The Perception of Breathiness in the Voices of Pediatric Speakers

***Lisa M. Kopf, †Mark D. Skowronski, ‡Supraja Anand, ‡David A. Eddins, and §Rahul Shrivastav,** **Cedar Falls, Iowa, †Boca Raton and ‡Tampa, Florida, and §Athens, Georgia*

Summary: Background. The perception of pediatric voice quality has been investigated using clinical protocols developed for adult voices and acoustic analyses designed to identify important physical parameters associated with normal and dysphonic pediatric voices. Laboratory investigations of adult dysphonia have included sophisticated methods, including a psychoacoustic approach that involves a single-variable matching task (SVMT), characterized by high inter- and intra-listener reliability, and analyses that include bio-inspired models of auditory perception that have provided valuable information regarding adult voice quality.

Objectives. To establish the utility of a psychoacoustic approach to the investigation of voice quality perception in the context of pediatric voices?

Methods. Six listeners judged the breathiness of 20 synthetic vowel stimuli using an SVMT. To support comparisons with previous data, stimuli were modeled after four pediatric speakers and synthesized using Klatt with five parameter settings that influence the perception of breathiness. The population average breathiness judgments were modeled with acoustic measures of loudness ratio, pitch strength, and cepstral peak.

Results. Listeners reliably judged the perceived breathiness of pediatric voices, as with previous investigations of breathiness in adult dysphonic voices. Breathiness judgments were accurately modeled by loudness ratio ($r^2 = 0.93$), pitch strength ($r^2 = 0.91$), and cepstral peak ($r^2 = 0.82$). Model accuracy was not affected significantly by including stimulus fundamental frequency and was slightly higher for pediatric than for adult voices.

Conclusions. The SVMT proved robust for pediatric voices spanning a wide range of breathiness. The data indicate that this is a promising approach for future investigation of pediatric voice quality.

Key Words: Listener perception–Breathiness–Matching task–Pediatric dysphonia.

INTRODUCTION

Abnormal voice quality is often the first sign of an underlying voice disorder. As such, formal evaluation of voice quality is an essential component of voice diagnostic evaluations and frequently contributes to critical treatment outcome measures. Approximately in children 6%–9% have a voice disorder or develop a voice disorder that ranges from mild to severe dysphonia.[1,2] Untreated, such disorders can lead to a variety of complications that may negatively impact speech intelligibility, conveyance of emotion, expression of personality, and the ability and willingness to communicate effectively. Together, these can influence a child's education, quality of life, well-being, and can have potential long-lasting effects into adulthood. For example, children with voice disorders may have difficulties in activities requiring a loud voice (eg, playground) or avoid participation in class activities (eg, providing a verbal answer in classroom) because of feelings of inferiority.[3–6] Further, prior research on listener attitudes indicate that such deviant vocal behaviors cause children with voice disorders to be perceived as withdrawn, less confident, less emotionally stable, and less intelligent.[7,8] Given these this negative stereotyping toward children with voice disorders, diagnostic accuracy and the ability to quantify voice treatment outcomes are of significance to this population.

Extensive research has focused on establishing objective measures of dysphonic voice quality that reliably and accurately capture the perception of human listeners.[9–13] Although many studies have assumed that voice quality measures designed for adult speakers are valid and accurate for pediatric voices,[14–16] only a few have directly evaluated this claim.[17–21] Methods used in the perceptual and acoustic evaluation of voice quality in pediatric patients are essentially identical to those used with adults despite the fact that there are marked differences between children and adults in anatomy and physiology that impact voice production.[22] Ideally, voice quality evaluation for pediatric patients would use measurement procedures and metrics that have been shown to provide accurate and reliable indices of voice quality perception for this population. Among the few studies that have developed evaluation metrics for pediatric patients, Campisi et al[17] and Maturo et al[20] developed normative databases of acoustic measures based on children aged 4–18 years. These researchers extracted several widely used acoustic variables (eg, frequency, perturbation) from the commercial *Multi-Dimensional Voice Program* (KayPentax, Montvale, NJ, USA). Each measure was significantly different for children at the younger end of the continuum and approached adult-like values with increasing age, particularly following puberty. The set of normative acoustic measures for pediatric voice was expanded to include cepstral measures (eg, cepstral peak prominence[21]). However, to our knowledge, only one study has directly compared pediatric and adult voices (both groups having vocal

Accepted for publication September 28, 2017.

From the *Department of Communication Sciences and Disorders, University of Northern Iowa, Cedar Falls, Iowa; †AventuSoft LLC, Boca Raton, Florida; ‡Department of Communication Sciences and Disorders, University of South Florida, Tampa, Florida; and the §Office of the Vice President for Instruction, University of Georgia, Athens, Georgia.

Address correspondence and reprint requests to David A. Eddins, Department of Communication Sciences and Disorders, University of South Florida, 4202 Fowler Avenue, PCD 1017, Tampa, FL 33620. E-mail: deddins@usf.edu

Journal of Voice, Vol. 33, No. 2, pp. 204–213

0892-1997

https://doi.org/10.1016/j.jvoice.2017.09.024

nodules) using conventional perceptual and acoustic measures.[18] In the current study, we consider whether a set of measurement and analysis methods successfully used to characterize breathiness in adult patients is practical and feasible for use with pediatric voices. Based on research such as that of Masaki[18] and Lopes et al,[15] we make an implicit assumption that, much like adult voices, dysphonic voices in pediatric patients may be described in a multidimensional perceptual space, with a continuum along "breathiness" being one of the dominant dimensions.

Comparing voice outcome measures in children and adults

There have been a number of investigations of dysphonia in pediatric patients, although only a few have investigated the potential to generalize outcome measures, such as questionnaires and subjective rating scales used with adults to the pediatric population. Quality of life indices are commonly used with adult patients (eg, Voice Related Quality of Life [VRQoL][23]), and several quality of life instruments have been developed or modified for use with children.[24–27] One example is the pediatric Voice Outcome Survey developed by Hartnick and colleagues. This is a brief, simple, parent-proxy quality of life tool that is convenient for clinical use. Similarly, the Pediatric Voice-Related Quality of Life (Pediatric VRQoL) was developed on the basis of the commonly used 10-item VRQoL questionnaire, with modifications to the wording suitable for parent-proxy reporting.[26] Finally, the Voice Handicap Index (VHI), developed for use in adult populations, was modified by Zur et al[27] to extend its use to the pediatric population, resulting in the pediatric Voice Handicap Index (pVHI[27]). Unlike these quality of life indices, in which adult-focused versions were modified for use with pediatric populations, perceptual and acoustic evaluations of pediatric voice quality largely consist of the same instruments used with adults. For example, the adult version of the The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)[28,29] has been used extensively with children without any modifications to the scales.[30,31] To evaluate whether such a practice is valid, Johnson et al[32] demonstrated that the relationship between CAPE-V scores and the VHI scale, previously established for adults, also holds for the CAPE-V and pVHI in a pediatric population,[32] providing some support for the use of the adult-based CAPE-V with the pediatric population (also see Zur et al[27]). However, the strength of such relationships decreased in pediatric patients with complex voice disorders such as children who have undergone airway reconstruction or other surgical procedures.[33]

A number of acoustic measures of dysphonic voice quality commonly used with adult speakers have also been used in pediatric speakers without any special modifications. Presumably, these measures inherit the same advantages and limitations in the pediatric population as seen with adult speakers. Such correlates of breathy voice quality may be broadly classified into the following groups[11,34]: measures of noise (eg, signal-to-noise ratio; harmonic-to-noise ratio[15,18,35–37]), measures of perturbation (eg, jitter, shimmer[15,36]), measures related to spectrum or cepstrum (eg, cepstral peak prominence[18]), and composite measures such as Acoustic Voice Quality Index.[16] Most of these acoustic measures take advantage of the quasi-periodic nature of the vocal acoustic signal, with the assumption that a departure from periodicity is indicative of greater severity of dysphonia. Unfortunately, this basic assumption works well only for normal voices and to some degree, to those with mild to moderate dysphonia. For the more severely dysphonic voices, this assumption fails for two major reasons. First, the vocal acoustic signals for voices with more severe dysphonia often do not demonstrate a clear fundamental frequency (f_0) and may instead show bifurcations or chaotic f_0 patterns and may be described as type 2 or type 3 voices.[38] Indeed, in one study, Kelchner et al[39] revealed that 20 of 21 voice signals recorded from children post-airway reconstruction were either type 2 or type 3.[39] Acoustic measures that rely on the estimation of signal f_0 will generally fail to accurately estimate the severity of dysphonia for such voices. A second reason is that there generally is a nonlinear relationship between acoustic changes and the perception of quality. For example, the same change in the level of noise (dB signal-to-noise ratio) applied to two different baseline levels of breathiness can have a very different impact on the perceived change in breathiness.[34] Among the studies that have shown correlations between perceptual and acoustic measures, two have considered pediatric voices.[15,16] Similar to adult voices, both reported moderate correlations.

Psychoacoustic approach to the study of voice quality

Although direct acoustic analysis is by far the most common approach in the clinical environment, such an approach often does not accurately predict clinician perception of voice quality. Isshiki et al[40] commented that "the human ear is most suitable, at least as a first step and at this stage of development of electronic instruments, for the purpose of differentiation of hoarseness." Despite the development of technology, from this, one can glean that computational metrics based on a signal that has undergone filtering and other transformations similar to those involved in processing sound by the human auditory system may provide a more accurate prediction of voice quality than estimates based strictly on the vocal acoustic signal.

To address the limitation of conventional schemes to objectively measure dysphonic voice quality, one may adopt a psychoacoustic approach to the study of voice quality perception. Such an approach involves four key principles: conceptualization with reference to principles governing sound quality in general; robust measurement methods with minimal bias; averaging across multiple listeners to reduce variance; and objective analyses that consider transformation of the acoustic signal by the auditory system as part of the perceptual process. In the case of breathy voice quality, the approach draws from research on the general auditory percept of tonality (tonal salience[41]). A psychophysical matching task that minimizes bias and results in ratio-level data with relevant physical units has been adopted and modified for the study of breathy voice quality.[42] To acoustically model the behavioral data, we have leveraged bio-inspired computational principles to account for transformations of the acoustic signal by the auditory system, which relates an acoustic stimulus to an estimate of the internal representation of that stimulus, producing a more accurate

model of the resulting perception revealed by behavioral measurement.[10,11,13,43,44] The principle differs from the more common approach that involves direct analysis of the acoustic signal to establish the relationship between acoustic properties (eg, jitter and shimmer) and perceptual judgments of the voices under study.

In the context of vocal breathiness in adult voices, Shrivastav and Sapienza[11] reported perceptual breathiness judgments, and showed that a single computational measure, partial loudness of harmonic energy (N'S_{IGNAL}), was able to account for a greater amount of variance in the perceptual judgments than any single acoustic measure among a large set of common acoustic measures. As described below, this approach seeks to estimate the internal (auditory system) representation of the spectrum of a waveform, following several stages of filtering and nonlinear transformation designed to loosely mimic transformations in the auditory system that lead to a representation of loudness as a function of frequency. In addition, Shrivastav and Camacho[45] demonstrated that η, the ratio of noise loudness to the partial loudness of a signal, was a better predictor of vocal breathiness than the cepstral peak prominence acoustic measure.[45] Although this partial loudness measure provided robust correlations with perceptual data, a weakness in its practical application is that computation of partial and noise loudness requires separation of the noise and harmonic energy, and this can be done with precision only using synthetic stimuli.

To overcome the requirement of synthetic speech, we have also considered an index known as pitch strength to evaluate dysphonia in natural and synthetic speech.[13,46] Borrowing from studies of the tonality conveyed by non-speech stimuli,[47] samples from dysphonic voices were evaluated in a series of perceptual studies and modeled using computational methods for pitch strength (ie, pitch salience from weak to strong). Shrivastav et al[13] reported a strong negative correlation (r = –0.989) between pitch strength judgments and loudness ratio, whereas Eddins et al[46] demonstrated that computational pitch strength estimates were strongly and negatively correlated with perceived breathiness. These studies show that pitch strength or pitch salience is low for breathy voices and pitch strength increases as the vocal breathiness decreases.

The current research attempts to adopt the psychoacoustic approach described above to measure the severity of dysphonia in pediatric voices. As a first step, we have developed a perceptual experiment using a set of pediatric voices that were analyzed and resynthesized to support investigation of the two bio-inspired algorithms described above (loudness ratio and pitch strength, along with cepstral peak). These objective measures were selected because these have been shown to be highly predictive of breathiness in adult voices.[45,46,48] The fidelity of the predictions of voice quality perception for pediatric voices from the current experiment was compared with analogous predictions of voice quality perception for adult voices from previous studies. In this initial study, we focus on breathiness because it is perhaps the most widely investigated and best understood of voice quality dimensions. Our ultimate goal is to establish reliable predictors of dysphonic voice quality that maintain a high level of accuracy across the lifespan.

METHODS

Listeners

A total of six female listeners (6F; mean age 20 years) consented to volunteer for the study following university institutional review board procedures. Listeners were students in the Department of Communicative Sciences and Disorders at Michigan State University who had taken at least one introductory course in speech-language pathology in the department. All listeners underwent a hearing screening to ensure that they had hearing within normal limits. The hearing screen consisted of otoscopy to ensure no ear canal blockage and a pure tone test at frequencies ranging from 0.125 to 8.0 kHz. All listeners completed the listening experiment.

Listening task

Breathiness judgments were obtained using a single-variable matching task,[42,49] in which listeners heard two stimuli in succession. The first stimulus, denoted the *standard*, was the voice to be evaluated. The second stimulus, denoted the *comparison*, was a noisy sawtooth waveform (described below) with a single variable of adjustment: noise-to-signal ratio (NSR) in dB. After the two stimuli were presented in succession, the listener indicated via button press in a software interface whether the comparison was more or less breathy than the standard stimulus. The presentation of the two stimuli and input of a single response constituted a single trial. Using an adaptive method of adjustment, the NSR of the comparison stimulus was either increased or decreased by 2 dB on the next trial to make the comparison stimulus more breathy or less breathy, respectively. Trials were presented until the listener indicated via button press that the comparison stimulus matched the standard stimulus in terms of perceived breathiness. Listeners were encouraged to explore a range of comparison values around the perceived matching value before indicating a match to ensure that the closest match was chosen.

Breathiness matching judgments for a given listener and a given standard stimulus were based on six blocks of trials. For three blocks of trials, the procedure began with the independent variable set to a high initial value (ie, high NSR of −5 dB) at the beginning of each block of trials. Each block of trials continued until the listener indicated that the perceived breathiness of the comparison matched that of the standard. For the other three blocks of trials, the procedure began with a low independent variable value (ie, low NSR of −30 dB). The final breathiness matching judgment was based on an average across the six blocks. Multiple blocks were used to minimize the impact of random errors resulting from changes in listener attention, fatigue, or other factors.[50] Thus, the matching task was completed in 120 blocks (three blocks beginning with a high IV, three blocks beginning with a low IV, and 20 standard voice tokens). The order of standard voice tokens was randomized across listeners, and the order of high and low IV for each standard voice token was randomized. Replicates for each standard voice token at each IV were tested consecutively.

Before the listening task, all listeners underwent a short practice session to familiarize themselves with the task. A set of three

adult voices, representing a range of breathiness, were used in the practice task. Each listener was provided feedback on the task itself but not on correctness of a response or responses.

During testing, listeners were seated in a single-walled sound booth in front of a computer monitor and mouse. Stimulus generation, presentation, response collection, and adaptive tracking were controlled by the TDT SykofizX software application (Tucker-Davis Technologies, Inc., Alachua, FL). Stimuli were delivered to the right ear of each listener via the TDT RZ6 Multi I/O processor and an Etymotic ER-2 ear insert (Etymotic, Inc., Elk Grove Village, IL). The stimulus presentation level was 75 dB sound pressure level. Listeners used a computer mouse to make their selections. Each listener participated in two to three sessions of no more than 2 hours each. Participants were given a short break every 15 minutes and more often if requested.

Standard stimuli

A total of four /a/ vowels from pediatric speakers were chosen from the University of Florida Child Voice Database.[51] All children were between the ages of 3.0 and 4.0 years. Children's voices that had a higher average f_0 than those used in similar experiments evaluating breathiness in adult voices were chosen.[46,52] The highest average f_0 value in the previous experiments was 219.4 Hz for a female speaker.

Based on an initial acoustic analysis, the four speakers' voices were synthesized using a Klatt synthesizer with the Liljencrants-Fant model for the sound source.[53] Synthesis parameters can be seen in Table 1. All synthetic vowels were created with a sampling rate of 20 kHz and were up-sampled to 24,414 kHz, an acceptable sampling rate for the hardware used in the experiment. The voices represented a range of fundamental frequencies (291–373 Hz). The aspiration noise (AH) and open quotient (OQ) parameters, which are correlated with breathiness perception, were manipulated in a similar way to Shrivastav et al.[52] Briefly, the value of these parameters was varied to create a series of five equal steps ranging from little perceptible breathiness to the maximum breathiness that could be output using the synthesizer (Table 1). Therefore, a total of 20 stimuli were created (four speakers × five AH:OQ combinations). Note that for one speaker, Caro, the difference between AO1 and AO2 was smaller than the other steps because of synthesis error that was not discovered until the experiment was completed.

In addition to the stimuli used in the experiment as described above, two additional sets of speech waveforms were synthesized to allow for the calculation of loudness ratio following the approach of Shrivastav et al.[52] To establish the harmonic signal required for the loudness ratio calculations, one set of waveforms was synthesized using the parameters as stated above while setting the AH values to zero, effectively removing the "aspiration noise." To establish the noise signal required for the loudness ratio calculations, a second set of waveforms was synthesized using the parameters above while setting the amplitude of voicing value to zero so that only the AH was present.

Comparison stimuli

Comparison stimuli were identical to those described by Patel et al.[42] Specifically, the stimulus consisted of a sawtooth waveform (f_0 = 151 Hz) mixed with a Gaussian noise that was filtered with a first-order low-pass filter (cutoff frequency = 151 Hz) to match the −6 dB/octave roll off of the sawtooth waveform. The sawtooth and noise waveforms were then both filtered with a first-order low-pass filter (cutoff frequency = 151 Hz) to better approximate the long-term average spectral slope of speech. Both the sawtooth and the noise spectra deceased by 12 dB/octave after filtering. The choice of parameters was based on an acoustic

TABLE 1.
Synthesis Parameters for the Four Pediatric Voices

Speaker	Caro	Kade	Kath	Math
Fundamental frequency in Hz (f_0)	372.6	330.9	328	290.6
Amplitude of voicing in dB (AV)	60	60	60	60
Aspiration range in dB (AH)	60–80	55–80	55–80	55–80
Open quotient range in % (OQ)	77–99	71–99	71–99	71–99
Speed quotient (SQ)	300	350	400	300
Gain in dB (GN)	60	65	65	60
First formant (F1)	1080	1088	981	1097
First formant bandwidth (B1)	300	300	246	204
Second formant (F2)	1830	1401	1415	1793
Second formant bandwidth (B2)	414	333	275	217
Third formant (F3)	2425	2218	1750	2662
Third formant bandwidth (B3)	450	339	300	300
Fourth formant (F4)	3896	4002	3825	4264
Fourth formant bandwidth (B4)	700	512	350	350
Fifth formant (F5)	4766	4504	4441	4623
Fifth formant bandwidth (B5)	850	600	500	500
Sixth formant (F6)	4990	4990	4990	4990
Sixth formant bandwidth (B6)	1000	1000	1000	1000

Included are the names of the adjusted parameters with their synthesizer abbreviations in parentheses.

analysis of a large set of dysphonic voices from the KayPENTAX Disordered Voice Database (PENTAX of America, Inc., Montvale, NJ). The NSR was varied in the psychophysical task described above to achieve matches to the perceived breathiness of a wide range of dysphonic voices.

Acoustic analysis

Loudness ratio, cepstral peak, and pitch strength were calculated for each of the standard synthesized voices, and all acoustic measures were compared with the perceptual data using linear regression. Pitch height was previously shown to covary with breathy voice quality,[52] so the f_0 of each synthetic voice sample was included in the regression models. All acoustic measures and models were created using MATLAB scripts (The MathWorks, Natick, MA, version R2016a). Following description of the analysis methods below, an illustration of the relationship between each of the three acoustic measures and two voice samples is provided in Figure 1.

Loudness ratio was calculated as follows: each synthetic voice sample (created as separate periodic and aperiodic WAV files) was scaled such that the combined periodic and aperiodic signals were at 75 dB SPL. Next, the periodic and aperiodic components were converted to power spectral density functions using overlapping 50-ms Hann windows at 100 frames per second with 2048-point fast fourier transforms. The power spectral densities were scaled by outer and middle ear transfer functions[54] and analyzed with a perceptually motivated rounded exponential (roex) filter bank with filter center frequencies equally spaced by 0.1 equivalent rectangular bandwidth rate between 40 Hz and 15 kHz. The outputs of the filter bank for the periodic and aperiodic components—the excitation function (power units)—were converted to specific loudness (Sones/ERB rate), which accounted for the masking of each excitation function on the other and for the nonlinear compression of excitation.[55] Specific loudness was integrated over ERB rate to produce loudness (Sones). The noise loudness for the aperiodic component and periodic loudness for the periodic component were calculated as the mean loudness over all analysis frames. Finally, loudness ratio = noise loudness/periodic loudness. For modeling, loudness ratio was converted to dB: $\text{loudness ratio}_{\text{dB}} = 10\log_{10}(\text{loudness ratio})$.

Pitch strength was calculated using the Auditory SWIPE-prime algorithm (Aud-SWIPE′)[56] as follows: the combined periodic and aperiodic components of each synthetic voice sample were filtered by an outer and middle ear filter to flatten the spectral envelope and analyzed with a perceptually motivated filter bank. The output of each channel was half-wave rectified to approximate inner hair cell rectification. Each rectified channel signal was converted to a spectral magnitude, square root compressed, and summed across channels to approximate a specific loudness function. The frame size for FFT analysis was approximately eight fundamental periods of each pitch candidate value (50% overlap between adjacent frames), and pitch candidates were spaced between 80 and 400 Hz at a rate of 48 pitch candidates per octave. The specific loudness function was correlated with a sawtooth waveform specific loudness function for each pitch candidate, and the pitch candidate with the highest correlation (normalized between 0 and 1) was determined to be the pitch of the analysis frame. The correlation value was the pitch strength of the analysis frame. Pitch strength over all analysis frames was averaged and used in the perceptual models.

The cepstral peak was calculated using interpolation[57] as follows: the combined periodic and aperiodic components of each synthetic voice sample were analyzed using overlapping 50-ms Hann windows at 100 frames per second with 2048-point FFTs. The log spectral magnitude of each analysis frame was limited to 100 dB dynamic range to remove spectral nulls. The log spectrum was zero-padded by a factor of 8 and transformed via inverse FFT to the cepstral domain (zero padding in the log spectral domain produces interpolation of the cepstral peak in the cepstral domain), and the cepstral peak was detected within the f_0 range of 80–400 Hz. The cepstral peak of each frame was converted to dB for modeling,[38] and the mean

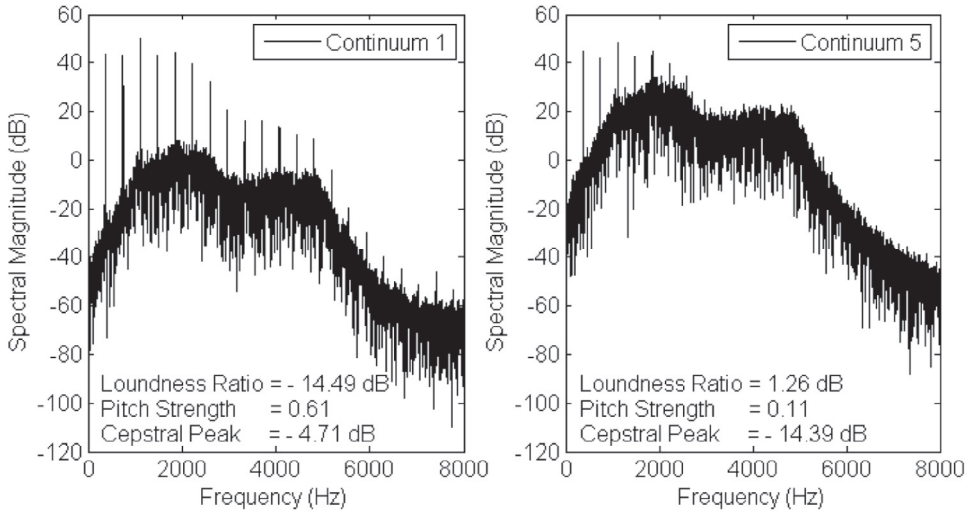

FIGURE 1. Fourier spectra of AH-OQ for two synthetic talkers. Both talkers are based on the same natural voice and represent the two end points of the breathy continuum (continuum positions 1 and 5), where position 1 represents a voice with a low degree of breathiness and position 5 represents a voice with high degree of breathiness. Values in the inset of each panel highlight the acoustic indices described in the text.

cepstral peak value over all frames was calculated and used in the perceptual models.

Figure 1 illustrates the relationship between the three analysis methods described above and actual stimuli used in the current experiment. This figure shows the magnitude spectrum of two of the 20 voice samples used in this study. These two samples are derived from the same natural talker but differ in that the voice shown in the left panel was synthesized to be at the low end of the breathy continuum, whereas the voice shown in the right panel was synthesized to be at the high end of the breathy continuum. Absolute values of the loudness ratio, pitch strength, and cepstral peak estimates based on the methods described above are shown in the inset of each panel. For these two example stimuli, it is evident that pitch strength and cepstral peak are lower and loudness ratio is higher for the stimulus on the right (perceived to be more breathy) than the stimulus on the left (perceived to be less breathy).

RESULTS

Effect of AH and OQ on matching thresholds for breathiness

Perceptual estimates of breathiness (mean ± standard error over listeners) are shown in Figure 2, with breathiness matching judgments on the ordinate and the AH:OQ continuum on the abscissa. Each symbol represents a different talker. Breathiness matching judgments are expressed as the NSR of comparison stimulus that matched the voice token. In this way, larger absolute values indicate greater perceived breathiness. The influence of increasing AH and OQ can be seen by considering the effect of continuing position, from left to right across the abscissa, corresponding to progressive increases in perceived breathiness as indicated by the matching judgments. A repeated measures analysis of variance with two within-subject variables (AH:OQ value, talker) revealed that AH:OQ value was a significant factor ($F_{4,20} = 24.7$, $p_{GG} = 0.0008$ is the P value with Greenhouse-Geisser correction for sphericity) while talker was not significant ($F_{3,15} = 0.15$, $p_{GG} = 0.83$) nor was the interaction of AH:OQ level and talker ($F_{12,60} = 1.22$, $p_{GG} = 0.34$). The lack of a difference among talkers is consistent with the interpretation that f_0 had little effect on breathy perception for this set of stimuli.

Rater reliability

Rater reliability was assessed via intraclass correlation (ICC).[58] To examine consistency within listeners, intra-ICC(2,k) was computed where k indicates three replicates (responses averaged from three high initial values and three low initial values). The range of ICC(2,k) values was 0.88–0.99 and when averaged across listeners was 0.95 ± 0.037 (standard deviation). To gauge consistency across listeners, inter-ICC(2,k) was computed where k indicates six listeners. The resulting ICC(2,k) value was 0.68. This analysis also indicates variability associated with different factors, including the stimulus ($\sigma_s^2 = 12.27$), listeners ($\sigma_r^2 = 22.51$), interactions ($\sigma_i^2 = 8.52$), and an error term ($\sigma_e^2 = 3.17$). The low value of the error term indicates that the ICC model accounted for nearly all of the variance of the dependent variable. The high intra-ICC values indicate high repeatability for each rater, and the moderately high inter-ICC value indicates moderately high agreement among the listeners.

Relationship between loudness ratio and breathiness matching values

Following the methods described above, the loudness ratio was computed for each voice sample and is shown by the black symbols in Figure 3 with perceptual breathiness matching values (dB NSR) on the ordinate and loudness ratio on the abscissa. Error bars indicate standard error over all listeners. In general, as the value of loudness ratio increases, the value of breathiness approaches zero. This relationship was assessed for pediatric

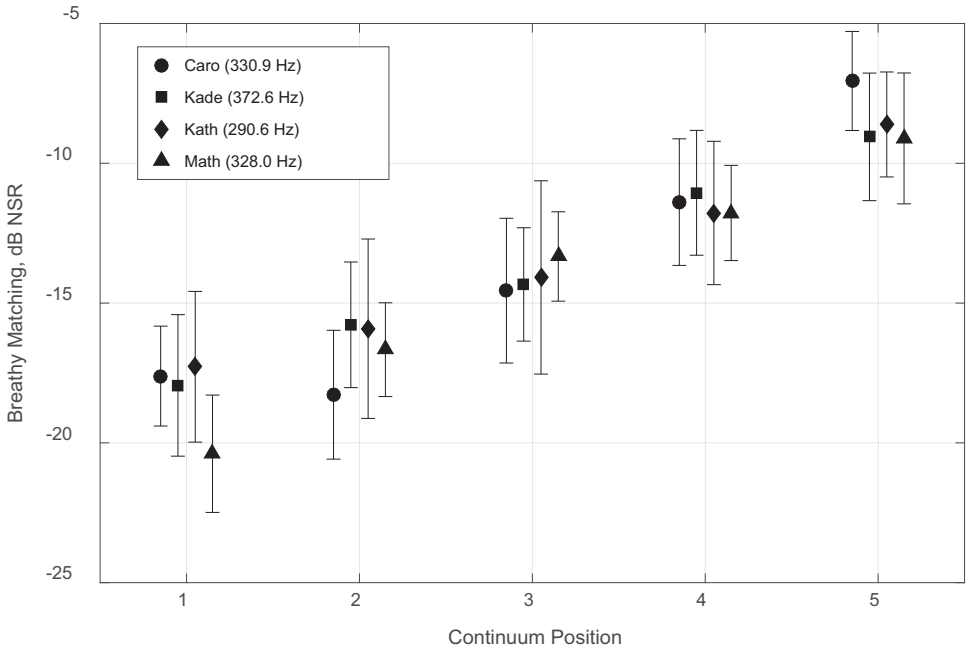

FIGURE 2. Matching NSR by continuum position. The mean ± SE are given, averaged over all listeners.

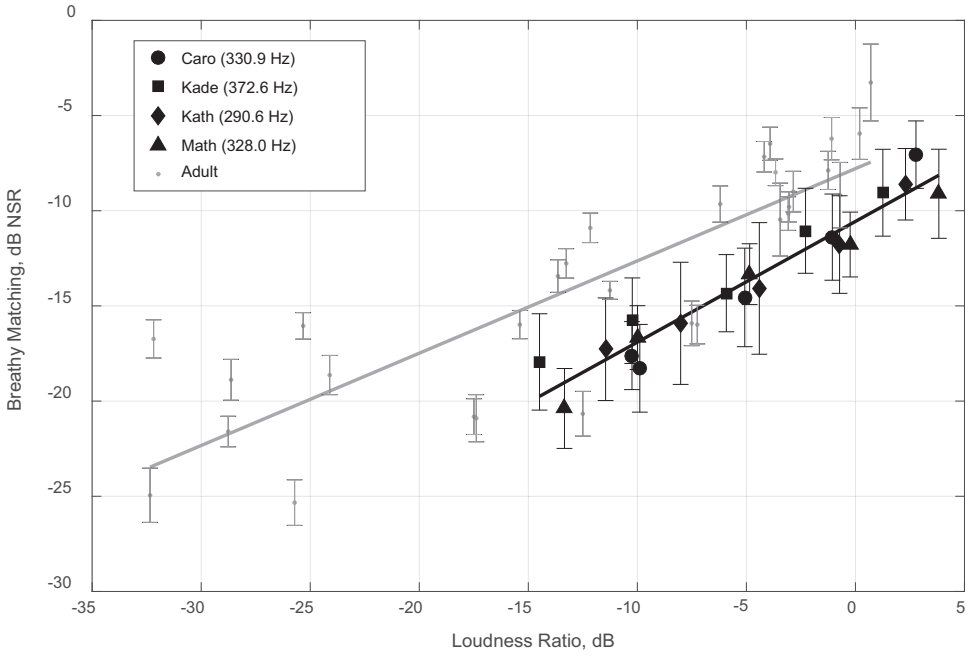

FIGURE 3. Matching NSR by loudness ratio in dB. The mean ± SE are given, averaged over all listeners. Black symbols are for pediatric voices, with symbol type separating the four different talkers. Gray symbols correspond to adult voices from Shrivastav et al.[52] Lines with the same shading scheme represent linear regression. In both cases, goodness of fit was high (pediatric voices: $r^2 = 0.93$; adult voices: $r^2 = 0.72$).

voices via linear regression, yielding a function $y = 0.64^* \times +0.01$, $r^2 = 0.93$. Including f_0 in the model did not affect the goodness of fit, further supporting the conclusion that variations in f_0 had little to no effect on breathiness perception for pediatric voices. For comparison, breathiness matching values for adult voices and a different set of listeners are shown in Figure 3 as gray symbols.[52] The relationship between loudness ratio and breathiness matching values for adult voices was similar to those for pediatric voices, although the range of matching values and the range of loudness ratios was larger for adult than for pediatric voices. A linear regression yielded the following model: $y = 0.49^* \times -7.8$, $r^2 = 0.72$. In this case, adding f_0 (Hz) improved the model accuracy: $y = 0.46^* \times -0.052^* f_0 - 0.62$, $r^2 = 0.88$. It should be noted that f_0 values spanned a much wider range for adult than for pediatric talkers. Furthermore, the model for adult voices indicates that an *increase* in f_0 corresponds to a *decrease* in breathy judgments, consistent with the observation that the pediatric breathy judgments in Figure 3 were *lower* than those for adult voices.

Relationship between pitch strength and breathiness matching values

To evaluate the potential relationship between breathiness matching values and an estimate of pitch strength (a surrogate measure of the tonality sound quality[47]), each standard stimulus was processed by Aud-SWIPE′ to estimate pitch strength as shown in Figure 4. The data for the pediatric voices of the current study are depicted by black symbols with breathiness matching value (dB NSR) on the ordinate and pitch strength estimate on the abscissa. The pitch strength estimates for this set of voice samples ranged from approximately 0.1 (very low pitch strength) to about 0.65 (moderately high pitch strength). Lower pitch strength estimates correspond to high breathiness values and as the pitch strength estimates increase, the perceived breathiness decreases. This inverse relationship is well-characterized by a linear function: $y = -19.0^* \times -7.2$, $r^2 = 0.91$. Importantly, adding f_0 to the regression model had no substantial effect on goodness of fit. For comparison, perceptual judgments for the same adult voices as shown in Figure 3 are displayed here by gray symbols along with corresponding pitch strength estimates. Again, the data are well characterized by a linear function: $y = -27.7^* \times 0.46$, $r^2 = 0.71$. Interestingly, for adult voices, including f_0 into the regression model improved the prediction: $y = -25.9^* \times 0.050^* f_0 + 6.0$, $r^2 = 0.85$.

Relationship between cepstral peak and breathiness matching values

As noted earlier, numerous studies have indicated a relationship between values of the cepstral peak and overall dysphonic severity as well as perceived breathiness.[48,59,60] For the current pediatric data, Figure 5 shows a similar relationship in which higher breathiness matching values corresponded to lower cepstral peak values. This relationship was well characterized by a linear function: $y = -0.96^* \times 23.8$, $r^2 = 0.82$. Cepstral peak values for the adult data from Shrivastav et al[52] also showed a linear relationship: $y = -1.22^* \times 26.2$, $r^2 = 0.91$. In neither case did adding f_0 into the respective linear regression models improve model accuracy.

DISCUSSION

This investigation was designed with several specific goals in mind. Chief among them was to establish whether the perceived breathiness of pediatric voices could be reliably measured using a single-variable matching task (SVMT) and whether that perception would follow the same general trends as breathiness

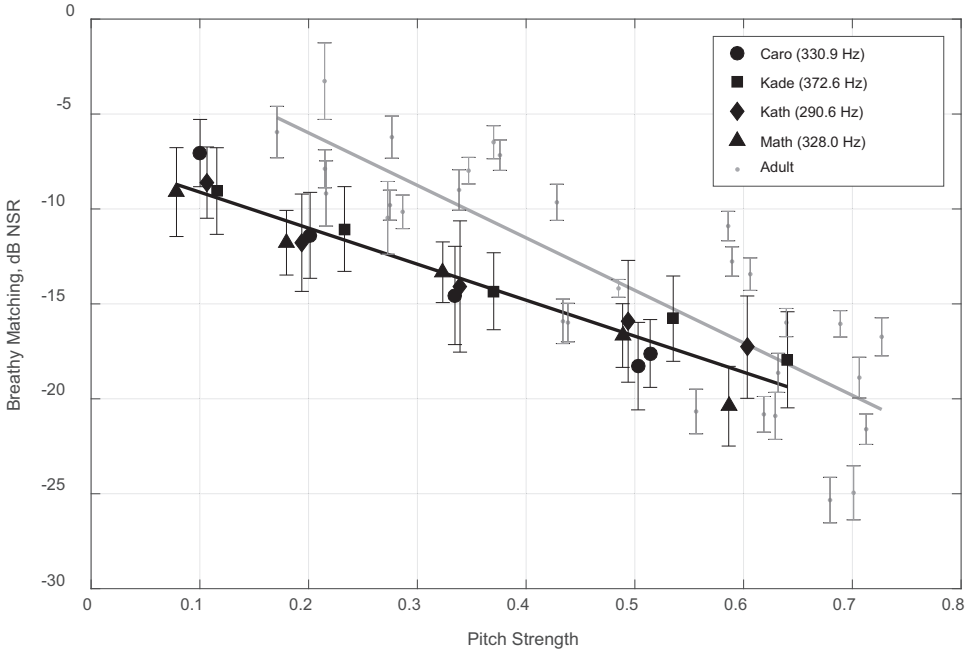

FIGURE 4. Matching NSR by pitch strength. The mean ± SE are given, averaged over all listeners. Black symbols are for pediatric voices, with symbol type separating the four different talkers. Gray symbols correspond to adult voices from Shrivastav et al.[52] Lines with the same shading scheme represent linear regression. In both cases, goodness of fit was high (pediatric voices: $r^2 = 0.91$; adult voices: $r^2 = 0.71$).

judgments for adult voices.[42,49] Although rating scales generally result in poor reliability between and within raters,[61] the interclass correlation values in the current study indicated moderately-high inter- and intra-listener reliability, analogous to data previously reported for adult dysphonic voices. Likewise, the range of breathiness matching values for the pediatric voices evaluated here was similar to, but slightly lower than, the range of breathiness matching values for the adult voices evaluated by Shrivastav et al.[52] Of course, the actual ranges are dependent
somewhat on the stimulus selection process as well as the choices of AH and OQ values chosen during stimulus generation. To determine actual population differences, random sampling and much larger set sizes would be required. Nevertheless, it is

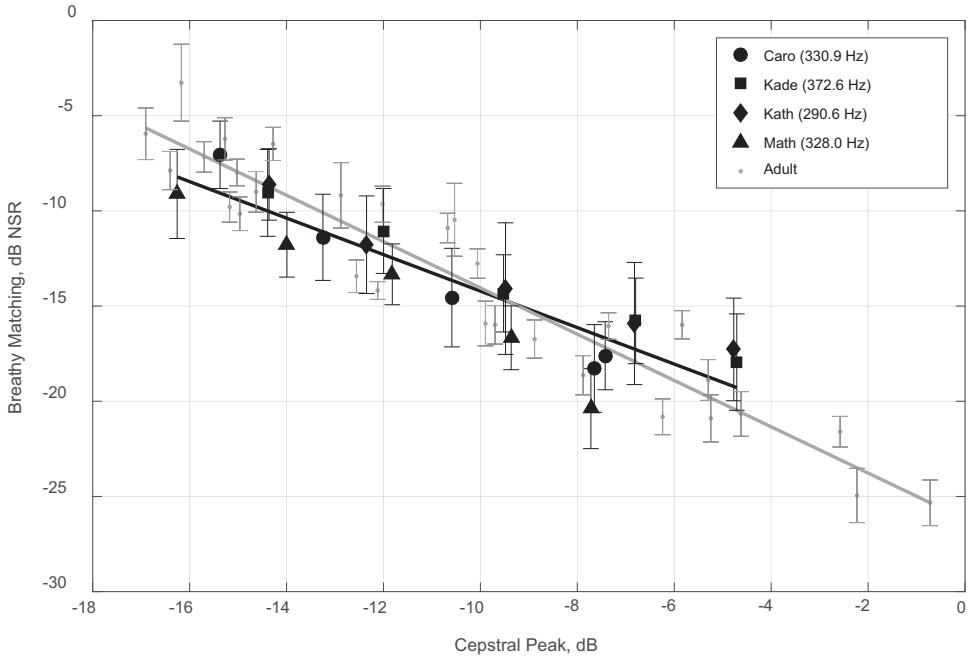

FIGURE 5. Matching NSR by cepstral peak in dB. The mean ± SE are given, averaged over all listeners. Black symbols are for pediatric voices, with symbol type separating the four different talkers. Gray symbols correspond to adult voices from Shrivastav et al.[52] Lines with the same shading scheme represent linear regression. In both cases, goodness of fit was high (pediatric voices: $r^2 = 0.82$; adult voices: $r^2 = 0.91$).

important that the SVMT measurement method was robust to a wide range of breathiness, and it is clear that the range of perceptual data reported here is similar to ranges reported for other data sets.[42,52]

We specifically chose to analyze and resynthesize vocal signals for pediatric talkers so that we could evaluate the loudness model that was used successfully to predict breathiness for adult voices ranging over a wide degree of perceived breathiness. A statistically significant effect of the combination of AH and OQ was found, indicating that this combination does affect listener perception of breathiness in pediatric voices, just as it did in adult voices.[52] In the case of the loudness ratio model (Figure 3), correlations between model predictions and perceptual data were strong for both pediatric and adult voices. However, for adult voices, the goodness of fit was markedly improved with the addition of an f_0 parameter into the regression model. Similarly, linear models relating pitch strength estimates to perceived breathiness did not require an f_0 parameter for pediatric voices, whereas model predictions for adult voices were much better with the addition of an f_0 parameter. In contrast, for the cepstral peak measure, the f_0 parameter did not improve the goodness of fit for either pediatric or adult voices. In all cases, an explanation may be that, for pediatric voices, model accuracy was already high for the simpler models ($r^2 > 0.9$), leaving little room for improvement by including f_0 in the models. Another contributing factor is that the pediatric voices spanned a narrow range of f_0 (~4 semitones), whereas the adult voices spanned a much wider range of f_0 (~12 semitones), thus limiting the distinction in f_0 among the pediatric voices.

In general, it would be advantageous if the computational indices associated with vocal breathiness were not strongly dependent on parameters such as f_0 because f_0 estimation often fails for type II voices and almost always fails for type III voices.[38] Considering the three analysis methods used here, the loudness ratio estimates were highly correlated with perception but are not appropriate for natural voices and have some f_0 dependence. Pitch strength estimates worked well with natural voices but, at least for the Shrivastav et al data set, showed an f_0 dependence that would be challenging for severely dysphonic type III voices. The cepstral peak measure, on the other hand, was highly correlated with perception and was f_0 independent for the current pediatric data set, as well as the Shrivastav et al dataset with adult voices.

CONCLUSIONS

This study is the first to apply a psychoacoustic framework to the perception of breathiness in pediatric voices that includes robust psychophysical methods through an SVMT and analysis via bio-inspired algorithms. Listener evaluation of breathiness using the matching task had moderately high inter- and intralistener reliability. Furthermore, it was shown that perceptual measures were strongly correlated with acoustic measures of loudness ratio, pitch strength, and cepstral peak. For the pediatric voices in the current study, none of those relationships were strongly dependent on estimates of f_0. In comparison with breathiness measures from adult voices from Shrivastav et al,[52] which served as a model for the present investigation, the perceived breathiness of the current voices spanned a slightly smaller range. For the adult voices, the strongest correlations were obtained when f_0 was included in linear models based on loudness ratio or pitch strength measures, but not with cepstral peak values. These results are positive indicators that listeners can judge breathiness with high reliability in both pediatric and adult dysphonic talkers. Further, these data set the stage for using the same approach in a larger set of natural voices that span a wide range of ages, fundamental frequencies, and types of dysphonia.

Acknowledgment

This research was supported by a grant from NIH (Grant No. R01 DC009029).

REFERENCES

1. Carding PN, Roulstone S, Northstone K, et al. The prevalence of childhood dysphonia: a cross-sectional study. *J Voice*. 2006;20:623–630.
2. Bhattacharyya N. The prevalence of pediatric voice and swallowing problems in the United States. *Laryngoscope*. 2015;125:746–750.
3. Andrews M. *Voice Treatment for Children and Adolescents*. San Diego, CA: Singular Publishing Group, Inc.; 2002.
4. Ruddy BH, Sapienza CM. Treating voice disorders in the school-based setting. Working within the framework of IDEA. *Lang Speech Hear Serv Sch*. 2004;35:327–332.
5. Connor N, Cohen S, Theis S, et al. Attitudes of children with dysphonia. *J Voice*. 2008;22:197–209.
6. Ruddy BH, Lewis V, Sapienza CM. The role of the speech-language pathologist in the schools for the treatment of voice disorders: working within the framework of the Individuals with Disabilities Education Improvement Act. *Semin Speech Lang*. 2013;34:55–62.
7. Lass NJ, Ruscello DM, Bradshaw KH, et al. Adolescents' perceptions of normal and voice-disordered children. *J Commun Disord*. 1991;24:267–274.
8. Lass NJ, Ruscello DM, Stout LL, et al. Peer perceptions of normal and voice-disordered children. *Folia Phoniatr Logop*. 1991;43:29–35.
9. Dejonckere PH, Remacle M, Fresnel-Elbaz E, et al. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Rev Laryngol Otol Rhinol (Bord)*. 1996;117:219–224.
10. Shrivastav R. The use of an auditory model in predicting perceptual ratings of breathy voice quality. *J Voice*. 2003;17:502–512.
11. Shrivastav R, Sapienza CM. Objective measures of breathy voice quality obtained using an auditory model. *J Acoust Soc Am*. 2003;114(4 pt 1):2217–2224.
12. Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J Voice*. 2010;24:540–555.
13. Shrivastav R, Eddins DA, Anand S. Pitch strength of normal and dysphonic voices. *J Acoust Soc Am*. 2012;131:2261–2269.
14. Glaze LE, Bless DM, Milenkovic P, et al. Acoustic characteristics of children's voice. *J Voice*. 1988;2:312–319.
15. Lopes LW, Lima ILB, Almeida LNA, et al. Severity of voice disorders in children: correlations between perceptual and acoustic data. *J Voice*. 2012;26:819.e7–819.e12.
16. Reynolds V, Buckland A, Bailey J, et al. Objective assessment of pediatric voice disorders with the acoustic voice quality index. *J Voice*. 2012;26:672.e1–672.e7.
17. Campisi P, Tewfik TL, Manoukian JJ, et al. Computer-assisted voice analysis: establishing a pediatric database. *Arch Otolaryngol Head Neck Surg*. 2002;128:156–160.
18. Masaki A. Optimizing acoustic and perceptual assessment of voice quality in children with vocal nodules: Massachusetts Institute of Technology; 2009.
19. Abbott KV, Li NY, Hersan R, et al. Voice therapy for children. In: *Clinical Management of Children's Voice Disorders*. San Diego, CA: Plural Publishing; 2010:111–133.
20. Maturo S, Hill C, Bunting G, et al. Establishment of a normative pediatric acoustic database. *Arch Otolaryngol Head Neck Surg*. 2012;138:956–961.

21. Infusino SA, Diercks GR, Rogers DJ, et al. Establishment of a normative cepstral pediatric acoustic database. *JAMA Otolaryngol Head Neck Surg*. 2015;141:358–363.
22. Sapienza CM, Ruddy BH, Baker S. Laryngeal structure and function in the pediatric larynx: clinical applications. *Lang Speech Hear Serv Sch*. 2004;35:299–307.
23. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). *J Voice*. 1999;13:557–569.
24. Hartnick CJ. Validation of a pediatric voice quality-of-life instrument: the pediatric voice outcome survey. *Arch Otolaryngol Head Neck Surg*. 2002;128:919–922.
25. Hartnick CJ, Volk M, Cunningham M. Establishing normative voice-related quality of life scores within the pediatric otolaryngology population. *Arch Otolaryngol Head Neck Surg*. 2003;129:1090–1093.
26. Boseley ME, Cunningham MJ, Volk MS, et al. Validation of the pediatric voice-related quality-of-life survey. *Arch Otolaryngol Head Neck Surg*. 2006;132:717–720.
27. Zur KB, Cotton S, Kelchner L, et al. Pediatric Voice Handicap Index (pVHI): a new tool for evaluating pediatric dysphonia. *Int J Pediatr Otorhinolaryngol*. 2007;71:77–82.
28. Kempster GB, Gerratt BR, Abbott KV, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
29. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20:14–22.
30. Kelchner LN, Brehm SB, Weinrich B, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. *J Voice*. 2010;24:441–449.
31. Nuss RC, Ward J, Huang L, et al. Correlation of vocal fold nodule size in children and perceptual assessment of voice quality. *Ann Otol Rhinol Laryngol*. 2010;119:651–655.
32. Johnson K, Brehm SB, Weinrich B, et al. Comparison of the pediatric voice handicap index with perceptual voice analysis in pediatric patients with vocal fold lesions. *Arch Otolaryngol Head Neck Surg*. 2011;137:1258–1262.
33. de Alarcon A, Brehm SB, Kelchner LN, et al. Comparison of pediatric voice handicap index scores with perceptual voice analysis in patients following airway reconstruction. *Ann Otol Rhinol Laryngol*. 2009;118:581–586.
34. Shrivastav R, Sapienza CM. Some difference limens for the perception of breathiness. *J Acoust Soc Am*. 2006;120:416–423.
35. Oates JM, Kirkby RJ. An acoustic investigation of voice quality disorders in children with vocal nodules. *Aust J Hum Commun Disord*. 1980;8:28–39.
36. Niedzielska G, Glijer E, Niedzielski A. Acoustic analysis of voice in children with noduli vocales. *Int J Pediatr Otorhinolaryngol*. 2001;60:119–122.
37. Simões-Zenari M, Nemr K, Behlau M. Voice disorders in children and its relationship with auditory, acoustic and vocal behavior parameters. *Int J Pediatr Otorhinolaryngol*. 2012;76:896–900.
38. Titze I. *Workshop on Acoustic Voice Analysis: Summary Statement*. Denver, CO: National Center for Voice and Speech; 1994.
39. Kelchner LN, Weinrich B, Brehm SB, et al. Characterization of supraglottic phonation in children after airway reconstruction. *Ann Otol Rhinol Laryngol*. 2010;119:383–390.
40. Isshiki N, Okamura H, Tanabe M, et al. Differential diagnosis of hoarseness. *Folia Phoniatr Logop*. 1969;21:9–19.
41. Zwicker E, Fastl H. Pitch and pitch strength. In: *Psychoacoustics: Facts and Models*. New York, NY: Springer-Verlag; 1990:103–132.
42. Patel S, Shrivastav R, Eddins DA. Developing a single comparison stimulus for matching breathy voice quality. *J Speech Lang Hear Res*. 2012;55:639–647.
43. Eddins DA, Shrivastav R. Psychometric properties associated with perceived vocal roughness using a matching task. *J Acoust Soc Am*. 2013;134:El294–El300.
44. Eddins DA, Kopf LM, Shrivastav R. The psychophysics of roughness applied to dysphonic voice. *J Acoust Soc Am*. 2015;138:3820–3825.
45. Shrivastav R, Camacho A. A computational model to predict changes in breathiness resulting from variations in aspiration noise level. *J Voice*. 2010;24:395–405.
46. Eddins DA, Anand S, Camacho A, et al. Modeling of breathy voice quality using pitch-strength estimates. *J Voice*. 2016;30:774.e1–774.e1.
47. Fastl H, Zwicker E. *Psychoacoustics: Facts and Models*. New York; Berlin: Springer; 2007.
48. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. *J Speech Hear Res*. 1994;37:769–778.
49. Patel S, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: a comparison of psychophysical methods. *J Voice*. 2010;24:168–177.
50. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res*. 2005;48:323–335.
51. Brown WS Jr, Shrivastav R. Comfortable effort level in young children's speech. *Folia Phoniatr Logop*. 2007;59:227–233.
52. Shrivastav R, Camacho A, Patel S, et al. A model for the prediction of breathiness in vowels. *J Acoust Soc Am*. 2011;129:1605–1615.
53. Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am*. 1990;87:820–857.
54. Chen Z, Hu G, Glasberg BR, et al. A new method of calculating auditory excitation patterns and loudness for steady sounds. *Hear Res*. 2011;282:204–215.
55. Moore BC, Glasberg BR, Baer T. A model for the prediction of thresholds, loudness, and partial loudness. *J Audio Eng Soc*. 1997;45:224–240.
56. Camacho A. On the use of auditory models' elements to enhance a sawtooth waveform inspired pitch estimator on telephone-quality signals. In Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on (pp. 1080–1085). IEEE.
57. Skowronski MD, Shrivastav R, Hunter EJ. Cepstral peak sensitivity: a theoretic analysis and comparison of several implementations. *J Voice*. 2015;29:670–681.
58. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420.
59. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J Speech Hear Res*. 1996;39:311–321.
60. Awan SN, Roy N. Acoustic prediction of voice type in women with functional dysphonia. *J Voice*. 2005;19:268–282.
61. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.