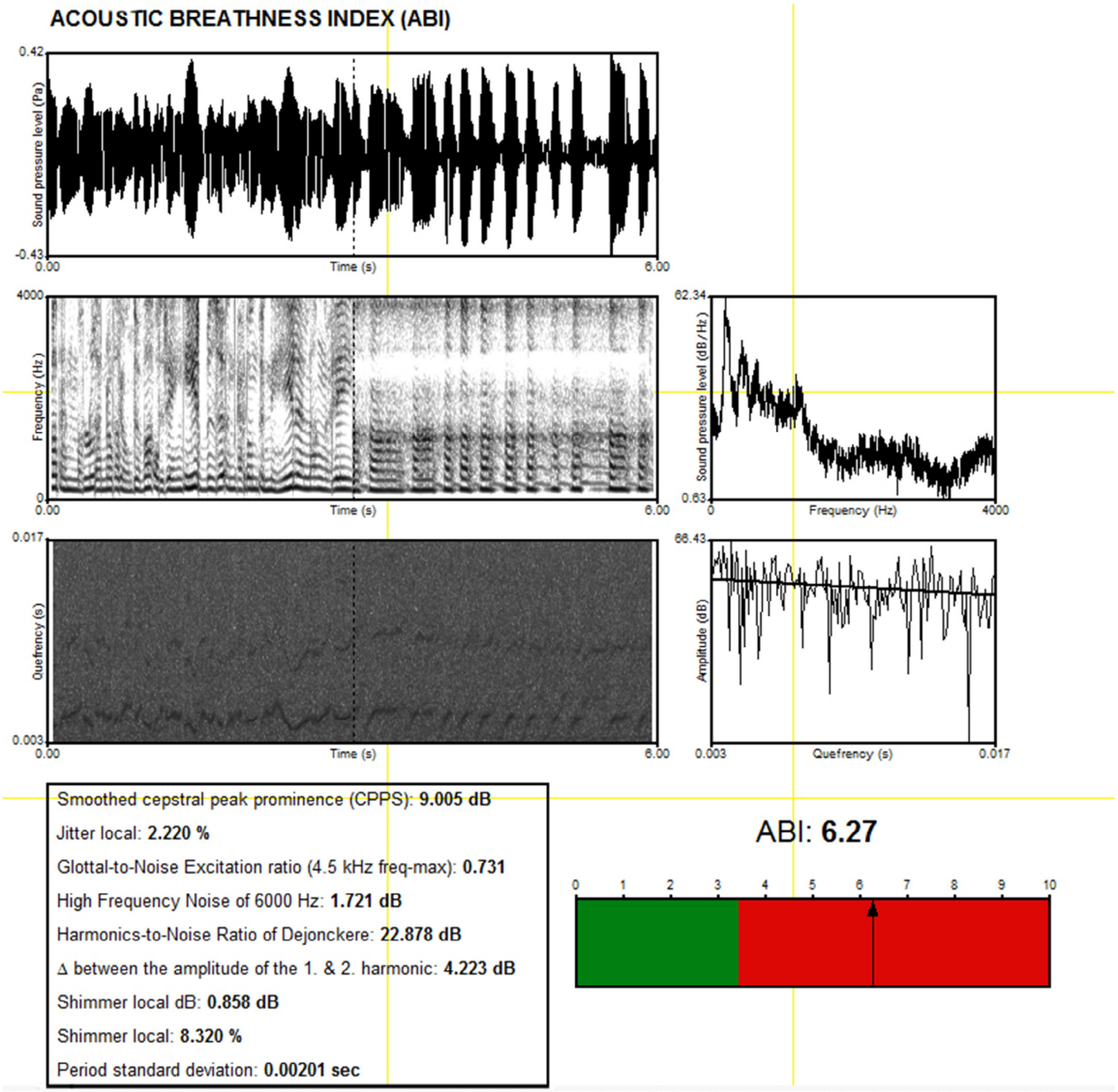

FIGURE 7. Example of the graphical output of the *Praat* script for ABI. Subject 241 was a 41-year-old woman with vocal fold scar. The ABI (6.27) confirms the B_{mean} that is equal to 2.25. Top graph: oscillogram. Center left graph: narrowband-spectrogram with window length = 0.03 seconds, time step = 0.002 seconds, and frequency step = 20 Hz. Center right graph: long-term average spectrum with frequency step = 1 Hz. Bottom left and right graphs, respectively: power cepstrogram and power cepstrum with time step = 0.02 seconds and ranging between 0.00303 seconds and 0.01667 seconds (ie, 330 Hz and 60 Hz, respectively). Finally, the table below illustrates the outcomes of the nine separate acoustic measures in the ABI model. The severity line from 0 to 10 demonstrates the ABI value beside the table. The higher an ABI score, the more breathy is the voice and *vice versa*. The green area corresponds to the absence of breathiness and the red area corresponds to breathiness in voice. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

regarding acoustic breathiness assessment with less voice samples and raters, rater reliabilities were at best comparable with those in the present study.[77–83] In future, to increase intra- and inter-rater reliability, an advanced and intensive training might be meaningful for the chosen judges. Furthermore, the addition of using narrowband spectrograms in the evaluation of breathiness increases the reliability as investigated by Martens et al,[81] Núñez-Batalla et al,[82] and Barsties et al.[83] Criteria of a continuum

breathiness severity in narrowband spectrograms could be used as in Sprecher et al.[74] This scheme showed a high correlation with the perceived breathiness judgment.[84]

Third, the criteria of a freeware single-software solution limited the inclusion of further strong acoustic predictors[3] such as L0 – L1 (ie, differences between the amplitude of the fundamental and formant 1 in the long-term average spectrum), the smoothed pitch perturbation quotient, normalized noise energy of the selection of 1–5 kHz (NNE 1000–5000 Hz), smoothed amplitude perturbation quotient, and CPP.

Fourth, the intra-cohort correlations were investigated on numerous subgroups of the same sample on which ABI was originally modeled. In future, an external validation study should be achieved to test the validity of ABI on reproducibility with alternative subjects and settings outside the initial study.

Fifth, further investigations might be useful to verify the accuracy between ABI and the perceived rating of breathiness in comparison with other multiparametric indices (eg, AVQI) and further perceived ratings of voice quality (eg, overall voice quality, roughness, strain). The aim of our investigation was to evaluate the originality of ABI that corresponds to the perceived rating of breathiness.

Sixth, more research is necessary as to what extent the phenomenon of breathiness (ie, perceptually or acoustically measured) corresponds to physiological aspects of vocal fold vibration like glottal closure. There are investigations that have shown moderate correlations between glottal closure and perceived breathiness as well as with acoustic parameters (eg, jitter, shimmer).[85] However, there is still a lack of research that has assessed the analyses of laryngeal imaging and perceived or acoustic measurements (eg, CPPs, various GNE parameters) at the same time. Therefore, in the current stage, no firm conclusion can be drawn between physiological aspects of vocal fold vibration and perceived or acoustic measurements.

CONCLUSION

ABI showed strong concurrent validity and high diagnostic accuracy for the evaluation of vocal breathiness. Furthermore, high ecological validity was accomplished in this multivariate model considering both continuous speech and sustained vowel segments. These two speech types were meaningful for voice quality analysis,[6,8,15,83] and the equalization of the proportions of these two speech types was considered.[8,39] The included subjects represented a voice clinic population reflecting different ages, genders, different types and degrees of dysphonia as well as nonorganic and organic laryngeal pathologies, and normophonia. The outcome of this study accomplishes an important step toward practical, reliable, and valid objective voice assessments. ABI supports the clinical and scientific assessment of breathiness in both subjects with voice disorder and vocally healthy subjects.

Acknowledgment

The authors thank Jopie Kuiper, Timmy Hartmann, Rudi Verfaillie, Prof. Dr. Marc De Bodt, Paulien Keim, Dr. Leo Meulenbroek, Gerti te Walvaart, Tinka Thede, Gertie Savelkoul, Jessica Fremdgen, Bertine Lefers, and Kim Rutten for their contributions in the perceptual judgment of the many concatenated voice samples.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at doi:10.1016/j.jvoice.2016.11.017.

REFERENCES

1. Verdolini K, Rosen CA, Branski RC. *Classification Manual for Voice Disorders-I. Special Interest Division 3, Voice and Voice Disorders, American Speech—Language-Hearing Association*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2006.
2. Shrivastav R, Sapienza CM. Objective measures of breathy voice quality obtained using an auditory model. *J Acoust Soc Am*. 2003;114:2217–2224.
3. Barsties B, Maryn Y, Gerrits E, et al. A meta-analysis: acoustic measurement of roughness and breathiness. 2016. In Review.
4. Hirano M. Psycho-acoustic evaluation of voice. In: Arnold GE, Winckel F, Wyke BD, eds. *Disorders of Human Communication 5. Clinical Examination of Voice*. Vienna, Austria: Springer-Verlag; 1981:81–84.
5. Kempster GB, Gerratt BR, Verdolini Abbott K, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
6. Barsties B, De Bodt M. Assessment of voice quality: current-state of the art. *Auris Nasus Larynx*. 2015;42:183–188.
7. Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatr Logop*. 2009;61:49–56.
8. Barsties B, Maryn Y. The influence of voice sample length in the auditory-perceptual judgment of overall voice quality. *J Voice*. 2016;In press.
9. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-based clinical voice assessment: a systematic review. *Am J Speech Lang Pathol*. 2013;22:212–226.
10. Vogel AP, Morgan AT. Factors affecting the quality of sound recording for speech and voice analysis. *Int J Speech Lang Pathol*. 2009;11:431–437.
11. Kisenwether JS, Sataloff RT. The effect of microphone type on acoustical measures of synthesized vowels. *J Voice*. 2015;29:548–551.
12. Friedrich G, Dejonckere PH. [The voice evaluation protocol of the European Laryngological Society (ELS)—first results of a multicenter study]. *Laryngorhinootologie*. 2005;84:744–752.
13. Carding PN, Wilson JA, MacKenzie K, et al. Measuring voice outcomes: state of the science review. *J Laryngol Otol*. 2009;123:823–829.
14. Maryn Y, Roy N, De Bodt M, et al. Acoustic measurement of overall voice quality: a meta-analysis. *J Acoust Soc Am*. 2009;126:2619–2634.
15. Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J Voice*. 2010;24:540–555.
16. Awan SN, Roy N, Dromey C. Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model. *Clin Linguist Phon*. 2009;23:825–841.
17. Awan SN, Roy N, Jetté ME, et al. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the CAPE-V. *Clin Linguist Phon*. 2010;24:742–758.
18. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. *J Speech Hear Res*. 1994;37:769–778.
19. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J Speech Hear Res*. 1996;39:311–321.
20. Michaelis D, Gramss T, Strube HW. Glottal-to-noise excitation ratio—a new measure for describing pathological voices. *Acustica/Acta Acustica*. 1997;83:700–706.
21. Dejonckere PH. Recognition of hoarseness by means of LTAS. *Int J Rehabil Res*. 1983;7:73–74.
22. Hammarberg B, Fritzell B, Gauffin J, et al. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngol*. 1980;90:441–451.

23. Hammarberg B, Fritzell B, Schiratzki H. Teflon injection in 16 patients with paralytic dysphonia: perceptual and acoustic evaluations. *J Speech Hear Disord*. 1984;49:72–82.
24. Wolfe VI, Steinfatt TM. Prediction of vocal severity within and across voice types. *J Speech Hear Res*. 1987;30:230–240.
25. Eskenazi L, Childers DG, Hicks DM. Acoustic correlates of vocal quality. *J Speech Hear Res*. 1990;33:298–306.
26. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33:103–115.
27. Wolfe V, Fitch J, Martin D. Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatr Logop*. 1997;49:292–299.
28. Fröhlich M, Michaelis D, Strube HW, et al. Acoustic voice analysis by means of the hoarseness diagram. *J Speech Lang Hear Res*. 2000;43:706–720.
29. Schönweiler R, Hess M, Wübbelt P, et al. Novel approach to acoustical voice analysis using artificial neural networks. *J Assoc Res Otolaryngol*. 2000;1:270–282.
30. Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurement. *J Voice*. 2004;18:299–304.
31. Awan SN, Roy N. Acoustic prediction of voice type in women with functional dysphonia. *J Voice*. 2005;19:268–282.
32. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20:527–544.
33. Madazio G, Leão S, Behlau M. The phonatory deviation diagram: a novel objective measurement of vocal function. *Folia Phoniatr Logop*. 2011;63:305–311.
34. Stráník A, Čmejla R, Vokřál J. Acoustic parameters for classification of breathiness in continuous speech according to the GRBAS scale. *J Voice*. 2014;28:653, e9-653.e17.
35. Maryn Y, De Bodt M, Roy N. The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders. *J Commun Disord*. 2010;43:161–174.
36. Barsties B, Maryn Y. [The Acoustic Voice Quality Index. Toward expanded measurement of dysphonia severity in German subjects]. *HNO*. 2012;60:715–720.
37. Reynolds V, Buckland A, Bailey J, et al. Objective assessment of pediatric voice disorders with the acoustic voice quality index. *J Voice*. 2012;26:672, e1-7.
38. Maryn Y, De Bodt M, Barsties B, et al. The value of the Acoustic Voice Quality Index as a measure of dysphonia severity in subjects speaking different languages. *Eur Arch Otorhinolaryngol*. 2014;271:1609–1619.
39. Barsties B, Maryn Y. The improvement of internal consistency of the Acoustic Voice Quality Index. *Am J Otolaryngol*. 2015;36:647–656.
40. Kankare E, Barsties B, Maryn Y, et al. A preliminary study of the Acoustic Voice Quality Index in Finnish speaking population. 11th Pan European Voice Conference; 2015 August 31–September 4, Florence, Italy.
41. Barsties B, Maryn Y. External validation of the Acoustic Voice Quality Index version 03.01 with extended representativity. *Ann Otol Rhinol Laryngol*. 2016;125:571–583.
42. Maryn Y, Kim HT, Kim J. Auditory-perceptual and acoustic methods in measuring dysphonia severity of Korean speech. *J Voice*. 2016;30:587–594.
43. Hosokawa K, Barsties B, Iwahashi T, et al. Validation of the acoustic voice quality index in the Japanese language. *J Voice*. 2016. doi:10.1016/j.jvoice.2016.05.010, S0892-1997(16)30078-9 [pii].
44. Uloza V, Petrauskas T, Padervinskis E, et al. Validation of the Acoustic Voice Quality Index in the Lithuanian language. *J Voice*. 2016. doi:10.1016/j.jvoice.2016.06.002, S0892-1997(16)30071-6 [pii].
45. Kay Elemetrics Corp. *Voice Range Profile (VRP) Model 4326: Software Instruction Manual*. Lincoln Park, NJ: Kay Elemetrics; 2003.
46. Barsties B. [Effects of different tasks on determination of the speaking fundamental frequency]. *HNO*. 2013;61:609–616.
47. Kay Elemetrics Corp. *Multi-Dimensional Voice Program (MDVP) Model 5105: Software Instruction Manual*. Lincoln Park, NJ: Kay Elemetrics; 2003.
48. Kay Elemetrics Corp. *Real-Time EGG Analysis Model 5138: Software Instruction Manual*. Lincoln Park, NJ: Kay Elemetrics; 2003.
49. Neiman GS, Edeson B. Procedural aspects of eliciting maximum phonation time. *Folia Phoniatr Logop*. 1981;33:285–293.
50. Speyer R, Bogaardt HC, Passos VL, et al. Maximum phonation time: variability and reliability. *J Voice*. 2010;24:281–284.
51. Rau D, Beckett RL. Aerodynamic assessment of vocal function using hand-held spirometers. *J Speech Hear Disord*. 1984;49:183–188.
52. Hirano M, Koike Y, Von Leden H. Maximum phonation time and air usage during phonation. *Folia Phoniatr Logop*. 1968;20:185–201.
53. Wuyts FL, De Bodt MS, Molenberghs G, et al. The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. *J Speech Hear Res*. 2000;43:796–809.
54. Jacobson B, Johnson A, Grywalski C, et al. The Voice Handicap Index (VHI): development and validation. *Am J Speech Lang Pathol*. 1997;6:66–70.
55. De Bodt M, Jacobson B, Musschoot S, et al. De Voice Handicap Index: Een instrument voor het kwantificeren van de psychosociale consequenties van stemstoornissen. *Logopedie*. 2000;13:29–33.
56. Deliyski DD, Shaw HS, Evans MK. Adverse effects of environmental noise on acoustic voice quality measurements. *J Voice*. 2005;19:15–28.
57. Deliyski DD, Shaw HS, Evans MK, et al. Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatr Logop*. 2006;58:274–288.
58. Van de Weijer JC, Slis IH. Nasaliteitsmeting met de nasometer. *Tijdschrift voor Logopedie en Foniatrie*. 1991;63:97–101.
59. Van Lierde K. Nasalance and nasality in clinical practice. Unpublished doctoral dissertation. Ghent, Belgium: University of Ghent; 2001.
60. Wuyts FL, De Bodt MS, Van de Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999;13:508–517.
61. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.
62. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45:111–126.
63. Dejonckere PH, Lebacq J. Harmonic emergence in formant zone of a sustained [a] as a parameter for evaluating hoarseness. *Acta Otorhinolaryngol Belg*. 1987;41:988–996.
64. Everitt BS. *The Cambridge Dictionary of Statistics*. 2nd ed. New York, NY: Cambridge University Press; 2002:202.
65. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378–382.
66. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
67. Van Belle S. Agreement between raters and groups of raters. Unpublished doctoral dissertation, University of Liège, Department of Mathematics, Liège, Belgium; 2009.
68. Frey LR, Botan CH, Friedman PG, et al. *Investigating Communication: An Introduction to Research Methods*. Englewood Cliffs, NJ: Prentice Hall; 1991.
69. Nau R. Statistical forecasting: notes on regression and time series analysis. 2016. [website for statistical computation]. Available at: http://people.duke.edu/~rnau/411home.htm. Accessed October 25, 2016.
70. Field A. *Discovering Statistics Using SPSS*. Los Angeles, CA: Sage; 2009.
71. Otto K. Anderson-Darling normality test calculator. 2005. [website for statistical computation]. Available at: http://www.kevinotto.com/RSS/templates/Anderson-Darling Normality Test Calculator.xls. Accessed October 25, 2016.
72. Portney LG, Watkins MP. *Foundations of Clinical Research, Applications to Practice*. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2000.
73. Dollaghan CA. *The Handbook for Evidence-based Practice in Communication Disorders*. Baltimore, MD: Brookes; 2007.
74. Sprecher A, Olszewski A, Jiang JJ, et al. Updating signal typing in voice: addition of type 4 signals. *J Acoust Soc Am*. 2010;127:3710–3716.
75. De Bodt M, Van den Steen L, Mertens F, et al. Characteristics of a dysphonic population referred for voice assessment and/or voice therapy. *Folia Phoniatr Logop*. 2015;67:178–186.
76. Bettens K, De Bodt M, Maryn Y, et al. The relationship between the Nasality Severity Index 2.0 and perceptual judgments of hypernasality. *J Commun Disord*. 2016;62:67–81.
77. Dejonckere PH, Remacle M, Fresnel-Elbaz E, et al. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with

acoustic measurements. *Rev Laryngol Otol Rhinol (Bord)*. 1996;117:219–224.
78. De Bodt MS, Wuyts FL, Van de Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:74–80.
79. Revis J, Giovanni A, Wuyts F, et al. Comparison of different voice samples for perceptual analysis. *Folia Phoniatr Logop*. 1999;51:108–116.
80. Webb AL, Carding PN, Deary IJ, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol*. 2004;261:429–434.
81. Martens JW, Versnel H, Dejonckere PH. The effect of visible speech in the perceptual rating of pathological voices. *Arch Otolaryngol Head Neck Surg*. 2007;133:178–185.
82. Núñez-Batalla F, Díaz-Molina JP, García-López I, et al. [The effect of anchor voices and visible speech in training in the GRABS scale of perceptual evaluation of dysphonia]. *Acta Otorrinolaringol Esp*. 2012;63:173–179.
83. Barsties B, Beers M, ten Cate L, et al. The effect of visual feedback and training in auditory-perceptual judgment of voice quality. *Logoped Phoniatr Vocol*. 2015;In press.
84. Barsties B, Hoffmann U, Maryn Y. [The evaluation of voice quality via signal typing in voice using narrowband spectrograms]. *Laryngorhinootologie*. 2016;95:105–111.
85. Uloza V, Vegienė A, Saferis V. Correlation between the basic video laryngostroboscopic parameters and multidimensional voice measurements. *J Voice*. 2013;27:744–752.