Multi-parametric evaluation of dysphonic severity

Estella P.-M. MA

Division of Speech Pathology,

The University of Queensland

and

Edwin M.-L. YIU

Voice Research Laboratory, Division of Speech and Hearing Sciences,

The University of Hong Kong

Address for Correspondence:

Estella Ma, Ph.D. Lecturer, Division of Speech Pathology, The University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia.

Tel.: 617-33652797. Fax: 617-33651877. E-mail: estella.ma@shrs.uq.edu.au.

A preliminary version of this manuscript was presented at the 33rd Annual Voice Foundation Symposium: Care of the Professional Voice, Philadelphia, Pennsylvania, June 2004.

MULTI-PARAMETRIC EVALUATION OF DYSPHONIC SEVERITY

ABSTRACT

Summary: In recent years, the multi-parametric approach for evaluating perceptual rating of voice quality has been advocated. The present study evaluates the accuracy of predicting perceived overall severity of voice quality using a minimal set of aerodynamic, voice range profile (phonetogram) and acoustic perturbation measures. One hundred and twelve dysphonic individuals (93 females and 19 males) with laryngeal pathologies and 41 normal controls (35 females and six males) with normal voices participated in this study. Perceptual severity judgement was carried out by four listeners rating the G (overall grade) parameter of the GRBAS scale[1]. The minimal set of instrumental measures was selected based on the ability of the measure to discriminate between dysphonic and normal voices, and to attain at least a moderate correlation with perceived overall severity. Results indicated that perceived overall severity was best described by maximum phonation time of sustained /a/, peak intra-oral pressure of the consonant-vowel /pi/ strings production, voice range profile area and acoustic jitter. Direct-entry discriminant function analysis revealed these four voice measures in combination correctly predicted 67.3% of perceived overall severity levels.

Key words: voice assessment – multiple measures – dysphonic severity – aerodynamic – voice range profile – acoustic perturbation

MULTI-PARAMETRIC EVALUATION OF DYSPHONIC SEVERITY

INTRODUCTION

Contemporarily, dysphonic severity is evaluated by perceptual judgement and instrumental measurements. Perceptual voice evaluation is regarded by clinicians and researchers as the "gold standard" for documenting voice impairment severity[2]. Since it involves listener's subjective judgement of voice quality and severity, it is susceptible to various sources of inter- and intra-listener variability (see review by Kreiman et al.[2]). The literature has shown that perceptual reliability can be affected by the type of rating scale used, the vocal quality and voice samples to be evaluated, the background and experiences of the listeners and the provision of external voice references as anchors for the listeners. Previous studies have also shown that variability for ratings of individual voices, indicated by the width of the 95% confidence interval, is higher for mild-to-moderately rough voices than for voices at the two endpoints (normal and extremely rough) on the rating scale[2-4].

Instrumental measurements, on the other hand, frequently involve instrumentation to quantify dysphonic severity. They are regarded as less subjective and hence a more reliable method to document vocal dysfunction. It is therefore not surprising to find the extensive literature identifying which instrumental measure can best predict perceptual severity, with the intention of replacing perceptual evaluation to document voice impairment severity. However, there has been an inconclusive finding of any single instrumental measure can consistently correlate strongly with perceptual judgement. Some researchers considered the multi-dimensional nature of voice and advocated using more than one type of instrumental measure to predict perceptual severity. This multi-parametric approach allows simultaneous

inclusion of different instrumental voice measures, and therefore enhances the power in differentiating perceptual severity levels[5].

Several authors have investigated the effectiveness of combining different instrumental measures to describe perceptual severity[5-9]. Such effectiveness is commonly evaluated in terms of the association (or concordance) between voice severity levels perceptually judged by listeners and predicted by instrumented measures of the same voice samples. The higher the concordance the stronger association between perceptual evaluation and instrumental measurements. Two different statistics have been used in these studies to evaluate concordance between perceptual and instrumental analysis. The most common is agreement, the percentage of voice samples whose severity levels measured by perceptual and instrumental analysis are the same[5, 6, 9]. Other statistics included correlation coefficient (Pearson's *r*) between the results of perceptual and instrumental analysis[8].

Giovanni et al.[9] employed two acoustic perturbation (jitter and signal-to-noise ratio) with two aerodynamic (voice onset time and glottal leakage) measures that were collected simultaneously using the EVA® system to predict perceptual severity ratings. Perceptual judgement was performed on a 5-point rating scale from '0' normal to '4' severe. Direct-entry discriminant function analysis revealed the four instrumental measures in combination achieved 66.1% (158 out of 239) concordance with perceptual severities. However, this concordance was based on voice samples perceptually rated as '0 (normal)', '2 (moderate)', '3 (intermediate)' and '4 (severe)'. Voice samples rated as '1 (very light or intermittent voice abnormalities)' were not included in the analysis because these samples did not show significant differences from Grade '0' and '2' voice samples. In other words, mildly impaired voice quality was not easily discriminated by the set of acoustic and aerodynamic measures.

Piccirillo and his colleagues carried out two studies in an attempt to develop a multi-parametric voice function index to describe dysphonic severity[7, 8]. They employed

multivariate logistic regression technique and identified a minimal set of four among 14 voice measures that could best distinguish between dysphonic and normal voices. The measures selected were estimated subglottal pressure, phonational frequency range, airflow rate measured at the lips and maximum phonation time. However, the correlation between the combination of four measures and perceived overall severity was only moderate (Pearson's r=0.58).

Wuyts et al.[5] devised the Dysphonic Severity Index using four out of 13 aerodynamic, voice range profile and acoustic perturbation measures. The four voice measures were statistically selected using stepwise logistic regression procedure and represented the minimal set of instrumental measures that could best predict perceptual severity. These four measures were jitter percent, maximum phonation time of sustained /a/, the highest frequency value and the minimum intensity level. Perceptual evaluation was performed on a four-point scale and was taken from the Grade component of the GRBAS scale[1]. However, an integration of these four measures achieved only 49.9% (193 out of 387 subjects) concordance with perceived overall severity.

Yu et al.[6] obtained 11 aerodynamic and acoustic perturbation measures using the EVA® system. Perceptual severities were taken from the overall grade of the GRBAS scheme. The authors employed stepwise discriminant function analysis and identified a set of six measures that could most clearly distinguish among perceptual severity levels. The measures selected were frequency range, the estimated subglottal pressure from /pa/ string, maximum phonation duration of sustained /a/, signal-to-noise ratio, fundamental frequency of sustained /a/ and the Lyapunov coefficient. This set of measures correctly predicted 86% (72 out of the 84 male subjects) of the perceptual severities, which was quite promising. However, the inclusion of male subjects only in their study limited the generalizability of the results to the whole dysphonic population.

The range of concordance between perceptual and instrumental measures reviewed above (49.9% to 86.0%) suggests that even a combination of instrumental voice measures may not reliably predict perceptual severity. A closer examination of the classification results in these studies revealed that a number of mildly and moderately dysphonic voice samples were misclassified by the instrumental measures (see Table 1). The aim of the present study was to evaluate the accuracy of predicting perceived overall severity of voice quality using a minimal set of instrumental measures. In this study, voice samples were judged perceptually on an 11-point equal-appearing interval scale. Pre-judgement perceptual training and external synthesized voice anchors were provided for the listeners. These perceptual protocols have been shown to increase intra-listener agreement and lower inter-listener variability[3, 10, 11]. Instrumental measurements comprised of voice range profile (phonetogram), acoustic perturbation and aerodynamic measures. These instrumental measures were chosen because they have been frequently reported in the literature for documenting dysphonic severity and therefore can facilitate clinical applicability[1, 12].

Put Table 1 here

METHODS

Participants

One hundred and twelve dysphonic subjects (93 females and 19 males) with various types of laryngeal pathologies (see Table 2) and 41 control subjects (35 females and six males) with normal voices participated in this study. All the subjects were native Cantonese speakers and were aged from 20 to 55 years. The mean age of the dysphonic and the control groups were not significantly different (p=0.51).

Put Table 2 here

The dysphonic subjects were consecutive patients recruited from the Voice Research Clinic at The University of Hong Kong and two public hospitals in Hong Kong. None of them had received any voice therapy. All the control subjects reported no history of voice disorders. They were judged by themselves and the first author (EM), who is a speech pathologist with over five years of experiences in assessing and treating voice patients on a daily basis, to be free from any voice problems. Subjects were excluded from this study if they had previous vocal training, hearing problems, speech or language problems, oro-facial abnormalities, or severe respiratory and allergies problems.

Procedures

Each subject undertook a number of voice recordings for perceptual, acoustic perturbation, voice range profile (phonetogram) and aerodynamic evaluation. The recording sequences were randomized among subjects to counterbalance any potential order effects of the recording sequence on the voice samples. All the recordings were carried out in a sound-treated laboratory with background noise kept under 35 dBA. Each subject was seated comfortably in an upright position on a straight-back chair.

Voice recording for perceptual and acoustic perturbation analysis

Each subject was asked to read the Cantonese sentence /ba ba da bɔ/ (meaning 'Father hits the ball') five times at his/her most comfortable daily conversational pitch and loudness. All the voice recordings were recorded directly into the Kay Elemetrics' Computerized Speech Lab Model 4300B Multi-dimensional Voice Program using a professional grade, dynamic microphone (Shure, Beta 87) at a 10 cm mouth-to-microphone distance. The middle trial of the five sentences recorded from each subject was used for perceptual and acoustic

evaluation. Therefore, there were altogether 153 (128 female and 25 male) experimental voice samples.

Perceptual evaluation

Four final year female speech pathology students (mean age=22.0 years) served as listeners for perceptual evaluation. A training program, consisting of rating a set of 25 training stimuli (16 female and nine male) with a wide range of severity levels, was presented to the listeners prior to the actual evaluation of experimental stimuli. These training stimuli were taken from the corpus of voice samples recorded at the Voice Research Laboratory of The University of Hong Kong and did not include any of the experimental stimuli. External synthesized voice anchors representing normal voice, just noticeable and severe levels of breathiness and roughness were given to the listeners as references. The perceptual training program used has been shown to improve inter- and intra-listener agreement[11]. All the listeners had to reach at least 80% agreement with each other in the training program. Once the listeners achieved the 80% criteria, they then rated the actual experimental stimuli independently in a sound-treated booth. Each listener listened to an experimental stimulus, rated the overall severity (i.e., G parameter of the GRBAS scale) of the stimulus and then moved on to the next stimulus. Overall severity was chosen because it contains all the voice quality information of the voice sample[5] and it has been shown to be the most reliable of the voice qualities of the GRBAS scale[13]. The overall severity was rated on an 11-point equal-appearing interval scale with '0' represented normal and '10' represented severely deviant.

In order to evaluate inter- and intra-judge reliability of perceptual voice evaluation, 39 voice samples (25% of the 153 experimental voice samples) were randomly selected and duplicated to incorporate with the experimental samples. This resulted altogether in 192 experimental stimuli (160 female stimuli and 32 male stimuli) for perceptual voice evaluation.

All the listeners rated the set of female stimuli first, followed by the set of male stimuli in another session, separated by one week. The ratings of each stimulus among the four listeners were averaged to get the final rating for that stimulus.

Acoustic perturbation analysis

The same trial of sentence /ba ba da bɔ/ used for perceptual evaluation for each subject was used for acoustic perturbation analysis. Acoustic perturbation analysis was done using the Multi-dimensional Voice Program of Computerized Speech Lab on the whole sentence from the onset of the first word /ba/ to the offset of the last word /bɔ/. Four parameters including mean fundamental frequency, jitter (relative amplitude perturbation), shimmer and noise-to-harmonic ratio values were obtained for each sentence.

Voice range profile (phonetogram) recording

Voice range profiles (phonetogram) were recorded using the Swell's real-time computerized phonetogram Phog 1.0 (AB Nyvalla DSP) program. Each subject was asked to produce a sustained /a/ at his/her minimum and maximum phonational intensity across his/her maximum frequency range. Two frequency measures (the highest and the lowest frequency) and two intensity measures (maximum and minimum intensity) were determined for each subject. The difference between the two frequency and intensity values gave rise to the frequency range and intensity range respectively. In addition, voice range profile area was calculated automatically by the Swell's program for each subject.

Aerodynamic evaluation

Aerodynamic evaluations were done using the Aerophone II (Model 6800, Kay Elemetrics Corp.). During recording, subjects were asked to firmly place a facemask, which

was connected to a transducer module, over the mouth and the nose to measure phonatory airflow and air pressure. Each subject was first instructed to take a deep breath and then produce a sustained /a/ as long as s/he could at his/her most comfortable pitch and loudness. Five trials of maximum sustained /a/ phonation were recorded from each subject. This task was repeated with the vowels /i/ and /u/. The trial of sustained phonation with the longest phonation time and its corresponding mean airflow rate value was obtained for each vowel. Each subject was also instructed to produce five trials of sustained /a/ for five seconds at his/her most comfortable pitch and loudness. This was to assess mean airflow rate in comfortable sustained phonation. The averaged value of the five mean airflow rates in comfortable phonation was collected for each subject.

Two more aerodynamic tasks were carried out to assess peak intraoral pressure. The subjects were asked to keep a flexible silicon rubber tube centrally over the top of the tongue when performing these two tasks. The rubber tube was connected to the air pressure transducer for measuring phonational air pressure. The first task involved producing seven repeated consonant-vowel syllables of bilabial plosive stop /p/ with the vowel /i/ continuously with equal stress on each syllable. Five trials of /pi/ strings were recorded for each subject. The peak intra-oral pressure values of the middle three /pi/ syllables were obtained from each trial. The final peak intraoral pressure value of /pi/ string for each subject was the averaged value of the 15 syllables (3 syllables per trial x 5 trials). The second task involved reading five trials of the Cantonese sentence /ba ba da bɔ/ (meaning 'Father hits the ball') using the subject's most comfortable pitch and loudness as in daily conversations. The peak intraoral pressure values taken from the second word /ba/ was obtained from each trial. The final peak intraoral pressure of sentence production was the averaged values of the five trials.

Statistical analysis

The first statistical procedure involved selection of appropriate voice measures as predicting variables for perceptual ratings. In order to be selected, the voice measure should be able to differentiate between dysphonic and normal voices and should attain at least a moderate correlation coefficient (Pearson's r=0.40 or higher) with perceptual severity ratings. Since a number of statistical tests were carried out for each instrumental measurement, the *alpha* level for each statistical test was re-calculated using Bonferroni adjustment in order to minimize any potential Type I error. This was done by dividing 0.05 by the number of statistical tests carried out for the set of data. Therefore, the *alpha* levels for the set of acoustic perturbation, voice range profile and aerodynamic data were adjusted to 0.0125 (0.05/4), 0.007 (0.05/7) and 0.006 (0.05/9) respectively.

The second statistical procedure involved evaluation of prediction accuracy using direct-entry discriminant function analysis. Since the classification by the discriminant function analysis requires categorical data, the perceptual severity ratings on the 11-point EAI scale were categorized into normal (mean ratings from 0.0 to 0.9), mild (mean ratings from 1.0 to 3.9), moderate (mean ratings from 4.0 to 6.9) and severe (mean ratings from 7.0 to 10.0) levels of severity before the discriminant function analysis was carried out.

RESULTS

Perceptual voice evaluation

The dysphonic group demonstrated significantly more severe voice quality (p=0.0001) than the control group. Pearson's correlation coefficient was used to evaluate the reliability in perceptual voice evaluation. Both inter- and intra-listener reliability were good. Inter-listener correlation coefficients ranged from 0.86 to 0.91 (p=0.0001) and intra-listener correlation coefficients were at least 0.90 (p=0.0001).

Instrumental measurements

Acoustic perturbation measures

Table 3 lists the mean acoustic perturbation values for the dysphonic and the control groups. The dysphonic group demonstrated significantly lower mean fundamental frequency values (p=0.005), significantly higher jitter and shimmer values (both p=0.0001) than the control group. However, the noise-to-harmonic ratio values of both groups were similar and were not significantly different (p>0.05).

Put Table 3 here

Voice range profile (phonetogram) measures

Table 4 shows the mean voice range profile values for the dysphonic and the control groups. The dysphonic group demonstrated significantly lower values in the highest frequency, frequency range, intensity range and profile area (all p=0.0001) than the control group (see Table 4). The dysphonic group also demonstrated significantly higher values in the lowest frequency, maximum intensity and minimum intensity (all p=0.0001) than the control group.

Put Table 4 here

Aerodynamic measures

Table 5 reports the mean aerodynamic values for the dysphonic and the control groups. The dysphonic group demonstrated significantly greater mean airflow rates than the control group for the most comfortable /a/ and maximum phonation of sustained /a/, /i/ and /u/ (all p<0.002) (see Table 5). The dysphonic group also demonstrated significantly higher peak intraoral pressure values of both the /pi/ string and sentence production than the control group

(both p=0.0001). The three maximum phonation durations of the dysphonic group were all significantly shorter (all p=0.0001) than those of the control group.

Put Table 5 here

Correlations between perceived overall severity and instrumental voice measures

Table 6 presents the Pearson's correlation coefficients between perceived overall severity and all the 20 instrumental voice measures. All voice measures except the lowest frequency, the maximum intensity and noise-to-harmonic ratio demonstrated significant correlation with perceived overall severity. Both peak intraoral pressures of /pi/ string and sentence production, the three maximum phonation time values, voice range profile area, jitter and shimmer values attained at least a correlation coefficient of 0.40 with perceptual severity.

Put Table 6 here

Discriminant function analysis

Among all the 20 instrumental measures, the following eight demonstrated statistically significant differences between dysphonic and normal voices, as well as attaining at least moderate and significant correlation (Pearson's r>0.40) with perceived overall severity:

1) Peak intraoral pressure of consonant-vowel /pi/ strings production (r=0.53)
2) Peak intraoral pressure of sentence production (r=0.50)
3) Maximum phonation time of sustained /a/ (r=-0.422)
4) Maximum phonation time of sustained /i/ (r=-0.412)
5) Maximum phonation time of sustained /u/ (r=-0.419)
6) Voice range profile area (r=-0.43)
7) Relative amplitude perturbation (r=0.75)
8) Shimmer percent (r=0.62)

These eight measures could be categorized into four categories: Peak intraoral pressure (measure 1 and 2), maximum phonation time (measure 3, 4 and 5), voice range profile measure (measure 6) and vocal fold vibratory perturbation measures (measure 7 and 8). The measures within each category accounting for the highest percentage of variance (r^2) was chosen as predicting variables for perceptual rating. This selection method maximized the clinical representativeness of the set of predicting variables to be selected with minimal redundancy among them. Therefore, four measures were selected including peak intra-oral pressure of the consonant-vowel /pi/ strings production, maximum phonation time of sustained /a/, voice range profile area and acoustic jitter. They were subsequently used as predicting variables for perceived overall severity.

Table 7 shows the number of subjects and the corresponding percentage predicted by the set of four voice measures for each dysphonic severity level. Correct classifications are listed along the diagonal of the table and are in bold typeface. The overall percentage of correct classifications was 67.3% (103 out of 153 subjects). Subjects who were perceptually rated as normal and severe were more accurately classified by the voice measures (82.5% and 71.9% respectively) than those who were rated as mild and moderate (67.9% and 36.0% respectively).

Put Table 7 here

DISCUSSION

Many researchers are of the view that there is not a single instrumental voice measure which can adequately quantify voice quality and severity[5, 6, 14, 15]. Therefore, multi-parametric evaluation of dysphonia has been advocated. This approach considers the multi-dimensional nature of voice and integrates different voice measures to describe dysphonia. The aim of the present study was to evaluate the accuracy of predicting perceived overall severity of voice

quality using a minimal set of aerodynamic, voice range profile (phonetogram) and acoustic perturbation measures.

In the present study, pertinent instrumental measures for describing perceptual severity were selected from 20 voice range profile, acoustic perturbation and aerodynamic measures. They were chosen based on their ability to discriminate between dysphonic and normal voices, and to attain at least a moderate correlation with perceptual rating. Four voice measures were selected as the minimal set of predicting variables for perceptual rating. These measures were maximum phonation time of sustained /a/, peak intra-oral pressure of the consonant-vowel /pi/ strings production, voice range profile area and acoustic jitter. They represent measures that are more sensitive and have higher discrimination ability for perceptual severity. Therefore, they should be prioritized in clinical voice assessment for evaluating dysphonia. Three of them (subglottal pressure, maximum phonation time and acoustic jitter) have also been reported to be pertinent for describing perceptual severity by other studies reviewed earlier in this paper. Subglottal pressure is the amount of pressure required to initiate a phonation cycle and has been reported to be more reliable than other aerodynamic measures such as mean airflow rate for measuring dysphonic severity[14-16]. Maximum sustained phonation time denotes the longest duration an individual can sustain a phonation after maximum inhalation and is an indicator of phonatory control[1]. Acoustic jitter measures the pitch-to-pitch variability of vocal fold vibration. Previous studies have demonstrated that jitter was sensitive enough to detect presence of dysphonia [17, 18] and voice quality change after vocal fatigue following intensive karaoke singing[19]. The results of this study enhance the clinical value of these three measures to evaluate dysphonia.

Mathematically, voice range profile area is the integration of an individual's intensity levels across his/her entire frequency range. Because of its two-dimensional nature, it is a more powerful measure for describing perceptual severity than either frequency or intensity

per se. Therefore it seems logical to be selected among other voice range profile measures as a significant predictive variable for dysphonia. Sulter, Schutte and Miller[20] found that profile area to be the most sensitive voice range profile parameter in distinguishing individuals with vocal training from those without vocal training. Results from the present study suggest that that profile area can also be a sensitive parameter for describing dysphonia.

Results from discriminant function analysis revealed percentage of concordance between the combination of selected set of voice measures and perceptual judgement was 67.3%. Such percentage of concordance was comparable to that reported by Giovanni et al.[9] (66.1%) and was far higher than that by Wuyts et al.[5] (49.9%). This might be due to the different perceptual evaluation protocols used among these studies. In the present study, perceptual training and external synthesized voice anchors were provided for listeners with an attempt to improve perceptual reliability.

However, the level of concordance in this study was below that reported by Yu et al.[6] (86%). In their study, six predictive variables were used as predictive variables for perceptual severity. On the contrary, only four predictive variables were used in this study. Moreover, the study by Yu et al.[6] comprised male subjects only. Gender-dependent biomechanics might reveal some explanations. It has been suggested that posterior glottal chink is more commonly found in females, accounting for the relatively more prominent perceived breathiness in female than male voices[21]. The interaction of perceptual breathiness with roughness in female voices might affect the predictability of instrumental measures for dysphonia and consequently leads to the lower correct classification rate in the present study. Further investigation of gender-related objective evaluation of dysphonia using multi-dimensional approach is warranted with a larger sample size for both gender groups.

A closer examination of the results of discriminant function analysis revealed that only 36.0% (9 out of 25 subjects) of moderate dysphonic level were correctly predicted by the set

of voice measures. As shown in Table 7, half (50%) of the moderately rated subjects were being under-predicted as either normal or mild. Another one-third (28%) of moderately rated subjects were over-predicted as severe quality. This suggested that the set of four instrumental voice measures as predictive variables for severity ratings in the present study could not adequately identify moderate dysphonia.

For a set of voice measures to be clinically useful in quantifying dysphonic severity, they have to unambiguously differentiate among different dysphonic severity levels. This is clinically critical because it ensures a valid evaluation of dysphonia and hence treatment outcomes. Despite the number of studies which attempted to identify the best combination of instrumental measures for predicting perceptual severity with the intention of quantifying dysphonic severity, reports in the literature revealed that the percentage of concordance between the two measurements could range from 49.9%[5] to 86.0%[6]. Such inconsistent levels of association between perceptual and instrumental measures point to a definite need of more evidence before one can confidently replace perceptual judgement with instrumental evaluation. Until more information on the validity of voice measures is available, one should not over-rely on instrumental voice measures to quantify dysphonic severity. Some authors suggested a comprehensive approach and considered instrumental measures as a complement for perceptual evaluation in assessing dysphonic severity (see the roundtable discussion by Orlikoff et al.[22]). Based on the results of previous and the present study, we agree with these authors and recommend both perceptual and instrumental voice measurements should be included in a clinical voice assessment protocol.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to all subjects who had participated in the present study. We would also like to acknowledge the two anonymous reviewers for their constructive comments on the manuscript. We are grateful to Ms. Ida Leung, Ms. Yuet-Ming Yuen and Mr. Bob Lo for their assistance in data analysis.

REFERENCES

1. Hirano M. *Clinical examination of voice*. Vienna: Springer Verlag, 1981.

2. Kreiman J, Gerratt BR, Kempster GB, Erma A, Berke GS. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research* 1993;36:21-40.

3. Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research* 1993;36:14-20.

4. De Krom G. Consistency and reliability of voice quality ratings for different type of speech fragments. *Journal of Speech and Hearing Research* 1994;37:985-1000.

5. Wuyts FL, De Bodt MS, Molenberghs G, Remacle M, Heylen L, Millet B, et al. The Dysphonic Severity Index: An objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language and Hearing Research* 2000;43:796-809.

6. Yu P, Ouaknine M, Revis J, Giovanni A. Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice* 2001;15(4):529-542.

7. Piccirillo JF, Painter C, Fuller D, Fredrickson JM. Multivariate analysis of objective vocal function. *Ann Otol Rhinol Laryngol* 1998;107(2):107-12.

8. Piccirillo JF, Painter C, Fuller D, Haiduk A, Fredrickson JM. Assessment of two objective voice function indices. *Annuals of Otology, Rhinology and Laryngology* 1998;107:396-400.

9. 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 Phoniatrica et Logopaedica* 1996;48:175-185.

10. Yiu EM-L, Ng C-Y. Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics and Phonetics* 2004;18(1):1-19.

11. Chan KM-K, Yiu EM-L. The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language and Hearing Research* 2002;45(1):111-126.

12. Hirano M. Objective evaluation of the human voice: clinical aspects. *Folia Phoniatrica* 1989;41(1):89-144.

13. Revis J, Giovanni A, Wuyts FL, Triglia J-M. Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopaedica* 1999;51:108-116.

14. Hillman RE, Holmberg EB, Perkell JS, Walsh M, Vaughan C. Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech and Hearing Research* 1989;32:373-392.

15. Hillman RE, Holmberg EB, Perkell JS, Walsh M, Vaughan C. Phonatory function associated with hyperfuntionally related vocal fold lesions. *Journal of Voice* 1990;4(1):52-63.

16. Goozee JV, Murdoch BE, Theodoros DG, Thompson EC. The effects of age and gender on laryngeal aerodynamics. *International Journal of Language and Communication Disorders* 1998;33(2):221-238.

17. Yiu EM-L. Limitations of perturbation measures in clinical acoustic voice analysis. *Asia Pacific Journal of Speech, Language and Hearing* 1999;4:155-166.

18. Yiu EM-L, Worrall L, Longland J, Mitchell C. Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding. *Clinical Linguistics and Phonetics* 2000;14(4):295-305.

19. Yiu EM-L, Chan RM-M. Effect of hydration and vocal rest on the vocal fatigue in amateur karaoke singers. *Journal of Voice* 2003;17(2):216-227.

20. Sulter AM, Schutte HK, Miller DG. Differences in phonetogram features between male and female subjects with and without vocal training. *Journal of Voice* 1995;9(4):363-377.

21. Higgins MB, Netsell R, Schulte L. Aerodynamic and electroglottographic measures of normal voice production: intrasubject variability within and across sessions. *J Speech Hear Res* 1994;37(1):38-45.

22. Orlikoff RF, Dejonckere PH, Dembowski J, Fitch J, Gelfer MP, Gerratt BR, et al. The perceived role of voice perception in clinical practice. *Phonoscope* 1999;2(2):89-106.

Table 1. Percentage of concordance between perceptual judgement and instrumental measurements reported in the literature

Authors (Year)	Perceptual rating scale	% of concordance				Overall % of accuracy
		Normal	Mild	Moderate	Severe	
Giovanni et al. (1996)	G component of GRBAS scale; 5-point ordinal scale.	83	N/A[#]	54.1[+]	61.3	66.1
Wuyts et al. (2000)	G component of GRBAS scale; 4-point ordinal scale.	80	45	54	47	49.9
Yu et al. (2001)	G component of GRBAS scale; 4-point ordinal scale.	96	83	74	100	86.0

[#] Voice samples that were perceptually rated as mild were not included in discriminant function analysis, only those samples rated as normal, moderate and severe were included (refer to text for further details).

[+] Since the voice samples were rated on a 5-point scale with '0' for normal and '4' for severe dysphonia. The 54.1% for moderate level was the averaged value of accuracy values for '2' moderate (57%) and '3' intermediate (51.2%) dysphonia.

Table 2. Types of laryngeal pathologies in the dysphonic group

Laryngeal pathologies	Number of dysphonic subjects
Vocal nodules	43
Thickened vocal fold(s)	37
Chronic laryngitis	13
Vocal fold edema	5
Vocal polyp	4
Vocal fold palsy	4
Miscellaneous/unspecified	6
Total	**112**

Table 3. Mean and standard deviations of acoustic perturbation values

Measures	Dysphonic (*N*=112)		Control (*N*=41)		Independent-*t* tests		
	Mean	**SD**	**Mean**	**SD**	***t***	***df***	***p*-level**
Mean fundamental frequency	196.72	38.01	216.03	34.09	-2.86	151.00	0.005*
Relative amplitude perturbation	1.81	0.99	0.98	0.38	7.48	150.56	0.0001*
Shimmer percent	9.71	3.66	6.25	7.58	8.15	147.11	0.0001*
Noise-to-harmonic ratio	0.24	0.07	0.24	0.04	0.06	113.19	0.96

* Significant at 0.0125 level (2-tailed)

Table 4. Mean and standard deviation of voice range profile values

Measures	Dysphonic (*N*=112)		Control (*N*=41)		Independent-*t* tests		
	Mean	**SD**	**Mean**	**SD**	***t***	***df***	***p*-level**
Frequency measures (Hz)							
Highest frequency	832.19	266.51	1141.35	311.38	-6.07	151.00	0.0001*
Lowest frequency	120.04	25.81	109.31	18.54	2.44	151.00	0.006*
Frequency range†	32.92	7.01	40.08	4.87	-7.09	102.42	0.0001*
Intensity measures (dBA)							
Maximum intensity	109.29	6.08	105.24	6.32	3.61	151.00	0.0001*
Minimum intensity	60.78	7.25	48.71	3.12	14.35	147.11	0.0001*
Intensity range	48.52	8.07	56.54	6.51	-5.71	151.00	0.0001*
Profile area (dBA x semitones)							
Profile area	938.32	285.60	1384.73	280.85	-8.60	151.00	0.0001*

* Significant at 0.007 level (2-tailed)

† Frequency range was measured in semitones

Table 5. Mean and standard deviations of aerodynamic values

Measures	Dysphonic (*N*=112)		Control (*N*=41)		Independent-*t* tests		
	Mean	SD	Mean	SD	*t*	*df*	*p*-level
Mean airflow rates of phonation (l/s)							
Maximum sustained /a/	0.15	0.08	0.11	0.04	3.79	135.39	0.0001*
Maximum sustained /i/	0.14	0.08	0.11	0.04	3.09	142.60	0.002*
Maximum sustained /u/	0.17	0.09	0.12	0.05	3.58	135.17	0.0001*
The most comfortable /a/	0.17	0.09	0.13	0.05	3.98	133.31	0.0001*
Peak intraoral pressures (cm H_2O)							
Consonant-vowel strings	16.95	5.49	9.75	1.85	12.21	150.48	0.0001*
Sentence	12.32	4.13	7.71	1.72	9.72	148.49	0.0001*
Maximum phonation time (s)							
Maximum sustained /a/	15.29	7.79	22.90	8.86	-5.15	151.00	0.0001*
Maximum sustained /i/	16.45	7.64	24.45	8.79	-5.51	151.00	0.0001*
Maximum sustained /u/	15.40	6.67	23.06	9.05	-4.96	56.66	0.0001*

* Significant at 0.006 level (2-tailed)

Table 6. Pearson's *r* between instrumental measures and perceptual severity rating.

Measures	Pearson's *r*
Aerodynamic	
Mean airflow rate of phonation (l/s)	
Maximum sustained /a/	0.28**
Maximum sustained /i/	0.31**
Maximum sustained /u/	0.35**
The most comfortable /a/	0.33**
Peak intraoral pressure (cm H_2O)	
Consonant-vowel strings	*0.53***
Sentence	*0.50***
Maximum phonation time (s)	
Maximum sustained /a/	*-0.42***
Maximum sustained /i/	*-0.41***
Maximum sustained /u/	*-0.42***
Voice range profile	
Frequency measures (Hz)	
Highest frequency	-0.34**
Lowest frequency	0.09
Frequency range †	-0.37**
Intensity measures (dBA)	
Maximum intensity	0.02
Minimum intensity	0.38**
Intensity range	-0.35**
Area (dBA x semitones)	
Profile area	*-0.43***
Acoustic perturbation	
Mean fundamental frequency	-0.18*
Relative amplitude perturbation	*0.75***
Shimmer percent	*0.62***
Noise-to-harmonic ratio	0.13

Pearson's *r* that demonstrated at least a moderate (*r*=>0.40) was italicized.

* Significant at 0.05 level (2-tailed)

** Significant at 0.01 level (2-tailed)

† Frequency range was measured in semitones.

Table 7. Number of subjects (percentage) predicted by four predicting variables[†] into the four severity levels using discriminant function analysis.

Perceptual severity level	Number of subjects (percentage) predicted by instrumental measures				
	Normal	**Mild**	**Moderate**	**Severe**	**Total**
Normal	**33 (82.5)**	5 (12.5)	2 (5.0)	0 (0.0)	40 (100.0)
Mild	13 (23.2)	**38 (67.9)**	5 (8.9)	0 (0.0)	56 (100.0)
Moderate	4 (16.0)	5 (20.0)	**9 (36.0)**	7 (28.0)	25 (100.0)
Severe	0 (0.0)	0 (0.0)	9 (28.1)	**23 (71.9)**	32 (100.0)
Total	50	48	25	30	**153**

Overall prediction accuracy was 67.3%.

Figures in **bold typeface** represent correct predictions and corresponding percentages.

[†] The four predicting variables included maximum sustained /a/ phonation time, peak intra-oral pressure of the consonant-vowel /pi/ strings production, voice range profile area and relative amplitude perturbation.