Toward Improved Ecological Validity in the Acoustic Measurement of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels

*,†,‡Youri Maryn, †,‡Paul Corthals, ‡Paul Van Cauwenberge, §Nelson Roy, and ‡,‖Marc De Bodt, *Bruges, †‡§Ghent, and ¶Antwerp, Belgium and ‖Utah, USA

Summary: To improve ecological validity, perceptual and instrumental assessment of disordered voice, including overall voice quality, should ideally sample both sustained vowels and continuous speech. This investigation assessed the utility of combining both voice contexts for the purpose of auditory-perceptual ratings as well as acoustic measurement of overall voice quality. Sustained vowel and continuous speech samples from 251 subjects with (n = 229) or without (n = 22) various voice disorders were concatenated and perceptually rated on overall voice quality by five experienced voice clinicians. After removing the nonvoiced segments within the continuous speech samples, the concatenated samples were analyzed using 13 acoustic measures based on fundamental frequency perturbation, amplitude perturbation, spectral and cepstral analyses. Stepwise multiple regression analysis yielded a six-variable acoustic model for the multiparametric measurement of overall voice quality of the concatenated samples (with a cepstral measure as the main contributor to the prediction of overall voice quality). The correlation of this model with mean ratings of overall voice quality resulted in $r_s = 0.78$. A cross-validation approach involving the iterated internal cross-correlations with 30 subgroups of 100, 50, and 10 samples confirmed a comparable degree of association. Furthermore, the ability of the model to distinguish voice-disordered from vocally normal participants was assessed using estimates of diagnostic precision including receiver operating characteristic (ROC) curve analysis, sensitivity, and specificity, as well as likelihood ratios (LRs), which adjust for base-rate differences between the groups. Depending on the cutoff criteria employed, the analyses revealed an impressive area under ROC = 0.895 as well as respectable sensitivity, specificity, and LR. The results support the diagnostic utility of combining voice samples from both continuous speech and sustained vowels in acoustic and perceptual analysis of disordered voice. The findings are discussed in relation to the extant literature and the need for further refinement of the acoustic algorithm.

Key Words: Overall voice quality–Multivariate acoustic measurement–Perceptual rating–Cepstral measure–Amplitude perturbation–Spectral measure–Harmonics-to-noise ratio–Sustained vowel–Continuous speech.

INTRODUCTION

Clinical assessment of dysphonia often relies on a combination of perceptual and acoustic measurement techniques. In general, there is the clinician's perceptual evaluation of voice quality, which is considered to be the gold standard upon which other methods are validated. Different kinds of rating scales and various rating systems, such as Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS)[1] or consensus auditory-perceptual evaluation of voice (CAPE-V),[2,3] have been proposed to standardize and quantify this perceptual assessment and to enhance its reliability.[4–6] Because of the subjective nature of perceptual methods, however, there are potentially several internal and external sources of bias involved,[7] including, but not limited to (1) experience of the listener and his/her exposure to voice disorders,[8–12] (2) degree of the patients' dysphonia,[13] (3) type of auditory-perceptual rating scale[9,14–18] and, (4) speaking task or stimulus type.[12,19] Despite these problems, perceptual judgment of voice quality provides a measure that is readily accessible to all voice clinicians,[20] and it, therefore, remains an essential part of voice assessment.[21,22] The aforementioned problems, however, have lead clinicians and researchers to develop various kinds of instrumental methods to "objectively" quantify the degree of overall voice quality disruption. Among these methods, acoustic measurements have become especially attractive due to their noninvasiveness, relatively low cost, and ease of application.[23] Acoustic analyses often provide a numerical output, which potentially captures the degree of dysphonia severity, permits tracking of treatment outcomes, and provides a means to communicate this information relatively easily to all stakeholders, for example, voice clinicians, patients, third-party payers, and physicians.[24] However, perhaps one of the most compelling arguments for the use of acoustic measures is consistency[22] or the fact that, for a given voice sample, the outcome remains unaltered as long as the algorithm behind the measurement remains unchanged. Given these advantages, it is no surprise that there is a vast body of research that addresses acoustic analysis algorithms and methods (see Buder[25] for a comprehensive and complete overview) and investigates the relationship between acoustic measurement and the perceptual evaluation of voice quality (see Maryn et al[26] for a meta-analysis).

Accepted for publication December 31, 2008.

From the *Department of Otorhinolaryngology, Head and Neck Surgery, Speech-Language Pathology and Audiology, Sint-Jan General Hospital, Bruges, Belgium; †Faculty of Health Care Vesalius, University College Ghent, Ghent, Belgium; ‡Department of Otorhinolaryngology & Head and Neck Surgery and Speech-Language Pathology, Faculty of Medicine and Health Sciences, University of Ghent, Ghent, Belgium; §Department of Communication Sciences and Disorders, University of Utah, Salt Lake City, Utah; and the ‖Department of Otorhinolaryngology & Head and Neck Surgery and Communication Disorders, University Hospital, Antwerp, Belgium.

Address correspondence and reprint requests to Youri Maryn, Department of Speech-Language Pathology and Audiology, Sint-Jan General Hospital, Ruddershove 10, 8000 Bruges, Belgium. E-mail: youri.maryn@azbrugge.be

Journal of Voice, Vol. ■, No. ■, pp. 1-16
0892-1997/$36.00

doi:10.1016/j.jvoice.2008.12.014

In most voice clinics, acoustic measures are derived from sustained vowel samples and not from continuous speech samples. Several factors have contributed to this preference.[23,27,28] First, a sustained vowel represents relatively time-invariant phonation, whereas continuous speech involves rapid and frequent changes caused by glottal and supraglottal mechanisms. Second, in contrast to continuous speech, sustained mid-vowel segments do not contain nonvoiced phonemes, fast voice onsets and terminations, and prosodic fundamental frequency and amplitude fluctuations. Third, sustained vowels are not affected by speech rate, vocal pauses, phonetic context, and stress. Fourth, classic fundamental frequency or period perturbation and amplitude perturbation measures strongly rely on pitch detection and extraction algorithms. As a consequence, they lose precision in continuous speech analyses, in which perturbation is significantly affected by intonational patterns, voice onsets and offsets, and unvoiced fragments.[23] Fifth, sustained vowels can be elicited and produced with less effort and in a more standardized manner than that of continuous speech. Sixth, there is no linguistic loading in a sustained vowel, resulting in relative immunity from influences related to dialect and region, language, cognition, and so on.[19]

Although sustained vowel productions are certainly attractive for a variety of reasons, relying exclusively on this voice context does not seem to provide the most ecologically valid voice assessment, one that is truly representative of daily speech and voice use patterns.[12,23] Vocal fluctuations related to voice onset, voice termination, voice breaks, and so on, which are considered to be crucial in voice quality evaluation,[21] can have a relatively large impact on short signals. Furthermore, dysphonia symptoms usually emerge in conversational voice production instead of sustained vowels (with the exception of singing voice), and they are most often signaled by the patients themselves in continuous speech.[29] Additionally, certain voice disorders, such as adductor spasmodic dysphonia, can be characterized by relatively normal voice during sustained vowel productions, whereas voice produced in connected speech is often more severely compromised.[30] Stimulus type (sustained vowel vs continuous speech) is also an important issue in the perceptual evaluation of voice quality and has been investigated by several authors. Although de Krom[31] and Revis et al[32] reported no significant difference between the ratings of a sustained vowel and running speech, Wolfe et al[33] found a significant difference between the ratings of both sample types. The latter finding was supported in part by Zraick et al,[19] who reported a statistically significant difference between the judgments of sustained vowels and recordings of a picture description. Collectively, these findings highlight the need to base clinical voice assessment on more than just sustained vowel analyses, and it seems essential for perceptual and instrumental analyses to be based upon both sample types if it is to be considered ecologically valid.[21,29]

The relationship between acoustic measures and perceptual analysis of voice has received considerable attention in the literature. Researchers have traditionally reported bivariate correlations between specific acoustic measures and auditory-perceptual judgments of overall voice quality. For instance, in their meta-analysis examining the predictive power of specific acoustic correlates, Maryn et al[26] found evidence of moderate to strong correlations with overall voice quality for only a few acoustics markers (out of possible 69 acoustic measures). Aside from smoothed cepstral peak prominence (CPPs)[34] and pitch amplitude,[35] most of the other acoustic measures showed only moderate to very weak correlations with perceptual ratings of overall voice quality. In order to overcome the limited predictive power of single acoustic markers and also motivated by the multidimensionality of voice, several researchers have advocated and explored a multivariate approach for the prediction of voice quality and/or to discriminate among different perceptual categories/levels of dysphonia severity.[22,36–44] Table 1 summarizes relevant methodological items and the most salient outcomes of these multivariate studies. All studies used an equal-appearing interval scale (with a varying number of points, however) to measure the perceptual severity of dysphonia or G. With the exception of Yu et al,[42] the majority suggested a multivariate algorithm consisting of four (acoustic and/or aerodynamic) instrumental measures. The outcomes of these studies were expressed either as a correlation coefficient or in classification accuracy. The classification accuracy of four multivariate models ranged from 49.9% to 86.0%. The association between perception and instrumental measurement was investigated in two other studies, revealing absolute correlation coefficients between 0.58 and 0.88. Both statistics illustrate that the predictive validity of the multivariate approaches can vary from rather low to rather high. We reasoned that improved acoustic prediction of overall voice quality may be derived from combining both sustained vowels and connected speech contexts. There are very few studies in which concatenation of both stimulus types has been used for the clinical examination of overall voice quality and in which correlation coefficients (as a statistic for concurrent validity) as well as conventional measures of diagnostic test performance/precision, such as the ROC analysis, sensitivity, specificity, and LRs, have been presented.

Therefore, this study was undertaken to investigate the feasibility and utility of including both stimulus types in overall voice quality (ie, dysphonia severity) assessment consisting of perceptual and acoustic methods. The voiced segments of two sentences read aloud were concatenated with 3 seconds of the vowel /a/ into a single sound file. In a first experiment, the inter- and intrajudge reliability of perceptual overall voice quality ratings of the concatenated sound files were examined. In a second experiment, the criterion-related concurrent validity of several acoustic markers for the measurement of overall voice quality was studied. The individual correlations of acoustic markers with perceptual ratings were calculated, and the concurrent validity, as well as the internal consistency of a multivariate model based on stepwise linear regression, was investigated. Finally, the diagnostic precision of the model was assessed, using ROC analysis and estimates of sensitivity, specificity, and LRs.

TABLE 1.
Methodology and Outcome of Studies That Used a Multiparametric Approach in the Objective Measurement of Overall Voice Quality

				Perceptual Evaluation of Overall Voice Quality		Outcome	
Source	Number of Subjects	Multivariate Statistical Method	Objective Measures Included in Multivariate Model	Dimension	Scale	Absolute Correlation	Classification Accuracy (%)
Eskenazi et al[36]	16	Multiple linear regression analysis	Pitch amplitude Harmonics-to-noise ratio	Overall severity	EAI 7 points	0.75	/
Wolfe et al[37]	80	Stepwise multiple regression analysis	Relative average perturbation Fundamental frequency	Quality of phonation	EAI* 7points	0.56	/
Giovanni et al[38]	245	Direct-entry discriminant function analysis	Percent jitter Corrected spectrum Ratio of oral airflow to intensity (glottal leakage) Duration of the attack period	G, grade	EAI 5points	/	66.1
Wolfe et al[39]	51	Multiple regression analysis	Noise-to-harmonics ratio Standard deviation of fundamental frequency, percent jitter, relative average perturbation or pitch perturbation quotient	Severity of dysphonia	EAI 7points	0.61	/
Wolfe et al[39]	51	Multiple regression analysis	Noise-to-harmonics ratio Percent shimmer, shimmer in dB or amplitude perturbation quotient	Severity of dysphonia	EAI 7points	0.63	/
Piccirillo et al[40]	33	Logistic regression analysis	Subglottic pressure Airflow at lips Fundamental frequency range Maximum phonation time	G, grade	EAI 4points	0.58	/
Wuyts et al[41]	387	Stepwise logistic regression analysis	Maximum phonation time Highest fundamental frequency Softest intensity Percent jitter	G, grade	EAI 4points	/	49.9
Yu et al[42]	84	Stepwise discriminant function analysis	Fundamental frequency range Fundamental frequency Lyapunov coefficient Maximum phonation time Estimated subglottic pressure Total signal-to-noise ratio	G, grade	EAI 4points	/	86.0
Bhuta et al[43]	37	Stepwise multiple regression analysis	Voice turbulence index Noise-to-harmonics ratio Soft phonation index	G, grade	EAI 4points	0.66	/

Awan & Roy[42]	134	Stepwise multiple regression analysis	Ratio of the amplitude of the cepstral peak prominence to the expected amplitude of the cepstral peak Discrete Fourier transform ratio (energy$_{<4000\ Hz}$/energy$_{>4000\ Hz}$) Logarithm of shimmer Inverse square root of the pitch sigma	Severity of dysphonia	EAI 7points	0.88	/
Ma & Yiu[44]	153	Direct-entry discriminant function analysis	Maximum phonation time Peak intraoral pressure Voice range profile area Relative amplitude perturbation	G, grade	EAI 11points	/	67.3

* EAI, equal-appearing interval.

METHOD

Participants

Voice samples were provided by 22 vocally normal and 229 voice-disordered subjects on an informed consent basis. The voice-disordered subjects were recruited from the ENT caseload of the Sint-Jan General Hospital in Bruges, Belgium. All voice-disordered participants presented with a variety of etiologies and were referred for voice assessment by staff oto-laryngologists. Participants were selected consecutively over the course of a 2-year period. There were 149 females and 79 males, and ages ranged from 8 to 85 years with a mean of 38.9 years (SD = 19.5 y). The scores on the Dysphonia Severity Index[41] ranged from −16.50 to 9.67 with a mean of −0.46. The scores on the Voice Handicap Index[45] ranged from 0 to 106 with a mean of 39.8. Laryngological diagnoses were made with a flexible transnasal chip-on-tip laryngostroboscope (Olympus ENF-V). Table 2 summarizes the variety of voice disorders included in the sample. This group of subjects is considered to be representative of a clinical population of voice-disordered patients. It reflects 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. This study also included 19 females and 3 males without any voice disorder, aged from 19 to 48 years with a mean of 24.6 years. These subjects did not seek help, and since they had no actual voice complaint or history of voice,

TABLE 2.
List of Laryngeal Pathologies, With Their Absolute and Relative Occurrence in the Voice-Disordered Group of This Study

Voice Disorder	Absolute Number	Relative Number
Functional dysphonia	81	35.5
Nodules	42	18.4
Polypoid mucosa (edema)	29	12.7
Paralysis/paresis	18	7.9
Polyp	11	4.8
Cyst	8	3.5
Acute laryngitis	5	2.2
Hemorrhage	4	1.8
Granuloma	4	1.8
Leukoplakia	4	1.8
Mutational falsetto	3	1.3
Tumor	3	1.3
Presbylarynx	3	1.3
Ventricular hypertrophy	2	0.9
Sulcus glottidis	2	0.9
Post-radiotherapy	2	0.9
Web	2	0.9
Post-phonosurgery	1	0.4
Larynx trauma	1	0.4
Interarythenoidal pachyderm	1	0.4
Spasmodic dysphonia	1	0.4
Hyperkeratosis	1	0.4
Total	**228**	**100**

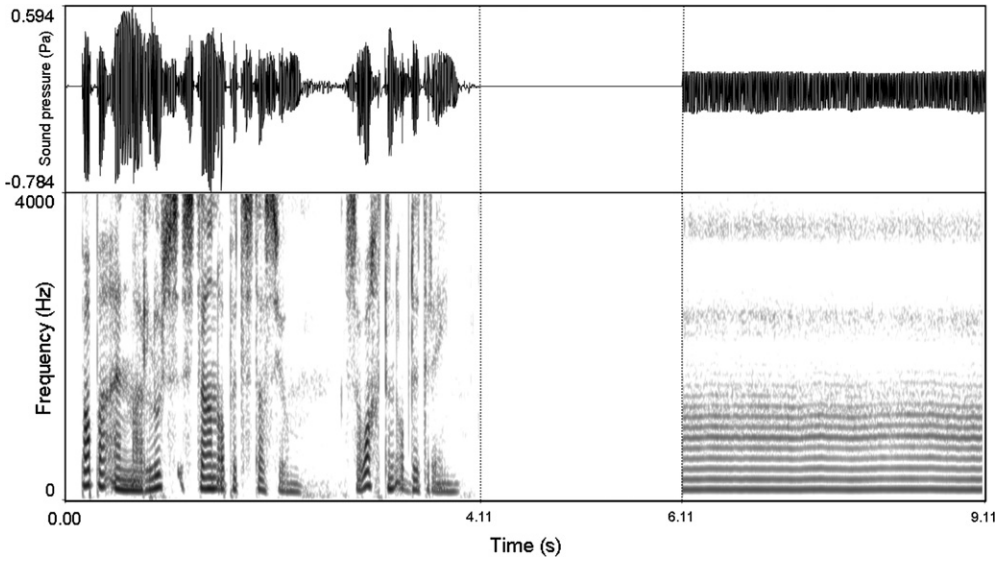

FIGURE 1. Oscillogram and narrowband spectrogram (window length = 0.03 s) of a concatenated voice sample (derived from subject 2), as used in the perceptual evaluations of this study. There are three areas. The left portion reflects the first two sentences of the "Papa en Marloes" text. The right area reflects the middle 3 s of a sustained /a/. Both samples were separated by 2 s of silence (area in the middle).

speech, or hearing problems, the assessment of these vocally normal subjects was limited to the recording of voice samples.

Voice samples

Every participant was asked to sustain the vowel /a/ for at least 5 seconds and to read aloud a phonetically balanced text[46,47] using a comfortable pitch and loudness. Both voice samples were recorded using an AKG C420 head-mounted condenser microphone[48] and digitized at 44100 samples per second,[49] that is, a sampling rate of 44.1 kHz and 16 bits of resolution using the Computerized Speech Lab (CSL model 4500).[50] For the voice-disordered subjects, this was done at the beginning of a standard voice assessment. The samples were saved in .wav format. The vowel samples used in this study were edited to include only the middle 3 seconds. The read text/connected speech samples were trimmed to include only the first two sentences. Finally, the voice samples were concatenated in the following order using *Praat*[51,52]: text segment, a pause of two seconds, followed by the 3 second sustained vowel segment. An example of the resulting concatenated waveform is given in Figure 1.

Overall dysphonia ratings

Five speech-language pathologists (two females and three males, with ages ranging from 27 to 59 y) were asked to rate each of the 251 concatenated voice samples. All listeners had previously participated more than once in post-academic courses on voice disorders, and they all had at least 5 years of clinical experience judging voice quality and overall dysphonia severity. The listening experiment was performed in a quiet setting. The listeners were seated in a circle, equidistantly around two loudspeakers that emitted the voice samples in opposite directions. All concatenated voice samples were presented in random order. All samples were judged within a 5 hour period of time. A 15 minute break was provided after each set of five rating sessions, ie, after 45 minutes. Before the beginning of the listening experiment, all judges had confirmed that one particular concatenated voice sample represented normal voice quality. In order to establish an external standard of normal voice quality, all five rating sessions started with listening to this "normal" voice sample as a referent to compare the 251 voice samples. By doing so, the authors intended to augment the reliability of the auditory-perceptual voice ratings.[53]

The five listeners were instructed to evaluate the severity of the perceptual dimension of overall dysphonia (or Grade, "G"). Before judging the samples, this perceptual dimension was operationally defined following the description of Kreiman and Gerratt.[54] The "G" dimension was rated on a four-point equal-appearing interval scale, as suggested by the Japan Society of Logopedics and Phoniatrics[1]: with a score of 0 representing the absence of a dimension and scores 1, 2, and 3, respectively, corresponding with a slight, moderate, and severe presence of G. Samples were repeated whenever one or more listeners were not confident in their judgment. At the end of the perceptual experiment, 25 randomized voice samples (ie, 10% of all samples) were repeated a second time to determine intrarater reliability.

Acoustic measures

Digital copies of the recordings used for the perceptual evaluations were selected for acoustic measurement. Since the majority of measures in this study pertain to voiced segments, a custom voicing detection algorithm was used to extract the voiced segments from the continuous speech files. The algorithm for detection and extraction of voiced segments was based on the three criteria proposed by Parsa and Jamieson[23(pp332)] and implemented in *Praat*. The programming script is provided in Appendix 1. Frames of 30 milliseconds were designated as voiced if (a) sound energy exceeded 30% of the overall signal energy, (b) zero crossing rate was below 1500 Hz, and

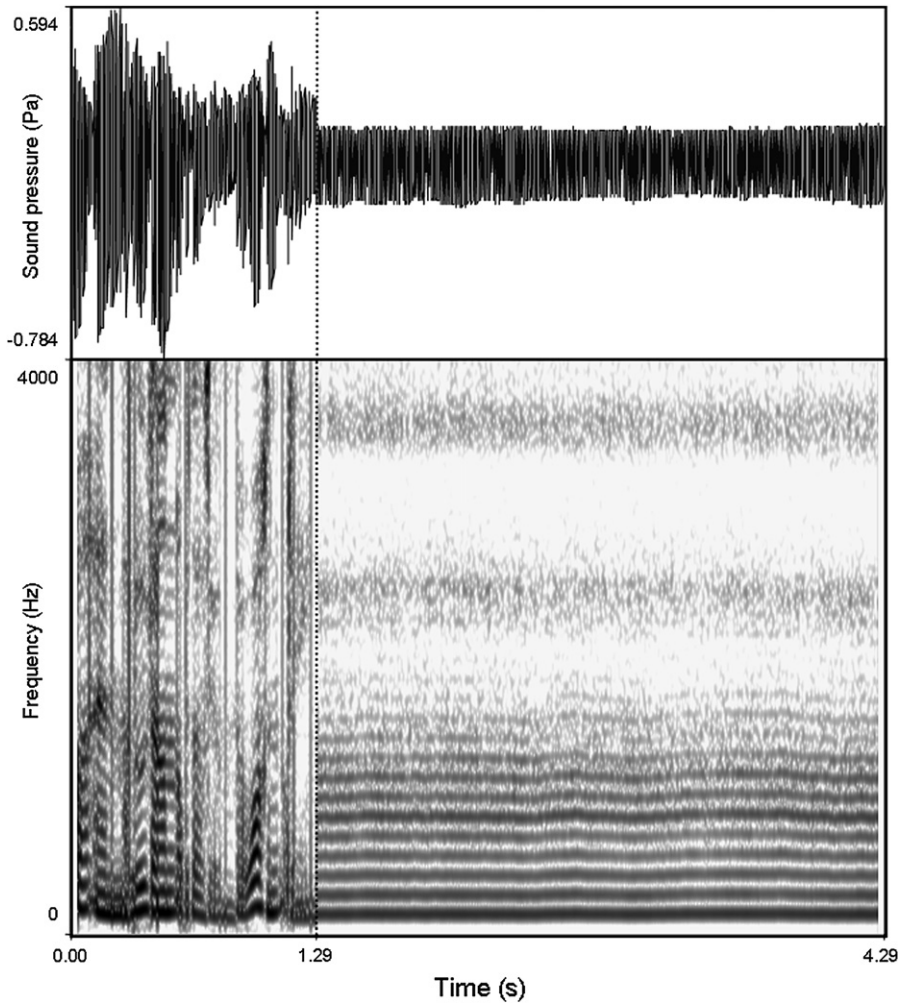

FIGURE 2. Oscillogram and narrowband spectrogram (window length = 0.03 s) of a concatenated voice sample (derived from subject 2), as used for the acoustic measures of this study. There are two areas. The left area reflects the concatenated voiced segments of the first two sentences of the "Papa en Marloes" text. The right area reflects the middle 3 s of a sustained /a/.

(c) the normalized autocorrelation peak was above 0.3. Afterwards, the voiced continuous speech samples were concatenated with the sustained vowel sample of the same patient. An example of the resulting waveform is shown in Figure 2. Thirteen acoustic measures were derived from this material. The following 11 acoustic measures were derived using *Praat*: slope of the long-term average spectrum (Slope), tilt of the trend line through the long-term average spectrum (Tilt), jitter local (a.k.a. percent jitter), jitter rap (a.k.a. relative average perturbation), jitter ppq5 (a.k.a. pitch perturbation quotient), shimmer local (a.k.a. percent shimmer), shimmer local dB (a.k.a. shimmer in dB), shimmer apq11 (a.k.a. amplitude perturbation quotient), mean autocorrelation (mACF), noise-to-harmonics ratio (NHR), and harmonics-to-noise-ratio (HNR). The programming scripts used to obtain these measures in *Praat* are provided in Appendix 2. In addition, the concatenated voice samples were analyzed using the computer program *SpeechTool*,[55] obtained from Hillenbrand et al[56] and Hillenbrand and Houde,[34] which provided two cepstral measures: the cepstral peak prominence (CPP) and CPPs. In short, the measures on the concatenated samples (with only voiced fragments) included two spectral measures (slope and tilt), six perturbation measures (jitter local, jitter rap, jitter ppq5, shimmer local, shimmer local dB, and shimmer apq11), three glottal noise measures (mACF, NHR, and HNR), and two cepstral measures (CPP and CPPs).

Statistical analysis

All statistical analyses were completed using *SPSS* for Windows version 12.0 (SPSS Inc., Chicago, Illinois, USA). In the first experiment, the intrarater and interrater reliability of perceptual evaluation of overall voice quality (G) in concatenations of continuous speech and sustained vowel fragments (Figure 1) was explored. Two coefficients were used to determine listener agreement or reliability. Both statistics are nonparametric, because G-ratings are on an ordinal scale. First, the Cohen kappa coefficient (κ) was calculated. This statistic, yielding values of $\kappa = 1$ for perfect agreement and $\kappa = 0$ when agreement is no better than that by chance, can be defined as a measure of the unanimity in the evaluations by multiple pairs of raters when they are rating the same object.[57] Guidelines for the interpretation of the κ statistic are

provided by De Bodt et al.[10] Second, the Spearman rank-order correlation coefficient (r_s) was determined. This statistic reflects the degree to which a monotonic relationship exists between variables.[58] Interpretation guidelines for r_s are provided by Frey et al.[59]

For the second experiment, the predictive validity of the acoustic measurement of overall dysphonia severity (G) in the concatenated voiced samples was assessed (see Figure 2), and the following statistics were used. First, r_s and the coefficient of determination (r_s^2) between G and the 13 acoustic measures were calculated as measures of concurrent validity. Second, stepwise multiple linear regression was executed to construct a statistical model representing the best combination of acoustic predictors for the overall degree of disordered voice. A multiple regression equation was constructed based on the unstandardized coefficients of the statistical model. In order to simplify clinical interpretation, the model was linearly rescaled in such a way that the outcomes of the equation resulted in a score between 0 and 10. This final model was called Acoustic Voice Quality Index (AVQI). Third, in order to investigate the criterion-related concurrent validity of AVQI, the correlation between G and AVQI was calculated with the Spearman rank-order correlation coefficient. Fourth, in order to examine the diagnostic utility of AVQI, several estimates of diagnostic precision were calculated.[24] For instance, the accuracy of a diagnostic test is commonly evaluated by the sensitivity and specificity of the test. Sensitivity is defined as the proportion of subjects with the disease (ie, cases) who have a positive test, whereas specificity is the proportion of subjects without the disease (ie, noncases) who have a negative test. In tests that yield continuous data such as the AVQI employed in this study, several values of sensitivity and specificity are possible, depending on the cutoff point chosen to define a positive test. This trade-off between sensitivity and specificity can be displayed graphically using a technique known as the ROC curve. To generate an ROC curve, the investigator selects several cutoff points and determines the sensitivity and specificity at each point. Sensitivity (or the true-positive rate) is plotted on the Y-axis as a function of 1−specificity (the false-positive rate) on the X-axis. An optimal diagnostic test is one that reaches the upper left corner of the graph. A test of no value follows the diagonal from the lower left to the upper right corners, suggesting that at any cutoff the true-positive rate is the same as the false-positive rate.

For the ROC-curve of AVQI, a voice was considered to be normal only when all five judges agreed on its normalcy (ie, mean G = 0.0). On the other hand, a voice was considered dysphonic if one judge evaluated it at least as slightly dysphonic or G1 ($0.2 \leq$ mean G ≤ 3). The ability of AVQI to discriminate between normal and dysphonic voices was represented by the "area under ROC," that is, A_{ROC}-statistic. The outcome of A_{ROC} is interpreted as a score between 1.0 (for perfect discrimination between normal and dysphonic voices) and 0.5 (for chance-level diagnostic accuracy).[24] ROC-statistics have been used previously to discriminate vocally normal from voice-disordered subjects in several studies.[23,60–62] In order to facilitate clinical interpretation of AVQI-scores, a threshold-score to distinguish normal from disordered voice quality was derived from the ROC-curve, and positive and negative LRs were also calculated.

LRs provide additional information about the value of a diagnostic test and help diminish problems with sensitivity, specificity related to the uneven number of normophonic and dysphonic subjects in the sample. The LR incorporates both the sensitivity and specificity of the test and provides a direct estimate of how much a test result will change the odds of having a disease. The LR for a positive result (LR^+) yields information regarding how the odds of the disease increase when the test is positive. Specifically, LR^+ is calculated by determining the ratio of true-positive cases (sensitivity) to false-positive cases (1−specificity) [ie, LR^+ = (sensitivity)/1−specificity] and gives information regarding the likelihood that an individual has a voice disorder. When LR^+ yields a number greater than 10, the value of the diagnostic test is high. If the LR^+ yields a value of 3, there is a moderate likelihood that the test suggests the person has the disorder but is not conclusive and, therefore, should be interpreted with caution. If the test yields an LR^+ of 1, the diagnostic test does not help to diagnose a specific disorder. LR^- produces an estimate that helps determine whether an individual does not have a particular disorder when the diagnostic test does not identify them as such. LR^- gives information regarding how much the odds of the disease decrease when a test is negative. It is calculated by determining the ratio of false-negative cases (1−sensitivity) to true-negative cases (specificity) [LR^- = (1−sensitivity)/specificity]. Because the LR statistics consider sensitivity and specificity simultaneously, they are less vulnerable to sample size characteristics and base-rate differences between vocally normal and voice-disordered participants.[63] Both LR^+ and LR^- were calculated for specific AVQI cutoff points (based on the ROC-curve).

Finally, a cross-validation procedure was undertaken. It is well known that when applied to a new set of data, different from the one upon which it was initially modeled, any predictive model may lose accounted variance (r_s^2) and concurrent validity. Therefore, correlation coefficients between G-scores and AVQI-scores were calculated for 30 randomly selected subgroups of 100, 50, and 10 voice samples. This method of cross-validation is similar to a method described by Awan and Roy.[22]

RESULTS

Reliability of auditory-perceptual ratings of concatenated samples

Figure 3 shows the frequency distribution of the mean G-ratings. The results for intrarater reliability, based on 25 of the 251 voice samples, are represented in Table 3. The κ-statistic shows an average of 0.60 and ranges from 0.49 to 0.71. The r_s-statistic indicates a mean of 0.85 and ranges between 0.77 and 0.90. These results confirm moderate to high intrarater reliability. The interrater agreement outcomes are shown in Table 4. The κ-statistic shows an average of 0.39 and ranges from 0.21 to 0.52. The r_s-statistic has a mean of 0.61 and ranges

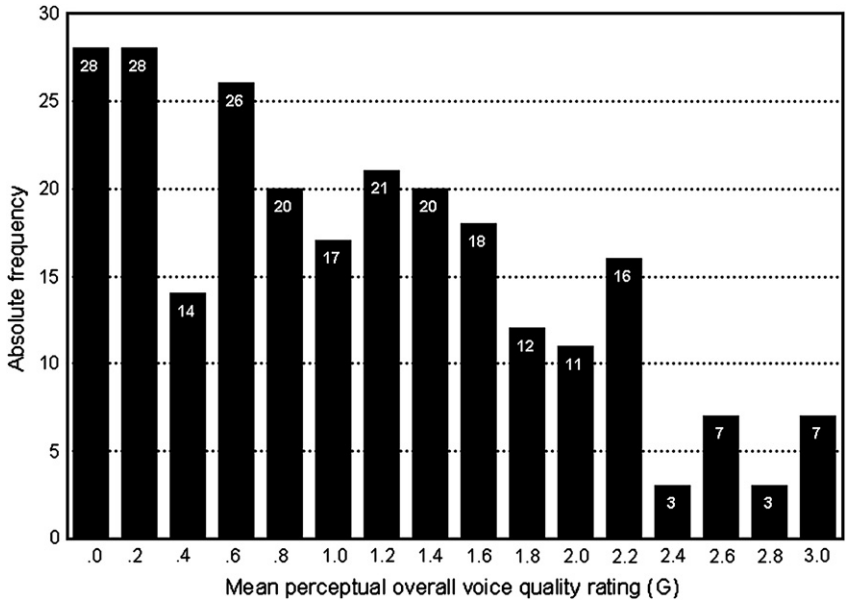

FIGURE 3. Frequency distribution of the mean auditory-perceptual overall voice quality ratings (average of G-scores of 5 experienced listeners) of the 251 concatenated voice samples.

between 0.51 and 0.73. These outcomes indicate fair to moderate interrater agreement.

Predictive validity of acoustic measures on concatenated samples

Table 5 lists the descriptive data for the 13 acoustic variables in the group of 23 vocally normal cases and the 228 dysphonic subjects. The correlations (r_s) and coefficients of determination (r_s^2) between overall voice quality ratings, and these 13 acoustic measures are shown in Table 6. The highest absolute r_s-value was found for CPPs ($r_s = 0.71$); followed by HNR ($r_s = 0.68$), shimmer local dB ($r_s = 0.66$), and CPP ($r_s = 0.65$). The lowest absolute r_s-values were found for the frequency perturbation measures, the glottal noise measures NHR and mACF, and the spectral measures: slope ($r_s = 0.01$), tilt ($r_s = 0.48$), NHR ($r_s = 0.51$), and jitter local ($r_s = 0.54$). The strongest correlation identified is for CPPs ($r_s = -0.71$), explaining approximately 50% of the variation of G. With the exception of slope, for which no significant correlation was found ($r_s = 0.01$), all correlations were significant at the alpha = 0.01 level. The stepwise multiple regression analysis revealed that a combination of six acoustic variables best predicted the overall dysphonia severity of voice recordings containing a concatenation of continuous speech as well as sustained vowels. The equation, based on the unstandardized coefficients of the regression, is as follows: $AVQI = 2.905 - 0.111 \times CPPs - 0.073 \times HNR - 0.213 \times shimmer\ local + 2.789 \times shimmer\ local\ dB - 0.032 \times slope + 0.077 \times tilt$.

The outcomes of this equation range from −0.39 to 3.50. For practical clinical application, however, the equation is linearly rescaled in order to fall on a scale with values between 0 and 10. The resulting equation is: $AVQI = (3.295 - 0.111 \times CPPs - 0.073 \times HNR - 0.213 \times shimmer\ local + 2.789 \times shimmer\ local\ dB - 0.032 \times slope + 0.077 \times tilt) \times 2.571$.

Inspecting the results, it is clear that there is a positive relationship between AVQI and G, and thus, the higher an AVQI score, the more disrupted the overall voice quality and vice versa. The correlation between the outcome of AVQI and the G-scores was 0.78, revealing high concurrent (or predictive) validity. This proportional relationship between G and AVQI

TABLE 3.
Intrarater Reliability of the Five Listeners who Rated Overall Voice Quality on the Concatenated Voice Samples: κ and r_s

Raters	κ	r_s
Rater 1	0.66	0.90
Rater 2	0.63	0.82
Rater 3	0.71	0.90
Rater 4	0.49	0.87
Rater 5	0.49	0.77

TABLE 4.
Matrix of InterRater Reliability Between the Five Listeners who Rated Overall Voice Quality on the Concatenated Voice Samples: κ and r_s

Raters	Statistics	Rater 2	Rater 3	Rater 4	Rater 5
Rater 1	κ	0.49	0.37	0.52	0.21
	r_s	0.64	0.54	0.68	0.51
Rater 2	κ		0.51	0.50	0.31
	r_s		0.63	0.73	0.56
Rater 3	κ			0.42	0.37
	r_s			0.62	0.61
Rater 4	K				0.23
	r_s				0.60

TABLE 5.
Average (M), Standard Deviation (SD), and Range (Min – Max) of the Outcomes of the Thirteen Acoustic Measures

Measures	Normal (N = 23)				Dysphonic (N = 228)			
	M	SD	Min	Max	M	SD	Min	Max
Slope (dB)	−23.31	5.03	−33.15	−14.17	−24.86	4.99	−37.84	−8.15
Tilt (dB)	−10.51	0.73	−12.14	−8.62	−9.45	1.38	−13.81	−5.00
Jitter local (%)	0.98	0.18	0.69	1.31	1.60	1.08	0.71	7.50
Jitter rap (%)	0.46	0.10	0.29	0.63	0.80	0.60	0.30	3.99
Jitter ppq5 (%)	0.50	0.10	0.34	0.71	0.84	0.64	0.34	5.07
Shimmer local (%)	3.18	0.91	1.53	5.09	5.47	3.66	1.51	22.05
Shimmer local dB (dB)	0.31	0.05	0.21	0.42	0.52	0.31	0.21	1.89
Shimmer apq11 (%)	2.31	0.74	1.11	4.19	3.87	2.77	1.22	19.59
mACF	0.97	0.00	0.97	0.98	0.95	0.06	0.55	0.98
NHR	0.04	0.01	0.02	0.06	0.08	0.11	0.03	0.88
HNR (dB)	22.92	2.09	18.89	26.41	18.66	4.71	0.99	28.92
CPP	16.77	2.08	13.57	21.78	13.80	2.45	8.65	21.74
CPPs (dB)	8.05	0.94	5.97	10.16	6.41	1.81	0.89	10.94

is illustrated in Figure 4. The coefficient of determination was 0.61. Figure 4 also shows that the AVQI scores for subjects with severe dysphonia are higher than expected and consequently raises the possibility of a nonlinear polynomial trend between AVQI and G. However, closer investigation of second- and third-order polynomial relationships for the present data revealed no statistically significant difference between the linear and the nonlinear models.

Cross-validation of AVQI

The 30 iterated cross-validations yielded mean correlations of 0.77, 0.75, and 0.80 for randomized subgroups of 100, 50, and 10 voice samples, respectively. These results are almost identical to the original correlation for all 251 voice samples. Figure 5 represents the distribution of these cross-validation correlations. For example, the 30 correlations for 100 randomly chosen voice samples show that the validity of AVQI can range from 0.670 to 0.857 (mean = 0.769; standard error = 0.009; standard deviation = 0.047). The correlations for 50 randomly chosen voice samples lie between 0.633 and 0.852 (mean = 0.751; standard error = 0.011; standard deviation = 0.058) and between 0.462 and 0.963 for 30 times 10 randomly selected voice samples (mean = 0.805; standard error = 0.021; standard deviation = 0.118). These results confirm the stability of the AVQI across subsets of voices.

Diagnostic accuracy of AVQI

To evaluate the diagnostic accuracy of AVQI and its ability to distinguish vocally normal from voice-disordered participants, an ROC-curve was constructed (Figure 6). The A_{ROC}, with the AVQI-scores as the test variable and the G-scores as the state variable, was 0.895, revealing relatively high discriminatory power to distinguish normal and pathological voices (with statistical significance at $P = 0.000$, under the assumption of an asymptotic distribution). The ROC-curve was also used to identify which cutoff point achieved the best balance between sensitivity and specificity and would provide optimal discrimination between the normal and pathological groups. In this regard, an AVQI cutoff score of 2.36 produced sensitivity and specificity estimates of 91% and 59%, respectively. Therefore, using this threshold, 91% of patients with dysphonia were correctly classified as being dysphonic (ie, pathological). However, only 59% of the normal subjects were correctly categorized as non–voice-disordered (ie, as having normal voice quality). This AVQI-score is accompanied by intermediate-range LRs: $LR^+ = 2.23$ and $LR^- = 0.15$. In contrast, using an AVQI cutoff criterion of 2.95 produced estimates of sensitivity of 74% and specificity of 96%. Only 74% of the dysphonic patients were classified correctly, but almost all subjects with normal voice quality were correctly classified as such. Likelihood analysis for this AVQI cutoff score resulted in much improved discriminatory power: $LR^+ = 19.98$ and $LR^- = 0.27$. To assist

TABLE 6.
Correlation Coefficients (r_s) and Coefficients of Determination (r_s^2) Between the Auditory-Perceptual Overall Voice Quality Ratings (G) and the 13 Acoustic Measures

Statistics	Slope	Tilt	Jitter Local	Jitter Rap	Jitter ppq5	Shimmer Local	Shimmer Local dB
r_s	0.01	0.48	0.54	0.56	0.55	0.64	0.66
r_s^2	0.00	0.23	0.29	0.31	0.31	0.40	0.44
	Shimmer apq11	mACF	NHR	HNR	CPP	CPPs	
r_s	0.61	−0.56	0.51	0.68	−0.65	−0.71	
r_s^2	0.37	0.31	0.26	0.46	0.43	0.50	

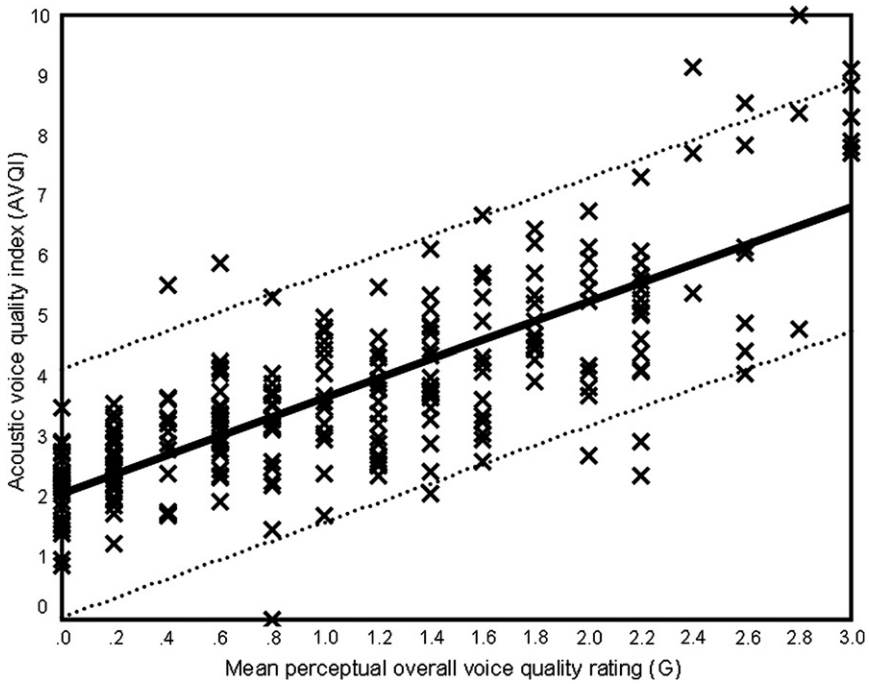

FIGURE 4. Scatter plot to illustrate the concurrent validity of AVQI (the two dotted lines above and under the regression fit line delineate the upper and lower boundaries, respectively, of the 95% prediction interval).

in the interpretation of LRs in this specific study, the higher the LR^+, the more confident the clinician can be that a person with a higher AVQI-score is voice-disordered/dysphonic. An $LR^+ \geq 10$ indicates that a positive AVQI-score (ie, >2.95) is very likely to have come from a dysphonic person. The lower the LR^-, the more confident the clinician can be that a person with a low AVQI-score (ie, <2.95) is normophonic. An $LR^- \leq 0.10$ indicates that a low AVQI-score is very likely to have come from a person without dysphonia.[63]

DISCUSSION

We investigated the utility of combining sustained vowels and continuous speech in dysphonia severity measurement. Although sustained vowels have long been preferred, both stimulus types are important in perceptual and acoustic measurement, and both contexts would seem necessary to improve ecological validity in voice assessment.[12,19,21,23,27–29] For this reason, the first two sentences of a commonly used Dutch text were concatenated with three seconds of a sustained /a/ vowel. The 251 concatenated samples were perceptually rated on a four-point equal-appearing interval scale of overall voice quality (ie, "G" from GRBAS). An acoustic analysis protocol, which contained a diverse set of acoustic variables, consisting of 13 frequency perturbation, amplitude perturbation, as well as spectral and cepstral measures, was applied. Absolute

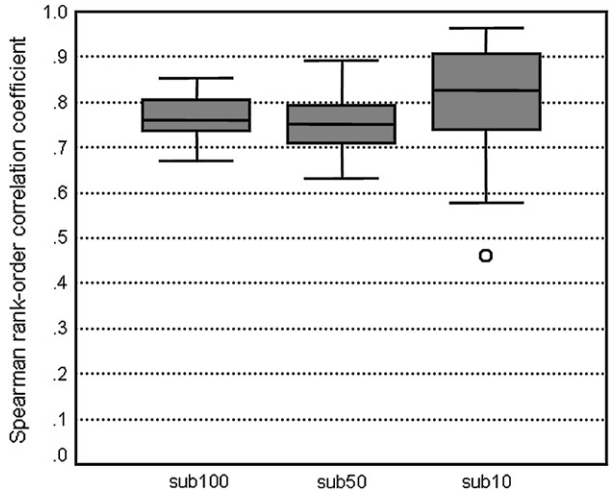

FIGURE 5. Box-and-whisker plots illustrating the cross-correlations between G and AVQI for 30 subgroups of 100, 50, and 10 randomly chosen voice samples.

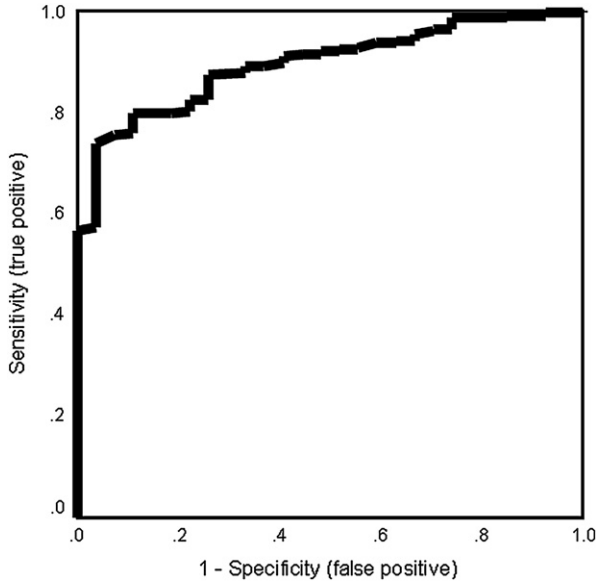

FIGURE 6. ROC curve to illustrate the diagnostic validity of AVQI.

correlation coefficients between these acoustic variables and G-scores were the highest for the cepstral, HNR, and amplitude perturbation measures.

The finding that cepstral measures were the most powerful predictor of G is compatible with existing reports in the literature. For instance, Heman-Ackah et al[64] completed a similar study but on separated samples of sustained vowels and continuous speech, and they reported that for sustained vowels, CPP measures were the best predictor of G ($r_s = -0.80$) as compared with perturbation and glottal noise measures. The outcome was even more impressive for CPPs applied to continuous speech ($r_s = -0.86$). However, these study results were based on a rather small group of 18 subjects. Similarly, Eadie and Baylor[12] also reported a strong association ($r_s = 0.806$) between CPPs and overall severity of dysphonia on sustained vowels, but again this finding was derived from a small number of subjects. In a study by Awan and Roy,[22] based on 134 subjects (Table 1), a cepstral measure called CPP/EXP (similar to the cepstral measures used in this study) was the most powerful contributor to their acoustic model for voice quality prediction. Finally, the cepstral peak magnitude, as reported in the study of Dejonckere and Wieneke,[65] was also the best predictor of hoarseness, compared to spectral and perturbation measures. In conclusion, the results from this investigation also confirm the strength of cepstral-based measures in predicting dysphonia severity and demonstrate the inferiority of specific perturbation measures such as jitter.[22,23,60,66]

In addition to bivariate analyses, a multiparametric analysis approach was employed to construct a weighted algorithm that would identify the most robust acoustic predictors of judgments of "G" or overall dysphonia severity.[22,36–44] Stepwise multiple regression analysis resulted in a model (ie, AVQI) consisting of six acoustic measures. With an initial r_s-value of 0.78 between G and AVQI, that is, a high degree of concurrent validity,[59] this model can be considered a strong predictor of overall voice quality. The AVQI model appears to perform better than other acoustic models reported by Eskenazi et al,[36] Wolfe et al,[37,39] Piccirillo et al,[40] and Bhuta et al[43] but not as favorably as the results reported by Awan and Roy.[22] In contrast to the other models that analyzed sustained vowels only, the performance of the AVQI is particularly compelling, given that AVQI incorporates continuous speech as well as sustained vowels.

Although the performance of the AVQI is very respectable ($r_s^2 = 0.61$), there still remains 39% of variance in G not accounted for by AVQI. This finding is confirmed by the rather wide 95% confidence interval illustrated in Figure 4. A narrower confidence interval would mean that there is less overlap in AVQI scores between adjacent perceptual levels of dysphonia severity, and the AVQI would better discriminate among or between these levels of severity. In this regard, one factor that likely attenuates the ability of *any* acoustic model to account for true variance in listener ratings of dysphonia severity is the "unreliability" of those perceptual judgments. Inter- and intrajudge *unreliability* ultimately contributes to increased error variance in the regression analysis, leaving less true variance to be explained or accounted for by the acoustic model. In this study, our interrater reliability was moderately low, perhaps reflecting differences in training and experience of the listeners. In light of the increased error variance related to only moderate levels of listener reliability, the amount of variance accounted for by the acoustic model is actually quite respectable. Perhaps more intensive training with external perceptual standards, as promoted by Chan and Yiu[53] and Kreiman and Gerratt,[67] could potentially improve the reliability of the ratings and thus the predictive power of the multivariate acoustic model.

This investigation represents the first attempt to investigate concurrent validity and diagnostic precision in the same study. Parsa and Jamieson,[23] for example, used only ROC analysis to investigate the diagnostic accuracy of several acoustic measures, and in contrast, Awan and Roy[22] focused only on the correlation between a multivariate acoustic model and the severity of dysphonia. In this investigation, diagnostic precision was studied using conventional estimates of diagnostic accuracy. In other words, we determined how accurate the AVQI was in determining whether someone does or does not have dysphonia. An ROC-curve was constructed (Figure 6), with an impressive $A_{ROC} = 0.895$. Since this statistic equals the probability of correctly discriminating between normal (normophonic) state and abnormal (dysphonic) states,[24] the result of this study indicates that, based on the AVQI, a clinician could correctly identify almost 90% of the cases. Unfortunately, Giovanni et al,[38] Wuyts et al,[41] Yu et al,[42] and Ma and Yiu[44] did not use A_{ROC} to investigate the classification accuracy of their multivariate constructs. Thus, it is difficult to compare our results with those from these previous studies. In addition, whereas the A_{ROC} in this study describes the discriminatory performance between two conditions (normophonic vs dysphonic), the values provided in Table 1 are based on the classification accuracy between more than two states (eg, normophonic vs slightly dysphonic, moderately dysphonic, severely dysphonic, etc), which also complicates comparisons. In general, however, it seems that based on the wide prediction interval around the regression line (Figure 4), the classification ability of AVQI between intermediate levels of dysphonia may be no better than that reported by other acoustic models.

The ROC-curve can also be used to decide which AVQI cutoff points would determine optimum diagnostic performance. For example, AVQI = 2.36 yields a sensitivity of 91% and a specificity of 59%. The high sensitivity indicates that the AVQI correctly identifies the majority of dysphonic subjects, whereas the lower specificity means that AVQI is less able to correctly identify normophonic subjects (controls). In a diagnostic setting, where one is especially interested in correctly labeling subjects as being dysphonic, this AVQI cutoff point could be proposed as a diagnostic threshold. However, once the results are adjusted for base-rate differences as in the LR analysis, the LRs associated with AVQI cutoff criterion suggest only intermediate-range LR^+ and LR^-, indicating weaker evidence of diagnostic accuracy. In contrast, if a clinician is primarily interested in correctly identifying normals or nondysphonics, such as in a screening test, a higher AVQI cutoff score of 2.95 may be more appropriate, since this score yields a sensitivity of 0.740 and a specificity of 0.963. The

improved specificity is reflected in the LR analysis for this particular cutoff score, which revealed excellent discriminatory accuracy for subjects who test positive (ie, AVQI > 2.95). However, even at this threshold level, the LR− results suggest that a clinician still cannot be sufficiently certain that subjects who test negative (ie, AVQI < 2.95) are indeed normophonic. In the final analysis, the results of the various indices of diagnostic precision are respectable and encouraging, but it is clear that the AVQI requires further refinement as a diagnostic index to distinguish vocally normal individuals from those with dysphonia.

Limitations and future directions

There are a number of limitations related to the analysis method employed and the results reported. First, the cross-validation was internally investigated on numerous subgroups of the same sample on which AVQI was originally modeled. Future investigations should externally confirm the validity of AVQI with new clinical voice samples and ratings. Second, although the perceptual ratings were made by experienced voice clinicians and external standards regarding normophonia were equalized across raters, there was only moderate inter-rater reliability, which likely attenuated the predictive power of the acoustic model. In order to increase the reliability of perceptual ratings, future methods should include multiple anchor stimuli representing different levels in the dysphonia continuum.[53,67] Instead of working with equal-appearing interval scales, future studies could probably benefit from the use of visual analog scales or a hybrid scale such as CAPE-V,[3] incorporating both equal-appearing and visual analog scales. Third, this study was the first to combine sustained vowels and continuous speech in the perceptual as well as the acoustic methodology. However, information regarding what precisely influences the final rating when combined stimuli are presented in this manner is unknown and deserves further attention. It is possible that the perceptual rating of a concatenated voice sample is primarily determined by one of the speaking tasks, for instance by the most dysphonic speaking task or by an average of the two speaking tasks; or alternatively, by a recency or primacy effect, to mention a few possibilities only. Future research should explore the influence of such variables when employing such concatenated samples. Fourth, like other studies,[16,64,68] this study used only the first two sentences of a reading passage. However, it is possible that longer samples of continuous speech will provide improved validity of acoustic and perceptual analysis results.

CONCLUSION

Voice quality assessment traditionally relies on measurement of sustained vowels. To improve ecological validity, acoustic and perceptual assessment of continuous speech should also be considered. The aim of this study was to investigate the feasibility and diagnostic precision of combining both voice contexts into one concatenated sample upon which auditory-perceptual ratings and acoustic measures could be completed. The results supported the viability of such an approach, with respective bivariate associations between listener ratings of dysphonia severity and specific acoustic variables. The diagnostic accuracy of a multivariate acoustic model (AVQI) was assessed, revealing respectable estimates of diagnostic precision. Further refinement of the acoustic algorithm is necessary.

Acknowledgments

The authors wish to thank Dr. James Hillenbrand (from Western Michigan University) for providing the software for the CPP and CPPs measures.

REFERENCES

1. 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.
2. Hillman R. Overview of the consensus auditory-perceptual evaluation of voice (CAPE-V), instrument developed by ASHA Special Interest Division 3. Presented at: *32nd Symposium of the Voice Foundation: Care of the Professional Voice;* June 4–8, 2003; Philadelphia.
3. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol* 2008; epub ahead of print.
4. De Bodt M, Van de Heyning PH, Wuyts FL, Lambrechts L. The perceptual evaluation of voice disorders. *Acta Otorhinolaryngol Belg*. 1996;50:283-291.
5. De Bodt M. *A framework for voice assessment, the relation between subjective and objective parameters in the judgment of normal and pathological voice* [unpublished doctoral dissertation]. Antwerp, Belgium: University of Antwerp; 1997.
6. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19:555-573.
7. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for research. *J Speech Hear Res*. 1993;36:21-40.
8. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33:103-115.
9. Kreiman J, Gerratt BR, Precoda K, Berke G. Individual differences in voice quality perception. *J Speech Hear Res*. 1992;35:512-520.
10. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. 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.
11. Wolfe VI, Martin DP, Palmer CI. Perception of dysphonic voice quality by naïve listeners. *J Speech Lang Hear Res*. 2000;43:697-705.
12. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20:527-544.
13. Rabinov CR, Kreiman J, Gerratt BR, Bielamowicz S. Comparing reliability of perceptual ratings of roughness and acoustic measure of jitter. *J Speech Hear Res*. 1995;38:26-32.
14. Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr.* 1993;45:76-83.
15. 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.
16. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *J Acoust Soc Am*. 2002;112:3014-3021.
17. Yu P, Revis J, Wuyts FL, Zanaret M, Giovanni A. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatr Logop*. 2002;54:271-281.
18. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21:576-590.

19. Zraick RI, Wendel K, Smith-Olinde L. The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *J Voice*. 2005;19: 574-581.
20. Orlikoff RF, Dejonckere PH, Dembowski J, et al. The perceived role of voice perception in clinical practice. *Phonoscope*. 1999;2:89-104.
21. Hammarberg B, Fritzell B, Gauffin J, Sundberg J, Wedin L. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngol*. 1980;90: 441-451.
22. Awan SN, Roy N. Toward the development of an objective index of dysphonia severity: a four-factor acoustic model. *Clin Linguist Phon*. 2006;20:35-49.
23. Parsa V, Jamieson DG. Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *J Speech Lang Hear Res*. 2001;44: 327-339.
24. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd Ed.). Upper Saddle River, NJ: Prentice-Hall; 2000.
25. Buder EH. Acoustic analysis of voice quality: a tabulation of algorithms 1902–1990. In: Kent RD, Ball MJ, eds. *Voice Quality Measurement*. San Diego, CA: Singular Publishing Group; 2000:119-244.
26. Maryn Y, De Bodt M, Van Cauwenberge P, Roy N, Corthals P. Acoustic measurement of overall voice quality: a meta-analysis. *J Acoust Soc Am*. 2008. Submitted for publication.
27. Murry T, Doherty ET. Selected acoustic characteristics of pathological and normal speakers. *J Speech Hear Res*. 1980;23:361-369.
28. Askenfelt AG, Hammarberg B. Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures. *J Speech Hear Res*. 1986;29:50-64.
29. Yiu E, Worrall L, Longland J, Mitchell C. Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding. *Clin Linguist Phon*. 2000;14:295-305.
30. Roy N, Gouse M, Mauszycki SC, Merrill RM, Smith ME. Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *Laryngoscope*. 1995;115:311-316.
31. de Krom G. Consistency and reliability of voice quality ratings for different types of speech fragments. *J Speech Hear Res*. 1994;37:985-1000.
32. Revis J, Giovanni A, Wuyts F, Triglia JM. Comparison of different voice samples for perceptual analysis. *Folia Phoniatr Logop*. 1999;51:108-116.
33. Wolfe V, Cornell R, Fitch J. Sentence/vowel correlation in the evaluation of dysphonia. *J Voice*. 1995;9:297-303.
34. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J Speech Hear Res*. 1996;39:311-321.
35. Prosek RA, Montgomery AA, Walden E, Hawkins DB. An evaluation of residue features as correlates of voice disorders. *J Commun Dis*. 1987;20: 105-117.
36. Eskenazi L, Childers DG, Hicks DM. Acoustic correlates of vocal quality. *J Speech Hear Res*. 1990;33:298-306.
37. Wolfe V, Fitch J, Cornell R. Acoustic prediction of severity in commonly occurring voice problems. *J Speech Hear Res*. 1995;38:273-279.
38. Giovanni A, Robert D, Estublier N, Teston B, Zanaret M, Cannoni M. Objective evaluation of dysphonia: preliminary results of a device allowing simultaneous acoustic and aerodynamic measurements. *Folia Phoniatr Logop*. 1996;48:175-185.
39. Wolfe V, Fitch J, Martin D. Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatr Logop*. 1997;49:292-299.
40. Piccirillo JF, Painter C, Fuller D, Haiduk A, Fredrickson JM. Assessment of two objective voice function indices. *Ann Otol Rhinol Laryngol*. 1998;107: 396-400.
41. 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 Lang Hear Res*. 2000;43:796-809.
42. Yu P, Ouaknine M, Revis J, Giovanni A. Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *J Voice*. 2001;15:529-542.
43. Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurement. *J Voice*. 2004;18:299-304.
44. Ma E, Yiu E. Multiparametric evaluation of dysphonic severity. *J Voice*. 2006;20:380-390.
45. Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS. The voice handicap index (VHI): development and validation. *Am J Speech Lang Pathol*. 1997;6:66-70.
46. Van de Weijer JC, Slis IH. Nasaliteitsmeting met de nasometer. *Tijdschrift voor Logopedie en Foniatrie*. 1991;63:97-101.
47. Van Lierde K. *Nasalance and nasality in clinical practice* [unpublished doctoral dissertation]. Ghent, Belgium: University of Ghent; 2001.
48. AKG Acoustics. *C420: User Instruction. MicroMic Series II*. München, Germany: AKG Acoustics Harman Pro; 2000.
49. Roark RM. Frequency and voice: perspectives in the time domain. *J Voice*. 2006;20:325-354.
50. KayPentax. *Multi-Speech and CSL Software: Software Instruction Manual*. Lincoln Park, NJ: KayPentax; 2004.
51. Boersma P. Praat, a system for doing phonetics by computer. *Glot Int*. 2001;5:341-345.
52. Boersma P, Weenink D [computer program]. *Praat: Doing Phonetics by Computer (Version 4.6.15)*. Amsterdam, The Netherlands: Institute of Phonetic Sciences; 2006. Available at: http://www.praat.org. Accessed February 20, 2007.
53. 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.
54. Kreiman J, Gerratt B. Measuring vocal quality. In: Kent RD, Ball MJ, eds. *Voice Quality Measurement*. San Diego, CA: Singular Publishing Group; 2000:73-101.
55. Hillenbrand J. *SpeechTool, Version 1.56* [computer program], 2006. Available at: http://homepages.wmich.edu/~hillenbr/. Accessed February 20, 2007.
56. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. *J Speech Hear Res*. 1994;37:769-778.
57. Cohen JA. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
58. Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: CRC Press LLC; 1997.
59. Frey LR, Botan CH, Friedman PG, Kreps GL. *Investigating Communication: An Introduction to Research Methods*. Englewood Cliffs, NJ: Prentice-Hall; 1991.
60. Parsa V, Jamieson DG. Identification of pathological voices using glottal noise measures. *J Speech Lang Hear Res*. 2000;43:469-485.
61. Heman-Ackah YD, Heuer RJ, Michael DD, et al. Cepstral peak prominence: a more reliable measure of dysphonia. *Ann Otol Rhinol Laryngol*. 2003;112:324-333.
62. Umapathy K, Krishnan S, Parsa V, Jamieson DG. Discrimination of pathological voices using a time-frequency approach. *IEEE T Bio-Med Eng*. 2005;52:421-430.
63. Dollaghan CA. *The Handbook for Evidence-Based Practice in Communication Disorders*. Baltimore, MD: MD Brookes; 2007.
64. Heman-Ackah YD, Michael DD, Goding GS. The relationship between cepstral peak prominence and selected parameters of dysphonia. *J Voice*. 2002;16:20-27.
65. Dejonckere PH, Wieneke GH. Cepstra of normal and pathological voices: correlation with acoustic, aerodynamic and perceptual data. In: Ball MJ, Duckworth M, eds. *Advances in Clinical Phonetics*. Amsterdam, The Netherlands: John Benjamins Publishing Co; 1996:217-226.
66. Kreiman J, Gerratt B. Perception of aperiodicity in pathological voice. *J Acoust Soc Am*. 2005;117:2201-2211.
67. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am*. 2007;122:2354-2364.
68. Qi Y, Hillman RE, Milstein C. The estimation of signal-to-noise ratio in continuous speech for disordered voices. *J Acoustic Soc Am*. 1999;105:2532-2535.

APPENDIX 1. Script for the detection and extraction of voiced segments in continuous speech, as scripted by the second author (P.C.) and to be used in the program Praat (version 4.6.15).

```
Resample… 22050 50
Rename… original
samplingRate = Get sampling frequency
intermediateSample = Get sampling period
Create Sound… onlyVoice 0 0.001 'samplingRate' 0
select Sound original
To TextGrid (silences)… 50 0.003 -25 0.1 0.1 silence sounding
select Sound original
plus TextGrid original
Extract intervals where… 1 no "does not contain" silence
Concatenate
select Sound chain
Rename… onlyLoud
globalPower = Get power in air
select TextGrid original
Remove
select Sound onlyLoud
signalEnd = Get end time
windowBorderLeft = Get start time
windowWidth = 0.03
windowBorderRight = windowBorderLeft + windowWidth
globalPower = Get power in air
voicelessThreshold = globalPower*(30/100)
select Sound onlyLoud
extremeRight = signalEnd - windowWidth
while windowBorderRight < extremeRight
   Extract part… 'windowBorderLeft' 'windowBorderRight' Rectangular 1.0 no
   select Sound onlyLoud_part
   partialPower = Get power in air
   if partialPower > voicelessThreshold
      call checkZeros 0
      if (zeroCrossingRate <> undefined) and (zeroCrossingRate < 3000)
            select Sound onlyVoice
            plus Sound onlyLoud_part
            Concatenate
            Rename… onlyVoiceNew
            select Sound onlyVoice
            Remove
            select Sound onlyVoiceNew
            Rename… onlyVoice
      endif
   endif
            select Sound onlyLoud_part
            Remove
            windowBorderLeft = windowBorderLeft + 0.03
            windowBorderRight = windowBorderLeft + 0.03
            select Sound onlyLoud
endwhile
select Sound onlyVoice
procedure checkZeros zeroCrossingRate
start = 0.0025
startZero = Get nearest zero crossing… 'start'
findStart = startZero
```

```
findStartZeroPlusOne = startZero + intermediateSample
startZeroPlusOne = Get nearest zero crossing… 'findStartZeroPlusOne'
zeroCrossings = 0
strips = 0
while (findStart < 0.0275) and (findStart <> undefined)
while startZeroPlusOne = findStart
findStartZeroPlusOne = findStartZeroPlusOne + intermediateSample
startZeroPlusOne = Get nearest zero crossing… 'findStartZeroPlusOne'
endwhile
distance = startZeroPlusOne - startZero
strips = strips +1
zeroCrossings = zeroCrossings +1
findStart = startZeroPlusOne
endwhile
zeroCrossingRate = zeroCrossings/distance
endproc
```

APPENDIX 2. Scripts for the calculation of 11 acoustic measures in the program Praat (version 4.6.15) used in this study (as scripted by Y.M.). The original sounds object received the name "Analysis."

2.A. SLOPE OF LTAS

```
select Sound Analysis
To Ltas… 1
ltasSlope = Get slope… 0 1000 1000 10000 energy
```

2.B. TILT OF TRENDLINE THROUGH LTAS

```
select Sound Analysis
To Ltas… 1
Compute trend line… 1 10000
ltasTrendlineTilt = Get slope… 0 1000 1000 10000 energy
```

2.C. FREQUENCY PERTURBATION MEASURES

```
select Sound Analysis
To Pitch (cc)… 0 75 15 no 0.03 0.45 0.01 0.35 0.14 600
select Sound Analysis
plus Pitch Analysis
To PointProcess (cc)
percentJitter = Get jitter (local)… 0 0 0.0001 0.02 1.3
percentJitter = percentJitter*100
relativeAveragePerturbation = Get jitter (rap)… 0 0 0.0001 0.02 1.3
relativeAveragePerturbation = relativeAveragePerturbation*100
pitchPerturbationQuotient = Get jitter (ppq5)… 0 0 0.0001 0.02 1.3
pitchPerturbationQuotient = pitchPerturbationQuotient*100
```

2.D. AMPLITUDE PERTURBATION MEASURES

```
select Sound Analysis
To PointProcess (periodic, cc)… 50 400
select Sound Analysis
plus PointProcess Analysis
percentShimmer = Get shimmer (local)… 0 0 0.0001 0.02 1.3 1.6
percentShimmer = percentShimmer*100
absoluteShimmer = Get shimmer (local_dB)… 0 0 0.0001 0.02 1.3 1.6
amplitudePerturbationQuotient = Get shimmer (apq11)… 0 0 0.0001 0.02 1.3 1.6
amplitudePerturbationQuotient = amplitudePerturbationQuotient*100
```

2.E. GLOTTAL NOISE MEASURES

```
select Sound Analysis
To Pitch (cc)… 0 75 15 no 0.03 0.45 0.01 0.35 0.14 600
select Sound Analysis
plus Pitch Analysis
To PointProcess (cc)
select Sound Analysis
plus Pitch Analysis
plus PointProcess Analysis
voiceReport$ = Voice report… 0 0 75 600 1.3 1.6 0.03 0.45
meanAutocorr = extractNumber (voiceReport$, "Mean autocorrelation: ")
nhr = extractNumber (voiceReport$, "Mean noise-to-harmonics ratio: ")
hnr = extractNumber (voiceReport$, "Mean harmonics-to-noise ratio: ")
```