Clinical Linguistics & Phonetics, September 2010; 24(9): 742–758

informa
healthcare

Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V

SHAHEEN N. AWAN[1], NELSON ROY[2], MARIE E. JETTÉ[3], GEOFFREY S. MELTZNER[4], & ROBERT E. HILLMAN[5]

[1]*Department of Audiology & Speech Pathology, Bloomsburg University of Pennsylvania, Bloomsburg, PA, USA,* [2]*Department of Communication Sciences and Disorders, The University of Utah, Salt Lake City, UT, USA,* [3]*University of Wisconsin-Madison, Wisconsin Institutes for Medical Research, Madison, WI, USA,* [4]*BAE Systems, Inc., Burlington, MA, USA, and* [5]*Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Surgery and Health Science and Technology, Harvard Medical School, Boston, MA, USA*

(Received 10 March 2010; Accepted 5 May 2010)

Abstract
This study investigated the relationship between acoustic spectral/cepstral measures and listener severity ratings in normal and disordered voice samples. CAPE-V sentence samples and the vowel /ɑ/were elicited from eight normal speakers and 24 patients with varying degrees of dysphonia severity. Samples were analysed for measures of the cepstral peak prominence (CPP), the ratio of low-to-high spectral energy, and their respective standard deviations. Perceptual ratings of overall severity were also obtained for all samples. Results showed that all acoustic variables combined in a four-factor model which correlated with perceived severity with $R = 0.81$ ($R^2 = 0.65$). For the vowel /ɑ/, a five-factor model incorporating all acoustic variables and gender correlated with perceived severity with $R = 0.96$ ($R^2 = 0.91$). Results indicate that a strong relationship between perceptual and acoustic estimates of dysphonia severity can be achieved in both continuous speech and vowel contexts using a model incorporating spectral/cepstral measures.

Keywords: *cepstrum, cepstral analysis, spectral analysis, continuous speech analysis, dysphonia severity*

Introduction

Auditory-perceptual assessment of voice quality and severity is an essential component of voice disorder evaluation. The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) scale was ostensibly developed to standardize clinical auditory-perceptual assessment of voice, and to describe the severity of perceptual attributes in a manner that would facilitate communication among clinicians (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, and

Correspondence: Shaheen N. Awan, PhD, Department of Audiology & Speech Pathology, Bloomsburg University of PA, Centennial Hall, 400 East Second St., Bloomsburg, PA 17815-1301, USA. E-mail: sawan@bloomu.edu

ISSN 0269-9206 print/ISSN 1464-5076 online © 2010 Informa UK Ltd.
DOI: 10.3109/02699206.2010.492446

Hillman, 2009). The CAPE-V elicits sustained vowels as well as connected speech productions in both sentence reading and spontaneous speech. By deliberately sampling a variety of voice contexts, the CAPE-V potentially reveals task-dependent differences in vocal performance. Although the CAPE-V represents an important step towards standardized clinical voice quality evaluation, it is common for voice clinicians to supplement auditory-perceptual evaluation with more objective assessment methods including acoustic analysis. Historically, acoustic analysis has been completed on sustained vowel contexts, and the CAPE-V vowel samples certainly lend themselves to traditional measures of voice analysis such as jitter, shimmer, and harmonic-to-noise ratios (HNRs). However, extending sustained vowel analysis methods to connected speech has been problematic for a variety of reasons. First, traditional perturbation measures such as jitter and shimmer will likely be artificially inflated due to the effects of intonation and unvoiced segments. Second, measures of jitter and shimmer can be negatively influenced by the relatively short duration vowel segments observed in continuous speech (Zhang and Jiang, 2008). Therefore, the combination of pitch and loudness variations, noise produced via true consonant production, and short voicing segments in connected speech may invalidate the results of traditional perturbation measurements.

In contrast to traditional time-based perturbation methods, several authors have attempted to use spectral-based acoustic methods to analyse normal and disordered voice in continuous speech. In particular, measures of the cepstrum are reportedly strong indicators of dysphonia severity in both sustained vowel and continuous speech contexts. The cepstrum (Noll, 1964) has been described as a Fourier transform of the logarithm power spectrum (Baken, 1987), and may be used to display and objectively determine the extent to which the dominant rahmonic (an anagram of 'harmonic', often associated with the vocal fundamental frequency) is individualized and emerges out of the background noise level. The principal advantage of spectral analysis methods (i.e., frequency-based analysis) is that estimates of aperiodicity and/or additive noise may be achieved without the identification of cycle boundaries.

Several investigators have evaluated the ability of cepstral measures to quantify the presence and severity of dysphonia in *sustained vowel* productions (Awan and Roy, 2005; 2006; 2009; Hartl, Hans, Vaissiere, Riquet, and Brasnu, 2001; Hillenbrand, Cleveland, and Erickson, 1994; de Krom, 1995; Wolfe, Martin, and Palmer, 2000). These studies have consistently reported that increased vocal severity/abnormality is associated with a decrease in amplitude of the cepstral peak (i.e., lower harmonic energy) and an increase in high frequency spectral energy. Furthermore, researchers have employed spectral/cepstral methods to analyse samples of normal vs disordered voice in *continuous speech*. For instance, Hillenbrand and Houde (1996) reported strong correlations between perceptual ratings of breathiness and measures of the cepstral peak prominence (CPP) obtained from samples of the second sentence of 'The Rainbow Passage' (Fairbanks, 1960). Qi, Hillman, and Milstein (1999) used linear prediction analyses to estimate the signal-to-noise ratio (SNR) of samples of 'The Rainbow Passage' obtained from 87 voice disordered speakers. The investigators reported an $r = -0.76$ between categorical severity ratings and speech SNRs and an $r = -0.78$ between continuous direct magnitude scaled ratings and SNRs. Parsa and Jamieson (2001) reported 96% correct classification accuracy (normal vs disordered) using overall spectral tilt, frequency domain HNR, and a measure of spectral flatness derived from samples of 'The Rainbow Passage' for 53 normal and 175 voice disordered patients. Heman-Ackah, Michael, and Goding (2002) reported that measures of the cepstral peak were the strongest individual correlates of overall dysphonia and ratings of breathiness in both continuous speech and in sustained vowel samples. Heman-Ackah, Heuer, Michael, Ostrowski,

Horman, Baroody, et al. (2003) reported an overall sensitivity of 87% and specificity of 90% in detecting overall dysphonia from continuous speech using measures of the cepstral peak prominence in 281 running speech samples (176 female; 105 male) that had been rated for severity using an undifferentiated 100 mm scale, and concluded that cepstral measures obtained from continuous speech had better sensitivity, specificity, and positive and negative predictive values than time-based measures of perturbation. Halberstam (2004) also analysed measures of the cepstral peak prominence (60 normal and disordered rated samples of 'The Rainbow Passage') and reported a strong correlation between perceived hoarseness and measures of the CPP. Halberstam also reported that correlations between cepstral measures from speech and perceived dysphonia were stronger than those observed between sustained vowel measurements and perceptual judgements. Laflen, Lazarus, and Amin (2008) examined the relative deviation of the fundamental frequency and the intensity of the cepstral peak in normal and voice disordered vowel, CVC syllable, and continuous speech samples ('How are you?'), and reported that spectral/cepstral measures from connected speech samples were effective discriminators between normal vs disordered samples. Recently, Awan, Roy, and Dromey (2009) reported that an acoustic model composed of spectral and cepstral measures (including the CPP) produced estimates of dysphonia severity that strongly correlated with perceived dysphonia severity ratings ($R = 0.85$; $R^2 = 0.73$) in a set of pre- and post-treatment continuous speech samples from 104 females with primary muscle tension dysphonia (MTD). These authors reported that spectral/cepstral estimates of dysphonia severity may be used as an effective treatment outcomes measure in pre vs post-treatment continuous speech samples. Collectively, the results of these and other studies suggest that spectral/cepstral measures offer promise as sensitive indices of dysphonia severity in both sustained vowels and connected speech (Youri, Roy, De Bodt, et al., 2009).

The CAPE-V is a relatively new tool for documenting auditory-perceptual features of the voice, and elicits speech/voice samples that lend themselves to objective acoustic analyses, including both sustained vowels and connected speech. However, analysis of sustained vowels alone may not capture all salient characteristics of a patient's voice. Several authors have asserted that continuous/running speech may (1) provide a more ecologically valid assessment of the patient's control of vocal parameters such as vocal quality, (2) reveal increased degrees of voice impairment, and (3) correlate better with perceptions of dysphonia (Eadie and Doyle, 2005; Halberstam, 2004; Laflen et al., 2008; Maryn, Corthals, Van Cauwenberge, Roy, and De Bodt, in press; Qi et al., 1999; Roy, Gouse, Mauszycki, Merrill, and Smith, 2005; Yiu, Worrall, Longland, and Mitchell, 2000). In addition, continuous speech incorporates important vocal attributes such as rapid voice onset and termination and variations in fundamental frequency and amplitude that may have a relatively large impact on short duration signals and, in turn, may be highly relevant to the perception of dysphonia in everyday situations and to clinical decisions regarding the voice quality of the patient (Hammarberg, Fritzell, Gauffin, Sundberg, and Wedin, 1980; Parsa and Jamieson, 2001). In this regard, the sentences contained in the CAPE-V aim to elicit voice characteristics such as soft glottal attacks and voiceless to voiced transitions ('How hard did he hit him?'), the presence of possible voiced stoppages or spasms and the ability to maintain consistent voicing ('We were away a year ago'), the presence of hard glottal attacks ('We eat eggs every Easter'), and the ability to transition easily between voiceless stop-plosive production and vowel production ('Peter will keep at the peak'). Differences in dysphonia severity across certain voice contexts have been considered the hallmark of specific voice disorders (Roy et al., 2005). Thus, the

CAPE-V has the potential to expose such task-dependent performance, thereby revealing the nature of the voice disorder.

It seems clear that auditory-perceptual and acoustic analyses should be conducted on *both* sustained vowel and connected speech samples. Yet, little is known regarding the sensitivity of acoustic analysis techniques to different levels of dysphonia severity, especially when applied to connected speech contexts. The purpose of this study was to investigate acoustic correlates of auditory-perceptual judgements of dysphonia severity collected via the CAPE-V. In particular, this study examined the relation between spectral/cepstral measures and severity ratings from normal, mild, moderate, and severely disordered male and female voice samples in both sustained vowel and sentence production tasks.

Methodology

Speech samples

Speech/voice samples from 32 speakers were chosen for auditory-perceptual and acoustic analyses. The speaker group was comprised of eight normal speakers and 24 voice disordered patients with varying degrees of dysphonia severity. Normal samples were from four males and four females between 25–32 years of age who had normal voice quality (as judged by a certified speech-language pathologist), no history of voice disorders, were native speakers of standard American English, and had passed a hearing screening (25dB at 500, 1000, 2000, 4000 Hz). The disordered voice samples were obtained from 12 males and 12 females between 21–78 years of age, divided equally in terms of gender distribution among three categories of dysphonia severity (mild, moderate, and severe). The voice samples used in this study were selected to represent the severity continuum encompassing normal and disordered voice—age as a variable was not controlled in this study. The disordered samples were selected from a database of ~ 300 clinical voice evaluation recordings to represent a variety of diagnoses and range of dysphonia severity (mild, moderate, severe) as judged by the examining clinician using the CAPE-V (see Table I). Recordings with detectable accented English, or with perceived loudness (too loud or too soft), pitch (too low or too high), nasality/resonance (hypernasal or hyponasal), or intelligibility (unintelligible) issues were excluded from the database. This was done to minimize any potential effects on listener ratings related to these factors. Two certified speech-language pathologists then divided the remaining database samples into male/female, and again into organic/non-organic diagnoses, with organic referring to a structural disorder caused by a lesion of the vocal fold(s) and non-organic

Table I. Summary of gender and diagnoses for subjects with different levels of dysphonia.

	Dysphonia severity		
Gender	Mild	Moderate	Severe
Male	muscle tension dysphonia	muscle tension dysphonia	unilateral paralysis
Male	unilateral paralysis	unilateral paralysis	unilateral paralysis
Male	polyp	papilloma	cancer
Male	papilloma	cyst	amyloidosis
Female	muscle tension dysphonia	unilateral paralysis	unilateral paralysis
Female	unilateral paralysis	unilateral paresis	muscle tension dysphonia
Female	cyst	nodules	nodules
Female	cyst	Reinke's oedema	cyst

including both neurogenic and functional diagnoses. Samples were then selected to maintain a gender balance, as well as to survey a wide range of dysphonia severity.

All subjects were required to perform a standard protocol used for clinical voice recordings that included the voice and speech tasks for the CAPE-V. Selected elements from the CAPE-V were used for the current study including the vowel /ɑ/ (sustained for 3–5 seconds at comfortable/constant pitch and loudness) and four sentences that are designed to elicit different laryngeal behaviours: (1) easy onset of phonation ('How hard did he hit him?'), (2) all voiced sounds ('We were away a year ago'), (3) hard glottal attack ('We eat eggs every Easter'), and (4) weighted with voiceless plosives ('Peter will keep at the peak'). All speech samples were recorded in the same sound-treated room (International Acoustics Corporation: ambient noise < 25 dBA). The speech acoustic signal was captured by a high quality, head-mounted condenser microphone (Sennheiser model MKE-2) that was held a constant distance of 15 cm from the lips. The microphone signal was amplified (Symetrix model SX202) and then digitized using the KayPentax Computerized Speech Laboratory (CSL model 4400) at a sampling rate of 50 kHz for sustained vowels and 25 kHz for sentences with 16 bits of resolution. All vowel samples were later down-sampled to 25 KHz to provide a common sampling rate prior to acoustic analyses. The total number of speech/voice samples analysed was 160 (32 speakers (8 Normal + 24 Disordered) × 5 samples (1 vowel + 4 sentences) each).

Acoustic analyses

All samples were analysed using a Windows-based computer program developed by the first author and used by Awan et al. (2009) as a predictor of vocal severity and treatment outcomes measure in continuous speech. The program applies spectral and cepstral analysis methods described by Hillenbrand et al. (1996) and Awan and Roy (2005; 2006; 2009), which have been used to characterize voice quality type and predict dysphonia severity in normal and disordered voice samples. All of the spectral/cepstral measures are derived from this single program with a common core of algorithms. Unlike the computer algorithms employed in previous studies by Awan and Roy (2005; 2006; 2009) which combined spectral/cepstral based acoustic methods with time-based measures such as shimmer and pitch sigma, the computer algorithms used in this study are solely spectral based and do not depend upon accurate identification of cycle boundaries for any of the measurements obtained.

The following describes the basic procedures used in the analysis of the continuous speech samples:

(1) The speech sample was divided into a series of 1024 point overlapping frames (75% overlap). For each analysis frame, a Hamming window was applied and a 1024 point discrete Fourier transformation (DFT) was computed. As described by Baken (1987), the DFT was then converted to the log power spectrum, followed by a second DFT. This procedure results in the cepstrum (essentially a Fourier transform of a Fourier transform). The cepstrum of a highly periodic signal is characterized by a prominent peak which is the dominant rahmonic (i.e., fundamental period) of the signal, and has been referred to as the cepstral peak prominence (CPP—Hillenbrand et al., 1994; Hillenbrand and Houde, 1996).

(2) As described by Hillenbrand and Houde (1996), a combination of time and quefrency averaging can aid in smoothing the cepstrum prior to identification of the CPP. In this study, a 7-frame cepstral averaging was carried out, with each smoothed

cepstral frame being calculated from the average of the current frame with the three previous and three subsequent cepstral frames. Cepstral averaging across time was followed by 11-bin quefrency averaging, in which each cepstral coefficient (i.e., data value observed on the abscissa of the cepstrum) was replaced by the average of the current coefficient with the five previous and five subsequent cepstral coefficients. Figure 1 provides an example of a smoothed cepstral frame.

(3) For each frame, several acoustic measures were computed. From the original unsmoothed window, a ratio of low/high frequency (L/H) spectral energy was calculated as a measure of spectral tilt (referred to as the DFT Ratio (DFTR) in Awan and Roy, 2005; 2006) using a 4000 Hz cut-off and reported in decibels. From the smoothed cepstral frames, the dominant rahmonic (i.e., CPP) was identified and the ratio of the CPP to the expected amplitude of the CPP (as estimated via linear regression analysis) was computed. For the purposes of this study, the search for the cepstral peak was restricted to quefrencies of 3.3–16.7 ms (300–60 Hz; Hillenbrand and Houde, 1996). Pilot work indicated that removal of signals which had normalized CPP values < 0 dB (i.e., dominant cepstral rahmonics that had an amplitude lower than the expected value as determined via subsequent linear regression analyses) helped to remove low amplitude highly aperiodic signals often associated with breath sounds and/or portions of unvoiced consonants.

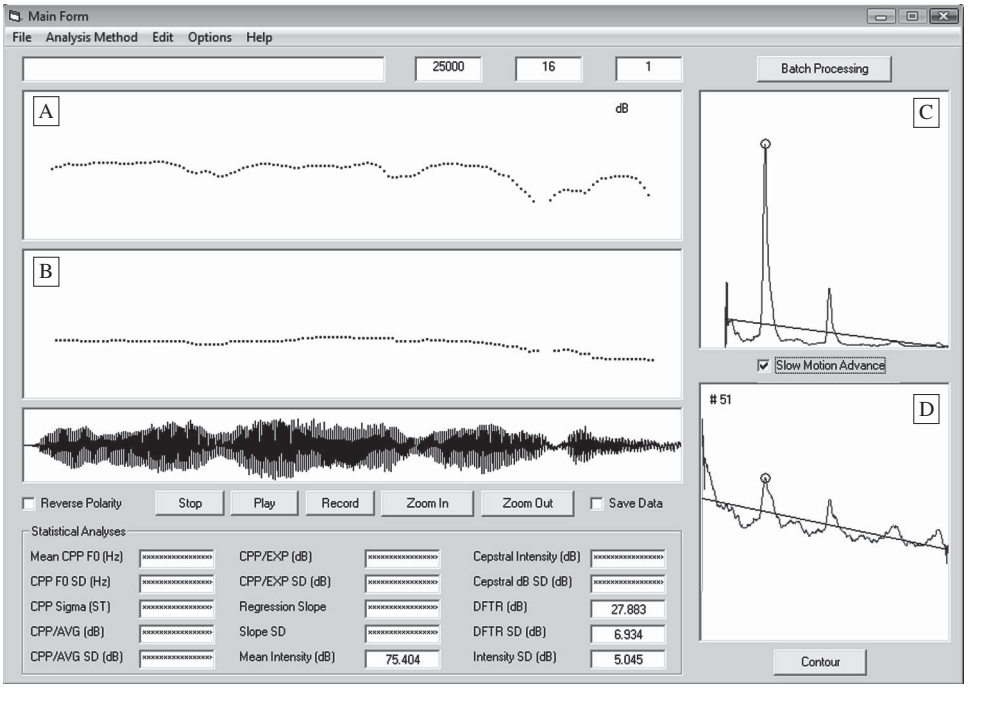

Figure 1. Application of spectral/cepstral analysis methods to a sample of continuous speech ('We were away a year ago'). The upper window ('A') shows the varying CPP (dB) over time; the lower window ('B') shows the raw fundamental frequency (Hz) contour. Window 'C' provides an example of a smoothed cepstral frame using raw cepstral coefficients for quefrencies > 2 ms. Window 'D' provides a smoothed cepstral frame computed in dB and extending up to quefrencies = .1 ms. In each window, the cepstral peak has been automatically identified, and linear regression lines have been computed for normalization of the CPP.

(4) Once analysis was completed across all of the analysis windows, the mean L/H spectral ratio and CPP, as well as their respective standard deviations (SDs), were calculated for the entire signal. Because the various spectral/cepstral measures were to be averaged across relatively long duration samples of non-stationary running speech affected by vowel-to-consonant transitions and intonation, it was reasoned that measures of the average variability (standard deviation) of each of the key variables would be important to collect. Furthermore, previous studies have indicated that measures of variability may be effective in characterizing the severity of the voice (Awan and Roy, 2006; 2009; Callan, Kent, Roy, and Tasko, 1999; Wolfe and Steinfatt, 1987). The computer program then displayed the varying CPP over time for the speech sample (see Figure 1) and saved all computed statistics in a data file.

Auditory-perceptual severity ratings

Twenty-five listeners (speech pathology graduate students with normal hearing and no reported history of significant voice disorder) were asked to use a computerized graphical user interface (GUI) version of 4 CAPE-V scales (perceptual attributes of overall severity, roughness, breathiness, and strain) to rate the 4 CAPE-V sentences (all voiced, easy onset, hard glottal attack, weighted with voiceless plosives) and sustained /ɑ/ vowel productions. The GUI (created using Matlab [Mathworks, Natick, MA]) mapped each CAPE-V scale to a visual analogue scale (VAS) whereby listeners used mouse-controlled sliders to indicate their ratings. Indicators for mildly deviant (MD), moderately deviant (MO), and severely deviant (SE) were placed at the exact distances from the ends of each scale as portrayed in the paper version of the CAPE-V. Each VAS was scaled from 0–100 and distance from the left end of the scale (i.e. 0) that the listener placed the slider was used as the perceptual rating. The use of a GUI version of the CAPE-V test simplified the measurement of the perceptual ratings by removing the need to physically measure marks made on a paper and pen scale.

Listeners participated in five separate listening sessions (one for each stimulus type) and a training session (examples of disordered voices played and consensus reached for ratings), and received payment of $50 per session. Each stimulus item for each speaker was presented five times in random order, with listeners able to repeatedly play each presentation as many times as needed to rate all four perceptual parameters (listeners were instructed to rate one parameter at a time and to listen to the stimulus a separate time for each rating). Each stimulus was repeated five times in a psychometric-based data analysis approach (Shrivastav, Sapienza, and Nandur, 2005) to better deal with listener reliability issues. For the purposes of this study, only the ratings of overall severity were used. The mean rating across all 25 judges $\times$ 5 judgements each (125 ratings) was computed for each of the 32 speakers for each of the CAPE-V sentences and the sustained vowel /ɑ/ sample.

Reliability

Inter-judge reliability for the overall severity ratings was assessed across all samples (four sentences and one vowel) using the intra-class correlation coefficient (ICC; McGraw and Wong, 1996), a measure of the degree of consistency among judges. Results indicated an Average Measures ICC = 0.993, $p < 0.001$ (95% confidence interval of 0.992–0.995) and a Single Measures ICC = 0.855, $p < 0.001$ (95% confidence interval of 0.826–0.882) for the rating of severity of vocal quality disruption. These results indicate that the judges were able to differentiate between different levels of dysphonia severity, and that the average rating of the

scores of the multiple judges was highly reliable, despite any apparent differences in severity rating. Intra-judge reliability was assessed by computing Pearson's *r* correlations among the repeated ratings for each sample. Across all judges, the mean intra-judge Pearson's $r = 0.91$ (SD = 0.06) and was considered to be acceptable for the purposes of this study.

Results

Statistical analyses were computed using *SPSS v.15.0* (SPSS Inc., Chicago, IL) and *SigmaPlot 10.0 for Windows* (Systat Software Inc., San Jose, CA). Stepwise multiple regression analysis was computed to ascertain the strength of correlation between listener perceived ratings and acoustic measures for (a) continuous speech samples (sentences) and (b) sustained vowel /ɑ/ productions. Separate analyses for continuous speech vs sustained vowel productions were conducted because preliminary regression analyses had shown substantial differences in the strength and direction of regression coefficients for these different sample types. In particular, the direction (+/− sign) of coefficients associated with the standard deviation variables (CPP SD and L/H spectral ratio SD) was observed to be negative for continuous speech samples vs positive for sustained vowel samples. Awan et al. (2009) have also reported that the variability of the CPP and the L/H spectral ratio reduce with increased dysphonia severity in continuous speech. In contrast, several examples in the literature (Awan and Roy, 2006; 2009; Callan, Kent, Roy, and Tasko, 1999) have indicated that increased variability in sustained vowel production is a common observation when the severity of dysphonia increases. For each stepwise regression analysis, the independent variables were the CPP; CPP SD; the L/H spectral ratio; and the L/H spectral ratio SD. In addition to the aforementioned acoustic variables, a qualitative variable was added to the stepwise regression analysis for continuous speech to account for Sentence type (four levels). A qualitative variable was also added to stepwise regression analyses for both continuous speech and vowels to account for Gender (two levels).

For the analysis of continuous speech samples ($n = 128$), all four acoustic variables significantly combined in a four-factor model which correlated with listener perceived severity with $R = 0.81$ ($R^2 = 0.65$; Adjusted $R^2 = 0.64$; see Table II). Although the CPP SD was observed to enter first into the stepwise regression procedure, the CPP was found to have the strongest beta coefficient, indicating that this variable was the strongest contributor to the overall R^2. The Sentence and Gender variables were not significant contributors to the final multiple regression model for the analysis of continuous speech. The stepwise multiple

Table II. Stepwise regression summary table for the analysis of CAPE-V sentence samples. The relative contribution of each acoustic variable to the multiple regression analysis is provided. Multiple correlation *R*, R^2, and change in R^2 values are provided.

Acoustic variable	*R*	R^2	R^2 change	*F* change	*df*	*p*	Standardized beta coefficient
CPP SD	.676	.457	.457	105.992	1,126	< .001	−.336
CPP	.763	.582	.125	37.239	1,125	< .001	−.446
L/H spectral ratio SD	.786	.618	.036	11.812	1,124	.001	−.305
L/H spectral ratio	.805	.647	.029	10.252	1,123	.002	−.241

CPP, the ratio of the amplitude of the cepstral peak prominence to the expected cepstral amplitude; L/H spectral ratio, low vs high frequency spectral energy; SD, standard deviation; *R*, multiple correlation; *df*, degrees of freedom.

Table III. Stepwise regression summary table for the analysis of the sustained vowel /ɑ/. The relative contribution of each acoustic variable to the multiple regression analysis is provided. Multiple correlation R, R^2 and change in R^2 values are provided.

Acoustic variable	R	R^2	R^2 change	F change	df	p	Standardized beta coefficient
CPP	.837	.701	.701	70.358	1,30	< .001	−.500
CPP SD	.892	.796	.094	13.405	1,29	.001	.274
L/H spectral ratio	.920	.701	.050	9.131	1,28	.005	−.242
Gender	.938	.879	.033	7.464	1,27	.011	−.217
L/H spectral ratio SD	.956	.914	.035	10.520	1,26	.003	.221

CPP, the ratio of the amplitude of the cepstral peak prominence to the expected cepstral amplitude; L/H spectral ratio, low vs high frequency spectral energy ratio; SD, standard deviation; R, multiple correlation; df, degrees of freedom.

regression analysis produced the following predictive equation for the predicted CAPE-V severity of sentences, S_s:

$$S_S = 148.68 - (5.91 \times CPP) - (11.17 \times \sigma_{CPP}) - (1.31 \times SR) - (3.09 \times \sigma_{SR}) \quad (1)$$

where σ_{CPP} is the standard deviation of the CPP, SR is the L/H Spectral Ratio, and σ_{SR} is the standard deviation of the L/H Spectral Ratio.

For the analysis of sustained vowel /ɑ/ productions ($n = 32$), a five-factor model incorporating all four acoustic variables and Gender was observed to correlate with listener perceived severity with $R = 0.96$ ($R^2 = 0.91$; Adjusted $R^2 = 0.90$; see Table III). In sustained vowel analysis, the CPP was again observed to have the strongest beta coefficient and strongest contribution to the overall R^2. The stepwise multiple regression analysis produced the following predictive equation for the predicted CAPE-V severity of the vowel /ɑ/, S_v:

$$S_v = 84.20 - (4.40 \times CPP) + (10.62 \times \sigma_{CPP}) - (1.05 \times SR) + (7.61 \times \sigma_{SR}) - (10.68 \times G) \quad (2)$$

where G is the gender variable (Male = 0; Female = 1).

Using the aforementioned equations, acoustic severity estimates were computed for each of the analysed CAPE-V sentences, as well as for the sustained vowel /ɑ/. Paired t-tests were computed to determine if significant differences existed between the acoustically estimated severity ratings vs the perceived severity ratings, and Pearson's r correlations were computed to determine the strength of association between estimated vs. perceived ratings (see Table IV). Results indicated that the use of a consistent set of acoustic variables and regression coefficients can result in strong correlations and a high degree of agreement with listener perceived severity ratings across a variety of samples. The strongest correlations between listener perceived severity and acoustically estimated severity ratings were observed for Sentence 3 ('We were away a year ago') and the sustained vowel /ɑ/ (see Table IV).

Effect of dysphonia severity

Figures 2 and 3 provide error bar charts for acoustically estimated vs listener perceived dysphonia severity ratings for each CAPE-V sample across the normal, mild, moderate, and severe subject groups. As expected from the previously reported correlation data, the greatest correspondence in mean estimated vs perceived ratings is found for sentence 3 ('We

Table IV. Mean perceived vs acoustically estimated severity ratings for the various CAPE-V samples (standard deviation are provided in parentheses). Paired *t*-results and Pearson's *r* correlations are also provided.

CAPE-V sample	Mean perceived severity rating	Mean estimated severity rating	Paired *t* results	Pearson's *r* correlation**
All sentences combined	32.16 (26.33)	32.17 (21.18)	$t = -.004$, $df = 127$, $p = .99$	$r = .81$
'How hard did he hit him?'	33.58 (27.57)	34.17 (20.36)	$t = -.21$, $df = 31$, $p = .83$	$r = .83$
'We were away a year ago'	30.93 (26.14)	28.41 (24.51)	$t = 1.04$, $df = 31$, $p = .31$	$r = .86$
'We eat eggs at Easter'	33.20 (25.73)	34.03 (21.22)	$t = -.29$, $df = 31$, $p = .77$	$r = .78$
'Peter will keep at the peak'	30.94 (26.98)	32.06 (18.73)	$t = -.36$, $df = 31$, $p = .72$	$r = .77$
Sustained vowel /ɑ/	36.11 (24.95)	36.12 (23.85)	$t = -.01$, $df = 31$, $p = .99$	$r = .96$

** All Pearson's *r* correlations are significant at $p < .001$.

were away a year ago') and for the sustained vowel /ɑ/. For all of the sentences, the figures indicate a tendency for acoustic estimations of severity to over-estimate mean listener perceived severity ratings for those samples that approach the normal end of the severity continuum, and to under-estimate mean listener perceived severity ratings for those samples that approach the more severe end of the continuum.

Receiver operating characteristic (ROC) analysis

The accuracy of an assessment tool can be evaluated by the sensitivity and specificity of the test. Sensitivity is defined as the proportion of participants with the disease (i.e., cases) who have a positive test, whereas the specificity is the proportion of participants without the disease (i.e., non-cases) who have a negative test. In tests that yield continuous data like those produced by the CAPE V severity rating scale used in this study, several values of sensitivity and specificity are possible, depending on the cut-off point chosen to define a positive test. This trade-off between sensitivity and specificity can be displayed graphically using a receiver operating characteristic (ROC) curve. To generate a ROC curve, the investigator selects several cut-off 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 (Zweig and Campbell, 1993). An optimal diagnostic test is one that reaches the upper left corner of the graph. A worthless test follows the diagonal from the lower left to the upper right corners, suggesting that at any cut-off the true-positive rate is the same as the false-positive rate. In this study, ROC analysis was computed to ascertain the degree to which the acoustic algorithm could separate dysphonic participants (i.e., cases) from non-dysphonic participants (i.e., normals or non-cases) as determined originally by auditory-perceptual judgements. In addition, the positive likelihood ratio (LR+) indicates the level of confidence that a positive test score truly reflects the presence of a disorder; in contrast, the negative likelihood ratio (LR-) indicates the level of confidence that a normal-range score truly reflects a non-disordered state. Using acoustically-derived severity estimates from all CAPE-V samples, ROC analysis indicated that an acoustically estimated severity rating cut-off of 22.67 would result in sensitivity = 72% and specificity of 80% (LR+ = 3.58; LR− = 0.35). The computed likelihood ratio scores are indicative of moderately strong

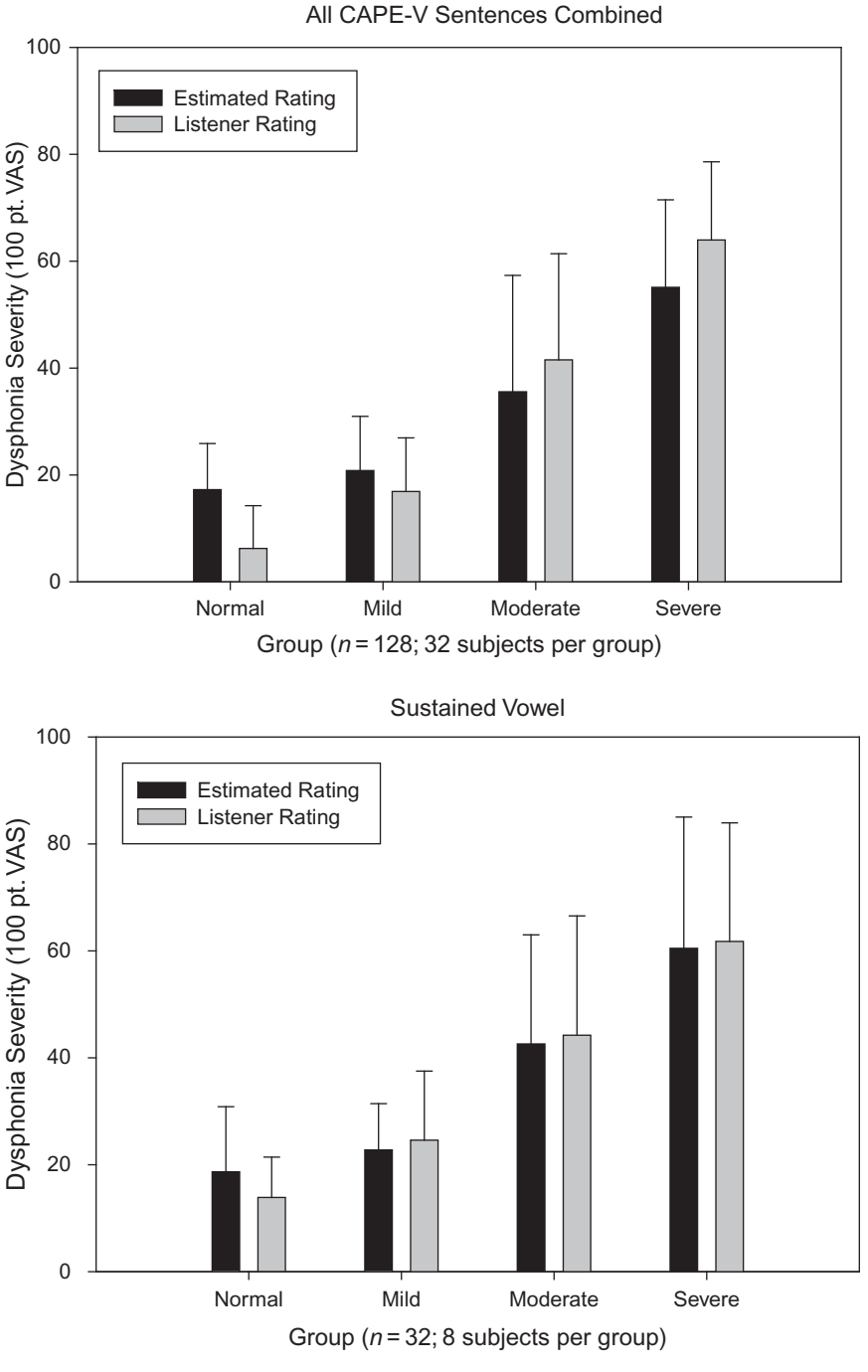

Figure 2. Mean acoustic severity estimates vs mean listener ratings per sample for male and female voices combined for all CAPE-V sentences combined and the sustained vowel /ɑ/.

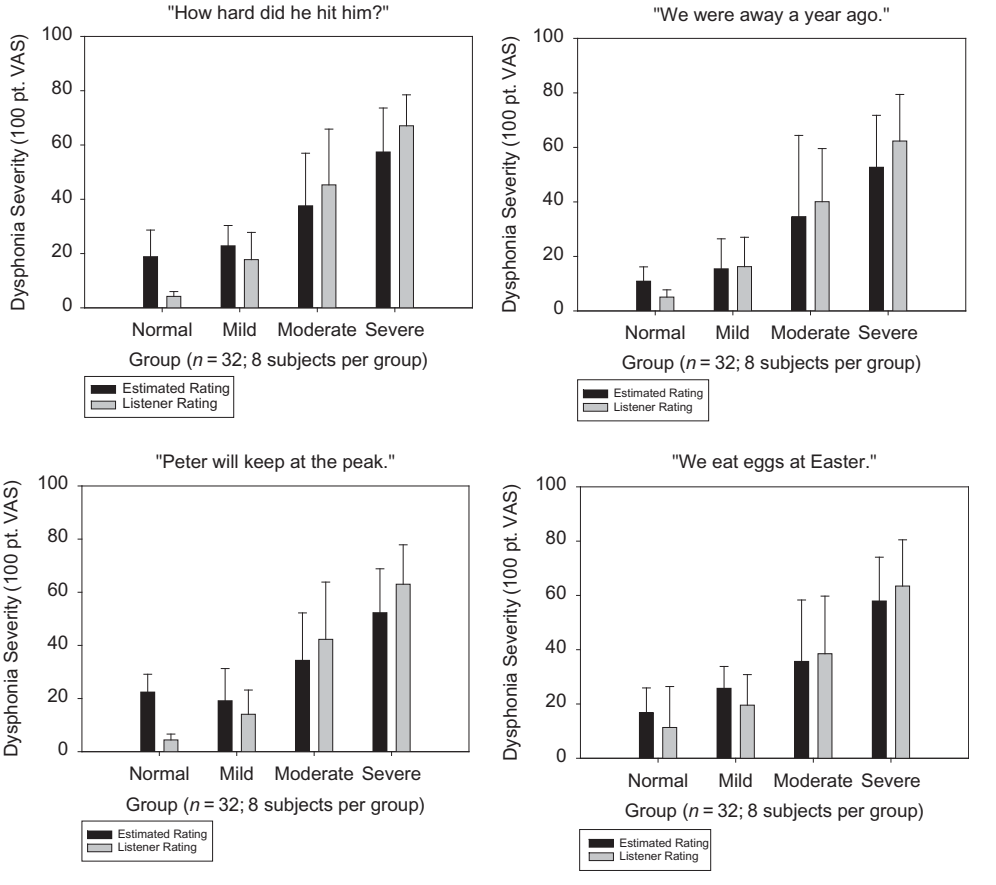

Figure 3. Mean acoustic severity estimates vs mean listener ratings per sample for male and female voices combined for individual CAPE-V sentences.

test results, with the LR+ results suggestive of the presence of voice dysfunction and the LR- scores suggestive of normality (Dollaghan, 2007). In addition, the accuracy of the test (i.e., how well the test separates the group being tested into those with and without the disease/disorder in question) may be measured via the area under the ROC curve. An area of 1 represents a perfect test; an area of .5 represents a worthless test. The area under the curve (AUC) of the ROC plot for all CAPE-V samples was 0.79, indicating respectable diagnostic precision (see Figure 4).

From previous analyses, it appeared that acoustically estimated severity ratings were particularly strong for sentence 3 ('We were away a year ago') and the sustained vowel /ɑ/. When the data from these two samples were combined, ROC analysis indicated that an acoustically estimated severity rating cut-off of 17.68 would result in sensitivity = 75% and specificity of 75% (LR+ = 3.00; LR− = 0.33). The area under the curve (AUC) of the ROC plot for Sentence 3 + Vowel /ɑ/ was .80 (see Figure 4). It should be noted that the sensitivity and specificity results must be tempered by the fact that the vocal status of the normal subjects was established by history and informal auditory screening of voice quality, and was not validated by laryngeal examination to rule out vocal fold pathology.

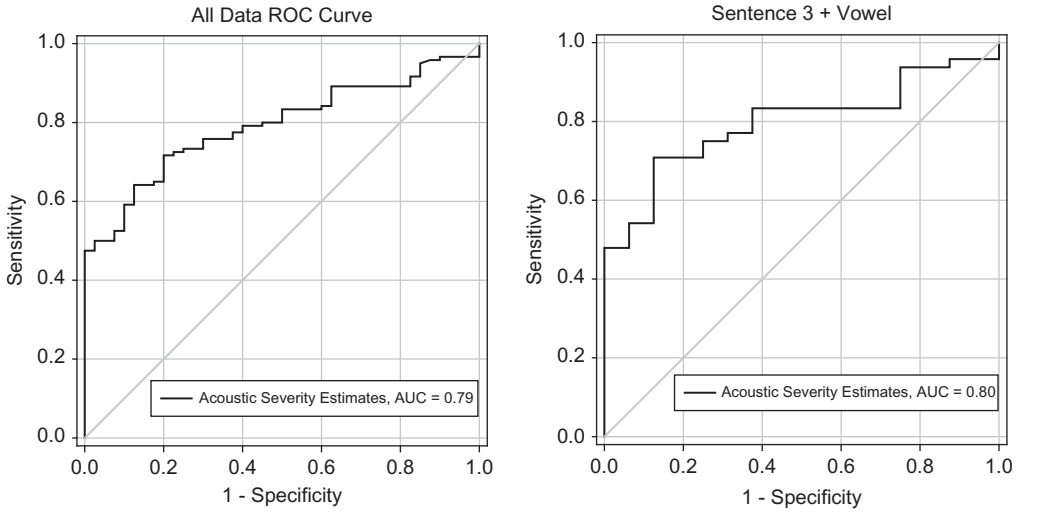

Figure 4. Receiver operating characteristic (ROC) curves using acoustic severity estimates from all combined CAPE-V samples, as well as for the CAPE-V sentence 3 ('We were away a year ago') and vowel /ɑ/ combined.

Discussion

The results of this study indicate that a strong relationship between perceived dysphonia severity and acoustic estimates of dysphonia severity can be achieved using a multivariable acoustic model incorporating both spectral and cepstral measures. This acoustic model related well to perceived dysphonia ratings for the types of sustained vowel and continuous speech samples elicited via the CAPE-V. Moreover, the addition of acoustic analyses of continuous speech samples adds considerably to the ecological validity of current voice assessment protocols since running/continuous speech samples are most representative of oral communication, and are highly relevant to the perception of dysphonia in everyday situations.

The use of a spectral/cepstral model permits estimation of dysphonia severity without the limitations associated with traditional time-based dysphonia measures such as jitter and shimmer. In the analysis of both sustained vowel and continuous speech samples, the cepstral peak prominence (CPP) was observed to be the strongest contributor to the multiple regression equations used to acoustically estimate dysphonia severity. However, the addition of other measures obtained via the same spectral/cepstral analysis procedures (L/H spectral ratio and the standard deviations of the mean CPP and mean L/H spectral ratio) significantly strengthened the predictions. Since all of these measures can be obtained efficiently via a common core of spectral/cepstral analysis procedures, it seems important to combine measures of the CPP with measures such as the CPP SD, and the L/H spectral ratio and L/H spectral ratio SD to strengthen dysphonia severity estimates across a wide range of dysphonia severities and types. However, based upon the results, it appears that these measures should be combined in different ways for sustained vowel vs continuous speech samples. In particular, the standard deviation variables (CPP SD and L/H spectral ratio SD) were observed to vary directly (i.e. increase in magnitude) with increased dysphonia severity in sustained vowels, but vary inversely (i.e., decrease in magnitude) with increased severity in continuous speech samples. Possible explanations for the increased CPP SD and L/H spectral ratio SD in

normal/near normal speech samples as compared to more severe levels of dysphonia may relate to the effect of the transitions from consonant to vowel and vowel to consonant productions in the continuous speech samples, as well as to the degree of F_0 variability observed in normal vs disordered voice production. In normal and near-normal voices, there was often a clear distinction and transition between relatively high energy, highly periodic vowel productions and surrounding weak amplitude semi-vowel productions (as observed in /w/ and /j/) or aperiodic or mixed aperiodic/periodic true consonant productions. These transitions may result in increased variability in the CPP and L/H spectral ratio and, therefore, increased CPP SD and L/H spectral ratio SD due to the presence of normal transitions from relatively aperiodic/unstable to periodic/stable (or vice versa) speech signal characteristics. In contrast, more severely dysphonic voices tend to be more consistently unstable, and, therefore, do not transition as markedly to or from quasi-periodic vowel production. The result is reduced variability in the CPP SD and L/H spectral ratio SD for more severely disordered voices. Secondly, it has been frequently reported that a tendency towards restricted F_0 variation and monopitch/monoloudness voice may occur in a variety of speech/voice disorders (e.g., Parkinson's disease; SLN paresis/paralysis, muscle tension dysphonia). In contrast to disordered voice, normal voice is characterized by free variation in the pitch and loudness of voice during normal prosodic alterations—this variation may also have a substantial effect on the relative amplitude of the CPP. Recent work by Awan, Giovinco, and Owens (2010) has demonstrated that increases in vocal loudness (often accompanied by increases in vocal pitch) result in significant increases in the CPP. Therefore, prosodic variations such as the production of stressed vs. unstressed syllabic patterns may be accompanied by corresponding increased variation in the CPP and, therefore, increased CPP standard deviation. The results of this study suggest that, in contrast to sustained vowel analysis, the effects of transitions from consonant to vowel or vice versa, as well as variations associated with normal prosodic patterns in speech, may represent valuable areas worth exploring when distinguishing normal vs. disordered voice.

Another difference in the analysis of sustained vowel vs. continuous speech samples observed in this study was the need to include Gender as a variable in the computed predictive multiple regression equation for the analysis of vowel samples. Although males and females did not differ significantly on any of the spectral/cepstral measures, observed differences between males and females in terms of CPP and L/H spectral ratio (males tended to have greater CPP and L/H spectral ratio values than females) may have combined to result in slightly different predictive equations for the genders. The tendency for differences between males vs females on the amplitude of the CPP and the magnitude of the low vs high frequency spectral ratio would appear to reflect males having relatively more spectral energy at the locations of the fundamental frequency and lower harmonics than do female speakers, and that this difference in spectral energy may be more prominent in sustained vowel productions than in running speech. In practice, these results suggest that any future computation of acoustically predicted dysphonia severity may simply require the examiner to indicate the gender of the speaker prior to analysis.

It is also important to acknowledge that the variable related to 'sentence type' was not a significant contributor to the stepwise multiple regression equation which estimated dysphonia severity during continuous speech. The removal of CPP values < 0 dB from the analyses may have removed much of the contribution of consonant productions (particularly unvoiced stop and fricative productions) to the elicited speech samples. These results suggest that a single predictive equation can be applied to a variety of connected speech samples, wherein strong predictions of perceived dysphonia severity were observed for all sentences using the

single multiple regression equation (*r*'s ranged between 0.77–0.86). Yet, the strongest prediction of dysphonia severity in speech was for the *all-voiced* CAPE-V sentence ('We were away a year ago'). Improved estimates of listener perceived severity (both in mean severity ratings and in overall correlations) may have been observed for this sentence because voiced vs voiceless/vowel vs true consonant segmentation is not as critical in the acoustic analysis of this particular sentence. While the transitions from true consonants to vowel and vice versa are areas where vocal control may be compromised in dysphonic states (as mentioned previously), the addition of vowel-to-consonant (VC) and consonant-to-vowel (CV) transitions to the analysis of dysphonia severity likely presents complications for both the listener and the acoustic algorithms employed. For the listener, the presence of true consonants in the speech/voice sample may potentially distract from a focused judgement of dysphonia severity. For the acoustic analyses, it is possible that noise from surrounding true consonants may be inadvertently added to the analyses, and thereby influence the computed severity estimates. This may be true particularly for sentences with relatively high frequency noise content (as observed in 'We eat eggs at Easter' and 'Peter will keep at the peak') vs relative low frequency noise observed in the glottal fricative productions in 'How hard did he hit him', in which a relatively strong correlation between acoustically predicted and listener severity was achieved ($r = 0.83$). In comparison, the all-voiced sentence represents a relatively easier context to analyse both acoustically and perceptually. Therefore, although the spectral/cepstral analysis for speech described in this study resulted in strong estimations of overall severity, future versions may investigate more stringent voiced/voiceless decision-making in the analysis of speech samples than the simple threshold method used in this study. The challenge will be to remove the central part of any unvoiced consonant production while retaining the transitions between consonant and vowel that may provide important information regarding the ability of the speaker to clearly differentiate between voiced and unvoiced productions. Perhaps cepstral analysis could be preceded by time-based F_0 tracking or spectrally guided voiced/voiceless decision-making, thereby allowing the cepstral analyses to focus primarily on the voiced portions of the speech sample under analysis. This is an area worthy of further investigation.

The acoustic algorithms described in this study tended to over-estimate the severity of normal and mild voice samples and under-estimate the severity of more severe samples. This tendency has been previously reported by Awan and Roy (2009) and Awan et al. (2009) and may be due in part to an 'end-effect' in listener ratings, with a tendency for perceived ratings to be either negatively skewed (in more severe cases) or positively skewed (in normal/close to normal samples). In contrast, acoustically estimated ratings have been reported to more closely approach normal distributions (Awan et al., 2009). However, it is clear that connected speech analysis is certainly more acoustically complicated, and thus contributes more error variance to any regression-based estimates of perceived dysphonia severity. The ability to effectively separate consonant noise components in both normal and disordered vowel segments in the acoustic analyses is a challenging prospect, and may be responsible, in part, for the discrepancies between acoustically estimated dysphonia severity and listener perceived severity ratings. As an example, it is possible that use of a 0 dB threshold for the analysis of cepstral peaks may have resulted in an omission of spectral/cepstral data that should have been included in the analyses of severe voice samples, and which would have resulted in increased acoustic severity estimates for the severe voice samples. Incorporating improved voice/voiceless detection algorithms which may be able to distinguish between 'intended' voicing (e.g., as may occur during an aphonic voice break) vs the purposeful voiceless segments of the speech sample, may result in improved acoustic estimation of severity in more severe dysphonic samples.

It is these types of difficulties with analysis of dysphonia in connected speech that likely led to the observation that the best estimates of listener perceived severity were still obtained with sustained vowel productions. This finding contradicts those of Heman-Ackah et al. (2003) and Halberstam (2004) and was somewhat surprising considering the view that continuous speech analysis may provide a more ecologically valid assessment of the patient's control of vocal parameters such as vocal quality. However, a key advantage of sustained vowel production is that it provides a focused environment for both acoustic analysis and listener judgements. It has been well documented that listeners may be distracted from focused judgements of characteristics such as dysphonia severity by other characteristics within the speech signal such as articulatory and intonation variations. As mentioned, these same characteristics may also result in variability in the spectral/cepstral estimates provided via acoustic analyses. In contrast, the steady pitch and loudness expected during a sustained vowel production allow both listener judgements and acoustic analyses to be arrived at in a relatively uncomplicated fashion, thus producing very strong acoustic predictions of listener perceived dysphonia severity. Our results indicate that assessment of sustained vowel productions remains a valuable voice context with excellent acoustic predictions of listener perceived severity.

In conclusion, the cepstral/spectral acoustic model presented here represents an important step forward in objective voice assessment. Because the model was sensitive to varying degrees of dysphonia severity in both connected speech and sustained vowel contexts, it offers promise as a means to objectively quantify dysphonia severity and potentially serve as a valid treatment outcomes measure. Future studies with larger samples of voice disorder types and severities are needed to further establish its clinical utility.

Acknowledgements

The authors would like to thank Dr P. Bohling (Bloomsburg University of PA) for contributions to the statistical analyses used in this study.

Declaration of interest: Dr S. N. Awan is currently working with KayPentax (Lincoln Park, NJ) on the development of commercial computer software including cepstral analysis of continuous speech algorithms. The authors report no further conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Awan S.N., Giovinco, A., & Owens, J. (2010). Effects of vocal intensity and vowel type on cepstral analysis of voice. Presented at the 39th Annual Symposium: *Care of the Professional Voice*, Philadelphia, PA.

Awan, S.N., & Roy, N. (2005). Acoustic prediction of voice type in adult females with functional dysphonia. *Journal of Voice*, *19*, 268–282.

Awan, S.N., & Roy, N. (2006). Toward the development of an objective index of dysphonia severity: A four-factor model. *Clinical Linguistics & Phonetics*, *20*, 35–49.

Awan, S.N., & Roy, N. (2009). Outcomes measurement in voice disorders: Application of an acoustic index of dysphonia severity. *Journal of Speech, Language, and Hearing Research*, 52, 482–499.

Awan, S.N., Roy, N., & Dromey, C. (2009). Estimating dysphonia severity in continuous speech: Application of a multiparameter spectral/cepstral model. *Clinical Linguistics & Phonetics*, *23*, 825–841.

Baken, R.J. (1987). *Clinical Measurement of Speech and Voice*. Boston, MA: Little, Brown and Co.

Callan, D.E., Kent, R.D., Roy, N., & Tasko, S.M. (1999). Self-organizing maps for the classification of normal and disordered female voices. *Journal of Speech and Hearing Research*, *42*, 355–366.

Dollaghan, C. (2007). Appraising diagnostic evidence. In C. Dollaghan (Ed.), *The Handbook for Evidence-based Practice in Communication Disorders*. Baltimore, MD: Brookes.

Eadie, T.L., & Doyle, P.C. (2005). Classification of dysphonic voice: acoustic and auditory-perceptual measures. *Journal of Voice*, *19*, 1–14.

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed; pp 124–139). New York: Harper & Row.

Halberstam, B. (2004). Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL*, 70–73.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica*, *90*, 441–451.

Hartl, D., Hans, S., Vaissiere, J., Riquet, M., & Brasnu, D. (2001). Objective voice quality analysis before and after onset of unilateral vocal fold paralysis. *Journal of Voice*, *15*, 351–361.

Heman-Ackah, Y., Heuer, R.J., Michael, D.D., Ostrowski, R., Horman, M., Baroody, M.M., et al. (2003). Cepstral peak prominence: A more reliable measure of dysphonia. *Acta Otol Rhinol Laryngol*, *112*, 324–333.

Heman-Ackah, Y.D., Michael, D.D., & Goding, G.S. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice*, *16*, 20–27.

Hillenbrand, J., & Houde, R.A. (1996). Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, *39*, 298–310.

Hillenbrand, J., Cleveland, R.A., & Erickson, R.L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, *37*, 769–778.

Kempster, G.B., Gerratt, B.R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R.E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18, 124–132.

de Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of speech and Hearing Research*, 38, 794–811.

Laflen, J.B., Lazarus, C.L., & Amin, M.R. (2008). Pitch deviation analysis of pathological voice in connected speech. *Annals of Otology, Rhinology and Laryngology*, *117*, 90–97.

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (in press). Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice*.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: A meta-analysis. *J. Acoust. Soc. Am.*, 126.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46.

Noll, A.M. (1964). Short-term spectrum and "cepstrum" techniques for vocal pitch detection. *Journal of the Acoustical Society of America*, *41*, 293–309.

Parsa, V., & Jamieson, D.G. (2001). Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research*, *44*, 327–339.

Qi, Y., Hillman, R.E., & Milstein, C. (1999). The estimation of signal to noise ratio in continuous speech for disordered voices. *Journal of the Acoustical Society of America*, *105*, 2532–2535.

Roy, N., Gouse, M., Mauszycki, S.C., Merrill, R.M., & Smith, M.E. (2005). Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *Laryngoscope*, 115, 311–316.

Shrivastav, R., Sapienza, C.M., & Nandur, V. (2005) Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*, 48, 323–335.

Wolfe, V., & Steinfatt, T.M. (1987). Prediction of vocal severity within and across voice types. *Journal of Speech and Hearing Research*, *30*, 230–240.

Wolfe, V.I., Martin, D.P., & Palmer, C.I. (2000). Perception of dysphonic voice quality by naïve listeners. *Journal of Speech and Hearing Research*, 43, 697–705.

Yiu, E., Worrall, L., Longland, J., & Mitchell, C. (2000). Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding. *Clinical Linguistics & Phonetics*, *14*, 295–305.

Zhang, Y., & Jiang, J.J. (2008). Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of Voice*, *22*, 1–9.

Zweig M.H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39, 561–577.