Modeling of Breathy Voice Quality Using Pitch-strength Estimates

*David A. Eddins, †Supraja Anand, ‡Arturo Camacho, and §Rahul Shrivastav, *Tampa, Florida; †West Chester, Pennsylvania; ‡San Jose, Costa Rica; and §Athens, Georgia

Summary: Background. The characteristic voice quality of a speaker conveys important linguistic, paralinguistic, and vocal health-related information. Pitch strength refers to the salience of pitch sensation in a sound and was recently reported to be strongly correlated with the magnitude of perceived breathiness based on a small number of voice stimuli.

Objective. The current study examined the relationship between perceptual judgments of breathiness and computational estimates of pitch strength based on the Aud-SWIPE (P-NP) algorithm for a large number of voice stimuli (330 synthetic and 57 natural).

Methods and Results. Similar to the earlier study, the current results confirm a strong relationship between estimated pitch strength and listener judgments of breathiness such that low pitch-strength values are associated with voices that have high perceived breathiness. Based on this result, a model was developed for the perception of breathy voice quality using a pitch-strength estimator. Regression functions derived between the pitch-strength estimates and perceptual judgments of breathiness obtained from matching task revealed a linear relationship for a subset of the natural stimuli. We then used this function to obtain predicted breathiness values for the synthetic and the remaining natural stimuli.

Conclusions. Predicted breathiness values from our model were highly correlated with the perceptual data for both types of stimuli. Systematic differences between the breathiness of natural and synthetic stimuli are discussed.

Key Words: Listener perception–Breathiness–Matching task–Pitch strength–Aud-SWIPE (P-NP).

INTRODUCTION

Auditory perceptual evaluation of dysphonic voice quality is one of the most common and valuable clinical tools for determining severity of voice disorders and measuring treatment outcomes. Terminologies such as "breathiness," "roughness," and "strain" are often used to represent perceptions of dysphonic voice quality and are widely used in standardized voice assessment such as Grade, Roughness, Breathiness, Asthenia, Strain[1] and Consensus Auditory-Perceptual Evaluation of Voice.[2] The present study will specifically focus on the development of a model to quantify listener perception of breathiness, a quality resulting from an incomplete glottal closure resulting in an audible air escape.[2]

Although listener judgments of breathiness are an integral part of voice assessment, there have been persistent concerns regarding their reliability and validity.[3] Many acoustic measures such as signal-to-noise ratio (SNR), relative amplitude of first harmonic, and spectral slope have been developed to quantify breathiness in an objective manner.[4–8] These acoustic measures often reflect the underlying physiology of modifications in the glottal area and in glottal flow.[9] Past research designed to establish a relationship between the vocal acoustic signal and the perception of breathiness has produced highly inconsistent results.[10] The poor correspondence between the acoustic measures and the listener judgments of breathiness might be attributed to the assumption of a linear relationship between acoustic and perceptual variables. Over the past decade, it has been shown that the relationship between a physical stimulus (dysphonic voice sample) and its perceptual attributes is often nonlinear, yet can be successfully described by analytical models that include many of the nonlinear properties of the auditory system.[11–15] For example, modeling breathiness using auditory measures derived from a loudness model[16] resulted in better correlations with perceptual data than traditional acoustic measures such as cepstral peak prominence.[12,17] In this model, the loudness elicited by the aperiodic components in the voice was computed as "noise loudness" (NL) and the loudness elicited by the harmonic energy of the vowel that is masked by the aperiodic components of the same voice was computed as "partial loudness" (PL). Breathiness was directly related to NL and inversely related to PL. A computational model of breathiness was developed[14] in which breathiness $\left(\hat{b}\right)$ was predicted from a ratio of NL/PL or ratio of noise loudness to partial loudness (η) and fundamental frequency (f_0) expressed in equivalent rectangular bandwidth units, or Φ, as shown in Equation 1.

$$\hat{b} = 0.45\eta ideal^{1/(7.054-0.78\phi)} + 0.026 \tag{1}$$

The model accounted for 76.2% of the variance in the perceptual data for natural stimuli. However, predictions were least accurate for stimuli at both low and high ends of the breathiness continuum.

More recently, Shrivastav and colleagues have shown that the perception of breathiness and the perception of pitch strength were strongly related to each other.[15] Here, pitch strength refers to how strong or how faint a listener perceives the sensation of pitch in a sound stimulus. This quality is also referred to as "pitch salience." For example, a sustained vowel like /a/ evokes a strong pitch sensation and thereby has higher pitch strength when

Accepted for publication November 20, 2015.

From the *Department of Communication Sciences and Disorders, University of South Florida, Tampa, Florida; †Department of Communication Sciences and Disorders, West Chester University, West Chester, Pennsylvania; ‡School of Computer Science and Informatics, University of Costa Rica, San Jose, Costa Rica; 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, Florida 32620. E-mail: deddins@usf.edu

Journal of Voice, Vol. 30, No. 6, pp. 774.e1–774.e7
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2015.11.016

compared with a fricative consonant like /s/. Pitch strength of voices has been evaluated only recently,[15] whereas pitch strength has been extensively investigated for a variety of tonal and noise stimuli.[18–21] The authors examined the reliability of listener judgments of pitch strength and the correlation between these judgments and perceptions of breathiness and roughness. The authors first established that listeners can indeed judge pitch strength in dysphonic voices (inter-rater reliability of 0.87 and intra-rater reliability of 0.80). Further, perceived pitch-strength judgments and perceived breathiness were highly correlated ($r = -0.99$; $P < 0.001$) for 11 breathy voice samples. Stimuli with greater breathiness were perceived to have lower pitch strength, supporting an inverse relationship. Collectively, these results indicate that listeners were able to judge the pitch strength of dysphonic voices in a reliable manner and that pitch strength decreases systematically as breathy voice quality increases. Given this strong relationship between perceived breathiness and perceived pitch strength, it is plausible that a computational model of pitch strength could *predict* listener perception of breathiness, perhaps with less variability than the loudness-based model described above and much less time-consuming than perceptual judgments. Therefore, the goals of this study were to (1) confirm the relationship between perceived breathiness and computational estimates of pitch strength in a larger set of breathy stimuli, and (2) develop a model for perception of breathiness using acoustic estimates of pitch strength. If the relationship between breathiness and pitch strength reported previously extends to a broader range of stimuli and to different perceptual tasks, then there will be a high correlation between perceived breathiness and computational estimates of pitch strength. If successful, the pitch-strength model of breathiness would offer advantages over the PL model[13,14] (although the two likely are correlated)[22] in that pitch-strength estimation may directly account for pitch dependencies seen in the loudness-based model, and does not require separation of the stimulus into periodic and aperiodic components. Furthermore, the algorithm used to compute pitch strength can be used for running speech, making it simpler to extend such a model from vowels to conversational speech samples.

To best achieve the goals stated above, we leverage the results obtained from previous experiments using three different psychophysical methods in which listeners judged breathiness: rating scale, magnitude estimation, and a single-variable matching task (SVMT). The primary focus is on data from the SVMT because it yields ratio-level data, minimizes the potential for contextual bias, and eliminates the need for numerical scaling and any related errors.[23] Briefly, rating scales are often characterized by "equal appearing intervals," with the accompanying unverified assumption that listeners are aware of the total range of breathiness variations and that they can accurately divide this breathiness range into subjectively equal perceptual distances. By definition, magnitude estimation tasks remove this assumption because listeners judge ratios of sensation rather than abstract intervals. However, both magnitude estimation and rating scales are highly dependent on context, and can be affected by number, range, and frequency of stimulus attribute(s). High chances of inherent perceptual biases prevalent in both these methods can be removed with matching tasks that provide listeners with a reference to compare and judge voice samples. As a result, comparisons of ratings from matching tasks are more reliable across stimuli and across experiments.

METHODS

Stimuli

Breathy voice stimuli were selected from prior perceptual studies.[14,17,23,24] Of these, 57 phonation samples of vowel /a/ that represented a wide range of breathiness (from nearly normal to severe breathiness) were selected from the Kay Elemetrics Disordered Voice Database (Kay Elemetrics, Inc., Lincoln Park, NJ), henceforth referred as "natural stimuli." These stimuli were recorded at a sampling rate of 50,000 Hz and were downsampled to 24,414 Hz with 16-bit quantization to match the hardware requirements for perceptual experiments. In addition, 10 phonation samples of vowel /a/ (five male and five female; henceforth referred to as "talkers") were generated using the *Klatt Synthesizer* (Sensimetrics Corporation, Malden, MA), with the Liljencrants-Fant (LF) model[25,26] as glottal source. The synthesizer parameters of these samples, including the f_0 and the first three formant frequencies (F1, F2, and F3), were set based on 10 natural voice samples selected from a pilot listening experiment (four listeners judged the breathiness level of 50 randomly selected voices from the Kay Elemetrics Disordered Voice Database. Then, these ratings were averaged and rank ordered into five levels of varying breathiness [mild to severe] and one male/one female sample from each breathiness level). Further, prior research has indicated that changes in the aspiration noise level (AH) and the open quotient (OQ) parameters of synthetic vowels impact breathiness of the synthesized vowels.[6,25] Therefore, source and filter parameters including AH, OQ, tilt, and formant bandwidths were also manipulated (Table 1). Specifically, AH and OQ were systematically manipulated to generate separate stimulus continua with 11 levels of increasing breathiness. A third continuum that covaried in both AH and OQ (referred to as AO) was also created. Each of the stimulus series (AH, OQ, and AO) consisted of 10 talkers and 11 "breathiness levels" corresponding to 11 AH (~0–80 dB), OQ (~25–99%), or combined AH/OQ values. This resulted in a total of 110 stimuli per series. Thus, a total of 330 "synthetic stimuli" (three series × 10 talkers × 11 levels) were created. These synthetic stimuli were sampled at a rate of 12,207 Hz with 16-bit quantization. They were edited to have 500 ms duration and were equated to have the same root mean square amplitude. Finally, they were shaped with a 20-ms cosine-squared onset and offset window to avoid any audible clicks.

Listener judgments of breathiness

Approximately 10 individuals between 22 and 25 years of age were tested for each of the perceptual experiments cited below. Individual listeners participated in only one of the perceptual studies cited below with one exception. One listener participated in the SVMT study with natural stimuli and in the pitch-strength study. All participants were native speakers of American English, were students enrolled in either the Communication Sciences or Disorders or the Linguistics program, had taken at least one course in Communication Sciences or Disorders, and had

TABLE 1.
Klatt Synthesizer Parameters Used to Synthesize 10 Talkers (Adapted from Shrivastav et al[14,15])

	Male Talkers					Female Talkers				
Parameter	1	2	3	4	5	1	2	3	4	5
F_0 (Hz)	133.1	113.7	115.5	117	134.4	220.4	209.0	209.1	195.5	200.7
AV (dB)	60	60	60	60	60	60	60	60	60	60
OQ (%) range	25–99	25–99	25–99	35–99	30–85	30–99	35–99	35–99	25–99	30–99
Step size	7.4	7.4	7.4	6.4	5.5	6.9	6.4	6.4	7.4	6.4
SQ (%)	200	200	200	200	200	200	150	350	200	200
TL (dB)	0	10	20	30	40	0	10	20	30	40
FL (%)	10	10	10	10	10	10	10	10	10	10
AH (dB) range	0–75	0–80	0–75	0–80	55–80	0–80	0–80	0–75	0–80	55–80
Step size	7.5	8	7.5	8	2.5	8	8	7.5	8	2.5
FNP (Hz)	180	180	180	180	180	180	180	180	280	180
BNP (Hz)	1000	1000	1000	1000	1000	1000	1000	40	90	30
F1 (Hz)	661	559	814	586	814	891	759	1050	977	957
B1 (Hz)	200	400	600	800	1000	200	400	600	800	1000
F2 (Hz)	1122	1214	1473	1187	1473	1587	1333	1410	1356	1619
B2 (Hz)	200	200	200	200	200	200	200	200	150	200
F3 (Hz)	2281	2340	2250	2463	2250	3083	2930	3000	2905	2877
B3 (Hz)	300	300	300	200	250	300	300	300	200	250

Abbreviations: AH, amplitude of aspiration; AV, amplitude of voicing; BNP, bandwidth of nasal pole; B1, bandwidth of F1; B2, bandwidth of F2; B3, bandwidth of F3; FL, flutter; FNP, frequency of nasal pole; F_0, fundamental frequency; F1, first formant frequency; F2, second formant frequency; F3, third formant frequency; OQ, open quotient; SQ, speed quotient; TL, spectral tilt.

hearing thresholds less than 20 dB HL *via* air conduction at frequencies of 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. All perceptual experiments were approved by the institutional review board. Listeners consented to the procedures and received compensation for their participation. All data acquisition was controlled using *Sykofizx* software and TDT System 3 hardware (Tucker-Davis Technologies, Inc., Alachua, FL). Stimuli were presented to the listeners in a single-walled sound booth using ER-2 insert earphones (Etymotic Research Inc, Elk Grove Village, IL). These earphones were chosen for their flat frequency response at the tympanic membrane. Stimuli were presented at 85-dB sound pressure level in the right ear, to avoid potential effects of binaural interaction. For each perceptual experiment, listeners made judgments of breathiness using one of the following psychophysical methods: (1) rating scale, (2) magnitude estimation, or (3) SVMT.

Using a rating scale task, 27 natural and 330 synthetic stimuli were evaluated. Each natural stimulus was presented to listeners 10 times in a random order for breathiness rating on a five-point scale (0: minimal breathiness, 4: maximum breathiness). Similarly, each synthetic stimulus was presented to listeners five times in a random order for breathiness rating on a seven-point scale (1: minimal breathiness, 7: maximum breathiness). The breathiness ratings for each stimulus were averaged across all trials and listeners to estimate group means for each voice stimulus. For the magnitude estimation task, 30 natural and 330 synthetic stimuli were tested. Here, listeners assigned each voice stimulus a number that reflected the magnitude of breathiness (1–1000, continuous scale including fractions, excluding zero) without anchors. Listeners were instructed that a stimulus perceived to be twice as breathy as another should be assigned double the score as the first one. Stimuli from each synthetic series (ie, AH, OQ, and AO) were presented five times in blocks of 10; stimuli were randomized within each block and across each listener. The breathiness ratings for each stimulus were averaged across all trials and listeners to estimate group means for each voice stimulus. The rating scale task took approximately 1–2 hours and the direct magnitude estimation took approximately 3–4 hours for each listener.

For the SVMT, listeners evaluated the degree of breathiness in dysphonic voices by comparing them with a synthetic comparison signal. The comparison stimulus was created by mixing a low-pass filtered sawtooth waveform (151 Hz; –7 dB/octave) with a speech-shaped noise with identical filter. The variable parameter was the relative level of the sawtooth waveform to the noise of the comparison signal (SNR). On a given listening trial, the reference sound (one of 330 voice samples) was presented followed by the comparison sound. The SNR of the comparison stimulus (in decibels; dB) was increased or decreased by the listener until the perceived breathiness of the two stimuli was judged to be equal. The SNR for each subject and condition was taken as the average of five ascending (low initial SNR) and five descending (high initial SNR) blocks of trials. (Refer to Patel et al[23] for a detailed description of each of these methods.) For the SVMT, only five of the 11 levels of synthetic stimuli were tested from each talker (10 talkers × five levels; N = 50) because this task was much more time-consuming than the rating scale and the magnitude estimation tasks, and because data obtained using rating or magnitude estimation indicated very little difference in magnitude of perceived breathiness between adjacent points along the stimulus continuum for some of the stimuli included here.

Pitch-strength estimates

Computational estimates of pitch strength $\left(\hat{\Psi}\right)$ were obtained for all stimuli (57 natural and 330 synthetic stimuli) from a sawtooth waveform-inspired pitch estimator with auditory front-end (Aud-SWIPE).[27,28] Briefly, the algorithm computes a stimulus spectrum using an auditory front end and correlates the spectrum with a family of sawtooth kernel functions constructed over a range of f_0 (pitch candidates). The pitch candidate producing the largest correlation is the output pitch of the algorithm, and the correlation value is referred to as the pitch strength. The auditory front end consists of transfer functions that model the outer and the middle ear (effectively flattening the spectral envelope), a time-aligned gammatone filter bank (simulating cochlear filtering), and half-wave rectifiers (simulating the mechanical to electrical transduction process). The rectified output of each channel is converted to the spectral domain using a Fast-Fourier Transform (FFT) and a Hanning window of approximately eight fundamental periods in length (based on pitch candidate period). The upper harmonics above the center frequency of each channel are suppressed, and the final auditory spectrum is computed by summing the square root magnitude spectra across channels. The sawtooth kernel functions, representing the spectra of sawtooth waveforms, are constructed for each pitch candidate by centering sinusoidal lobes at the f_0 and all prime-numbered harmonics, with each lobe spanning the range of $-\pi$ to $+\pi$ radians and a width equal to the pitch candidate. The sinusoidal lobes are used because they closely approximate the spectral leakage of harmonic pulses due to a Hanning window that is eight fundamental periods in length. The negative parts of each lobe perform inhibition on inter-harmonic regions, whereas the use of prime-numbered harmonics reduces the correlation of subharmonic pitch candidates. The amplitude of the lobes is scaled by the square root of inverse frequency, which follows the roll-off of a sawtooth spectrum and reduces the correlation of super-harmonic pitch candidates. Pitch and pitch strength were estimated using overlapping analysis frames and a frame rate of 100 frames per second.

The pitch candidate with the highest degree of similarity (a number between 0 and 1) between the sawtooth waveform and the spectrum of the input signal was taken as the estimated pitch height, and the actual correlation value was taken as an estimate of the pitch strength.

RESULTS

Relationship between breathiness judgments and computational estimates of pitch strength

The current study had two goals. The first goal was to confirm the relationship between perceived breathiness and pitch strength, as reported by Shrivastav et al.[15] To do so, perceptual judgments of breathiness were obtained using three psychophysical tasks and four sets of voices. A second goal was to develop a model for the perception of breathiness using computational estimates of pitch strength produced by Aud-SWIPE (P-NP) estimator. A summary of the relationship between the perceptual results and the pitch-strength estimates, organized by perceptual task and stimulus set, is shown in Table 2. Correlations (Pearson r) between perceived breathiness and pitch-strength estimates from Aud-SWIPE (P-NP) were consistently high, ranging from 0.82 to 0.97, in absolute value. Negative correlations on rating scale and magnitude estimation tasks indicate that stimuli perceived to have lower breathiness (ie, lower numbers on rating scale and magnitude estimation) were observed to have higher pitch strength, as estimated by Aud-SWIPE (P-NP). On the contrary, on the SVMT, lower decibel SNR values indicate a voice with high breathiness and positive correlation. The inverse relationship between pitch strength and perceived breathiness mirrors that reported by Shrivastav et al[15] in that voices perceived to be higher in breathiness have lower pitch strength. Based on these results, it is evident that pitch strength is strongly correlated with breathiness judgments.

TABLE 2.
Correlation Between Pitch-strength Estimates and Breathiness Judgments for Natural and Synthetic Stimuli Obtained Through Three Psychophysical Methods

Perceptual Task	Stimuli	Pearson r	R^2
Rating scale	Natural	–0.88	0.78
	Synthetic—AH	–0.94	0.89
	Synthetic—OQ	–0.82	0.67
	Synthetic—AO	–0.90	0.81
Magnitude estimation	Natural	–0.92	0.85
	Synthetic—AH	–0.97	0.94
	Synthetic—OQ	–0.94	0.88
	Synthetic—AO	–0.95	0.90
Matching	Natural	0.93	0.87
	Synthetic—AH (Naïve listeners)	0.94	0.89
	Synthetic—AH (Expert listeners)	0.90	0.82
	Synthetic—AO	0.91	0.83

Twenty-seven natural stimuli were judged on a rating scale. Thirty natural stimuli were judged on magnitude estimation and matching tasks.

Example of model predictions

The strong relationship between perceived breathiness and pitch strength estimated by Aud-SWIPE (P-NP) offers the possibility to predict perceived breathiness for a novel data set. It is most straightforward to predict perceptual data from the matching experiment (SVMT) because those are ratio-level data, have meaningful units, and are perhaps the most straightforward to interpret in terms of perceived breathiness. To evaluate this possibility for the available data, part of the natural dataset was chosen as the basis to estimate breathiness. This particular dataset included 15 natural voices varying along a continuum of low to high breathiness. The perceptual judgments of breathiness (averaged across 10 listeners) were plotted against the pitch-strength estimates from Aud-SWIPE (P-NP), as shown in Figure 1. It is evident that stimuli with lower pitch-strength estimates were perceived to have higher breathiness and *vice versa*. This relationship was well described *via* linear regression to the data ($R^2 = 0.87$) as follows:

$$\hat{b} = 24.78 \times \hat{\Psi} + 6.921 \tag{2}$$

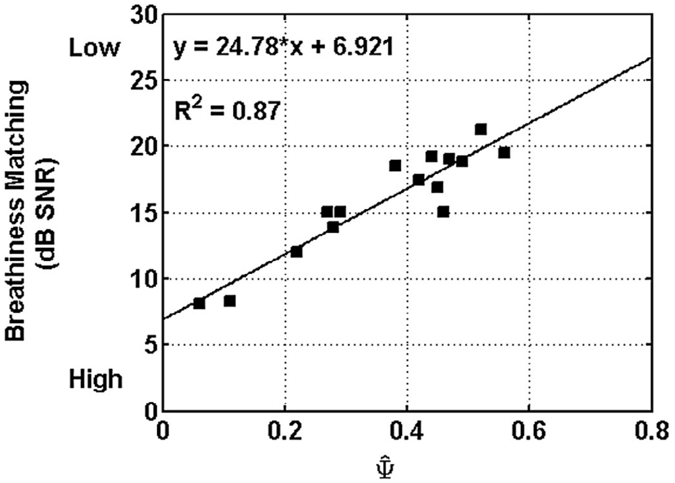

FIGURE 1. Pitch strength estimated with Aud-SWIPE (P-NP) $(\hat{\Psi})$ plotted against perceived breathiness obtained using SVMT for 15 natural stimuli along a continuum of vocal breathiness. Matching thresholds are reported in units of signal-to-noise ratio in decibel (see text for details) where values near 0 dB correspond to high perceived breathiness and values near 25 dB correspond to low perceived breathiness.

where $\hat{b}$ is perceived breathiness and $\hat{\Psi}$ is the pitch-strength estimate from Aud-SWIPE (P-NP). Based on this relationship, pitch-strength estimates from Aud-SWIPE (P-NP) can be used to predict perceived breathiness from the matching task in decibel SNR.

Predictions of the model using this regression equation were evaluated on (1) the remainder of the natural stimuli (N = 15), (2) synthetic AH series (N = 45)[a], and (3) synthetic AO series (N = 45)[b] for which perceptual breathiness matching data were obtained from 10 novel listeners for each data set. One means of assessing the quality of the predictions is to regress the predicted data onto the perceived data. Figure 2 compares perceived breathiness from SVMT with predicted breathiness from Equation 2. As shown in the figure, perceived and predicted breathiness were strongly correlated (r = 0.94, $P < 0.001$) for natural stimuli. This produced the following relationship:

$$\hat{b}_{actual} = 0.84 \times \hat{b}_{predicted} + 2.697 \tag{3}$$

which indicates that the breathiness model accounts for 88% of variance in the perceptual data of the natural stimuli (Figure 2).

To assess how well Equation 2 applies to synthetic data, we used it to predict breathiness ratings from the synthetic AH and AO series. As shown in Figure 3A and B, perceived and predicted breathiness were also strongly correlated to the AH (r = 0.94, $P < 0.001$) and the AO (r = 0.91, $P < 0.001$) synthetic stimuli. The average mean-squared error was 7.5 dB. Figure 4A and B depicts breathiness judgments *versus* pitch-strength estimates for the natural (black squares) and synthetic stimulus sets (AH unfilled circles in Figure 4A and AO unfilled triangles in Figure 4B). For the same perceived breathiness (ordinate), pitch-strength estimates (abscissa) were higher for synthetic stimuli than for natural stimuli. Furthermore, estimates of pitch strength for synthetic stimuli with the lowest perceived breathiness reached an asymptote at pitch-strength value of ~0.72.

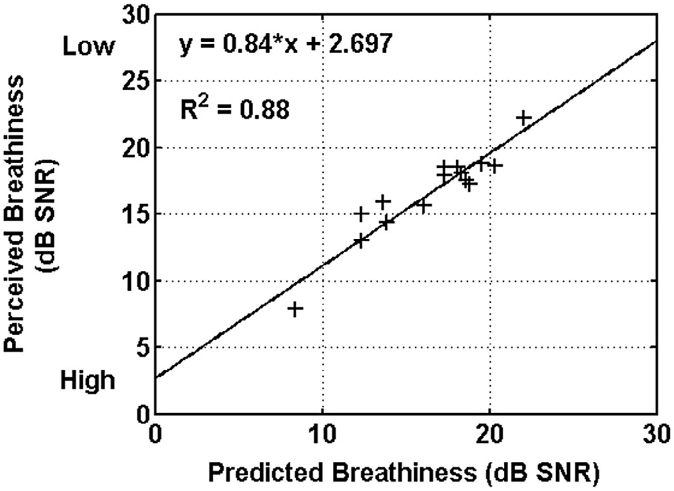

FIGURE 2. Breathiness predicted by the pitch strength model *versus* perceptual judgments of breathiness for the subset of natural stimuli.

DISCUSSION

Inspired by the strong relationship between perceptual judgments of vocal breathiness and perceptual judgments of pitch strength,[15] the current study was designed to evaluate the correspondence between perceived breathiness and a computational pitch estimator that yields an index of pitch strength (Aud-SWIPE [P-NP]).[28] As shown in Table 2, there was a strong relationship between estimated pitch strength and perceived breathiness that accurately characterized breathiness as measured with three different psychophysical tasks and multiple sets of voices, ranging from normal to extremely dysphonic. This relationship was observed over different stimulus types and sets (natural or synthetic voices, breathiness synthesized by manipulating AH, OQ, and AO). Hence, we derived a computational model for perception of breathiness analogous to the model proposed by Shrivastav et al[14] that was based on the PL model of Moore et al.[16] However, the nature of the perceptual data and the models used to predict breathiness differ in both studies. In terms of the perceptual data, Shrivastav et al[14] used a free direct magnitude estimation task rather than the SVMT used here. The matching task was preferred here over direct magnitude estimation because it yields ratio-level data, minimizes the potential for contextual bias, and eliminates the need for numerical scaling and any related errors.[23,24]

Shrivastav et al modeled breathiness as a power function of the η. Unfortunately, it was found that power varied with the f_0 of the stimulus, such that vowels with lower f_0 had a greater increase in breathiness for an equal change in η.[14] Therefore, the loudness ratio model had to explicitly account for changes in f_0. On the contrary, the breathiness model described in this study has only one predictive parameter: pitch strength (ie, the value of f_0 is not needed). Furthermore, the loudness ratio model

[a] Note that the results for one talker (t06) were discarded from this analysis as it was determined, after data collection, that the voice was diplophonic and that the synthetic version was highly unnatural sounding, resulting in a large deviation from the perceptual evaluations of the rest of the voices.

[b] See footnote a.

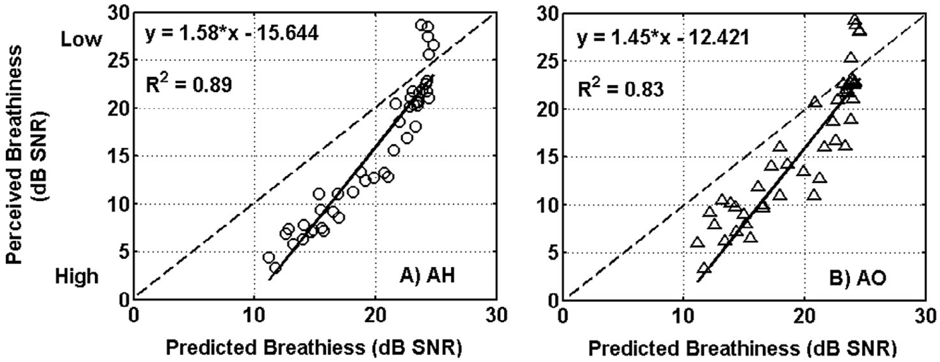

FIGURE 3. **A.** Breathiness predicted by the pitch strength model *versus* perceptual judgments of breathiness for synthetic AH stimuli. **B.** Breathiness predicted by the pitch strength model *versus* perceptual judgments of breathiness for synthetic AO stimuli.

required synthetic stimuli to compute the PL associated with the periodic and the aperiodic parts of the stimulus (those are not easily separable in natural stimuli), whereas pitch-strength estimates can be obtained on the composite synthetic or natural stimulus.

By definition, pitch strength is independent of pitch itself.[21] To evaluate this premise in voices, we computed f_0 using the *TF32* software (Madison, WI),[29] pitch using the Aud-SWIPE (P-NP) algorithm, and then computed correlations with estimated pitch strength from Aud-SWIPE (P-NP) and perceived breathiness from the matching task. For the natural stimuli, computed f_0 values were not significantly correlated with estimated pitch strength (r = 0.037, *P* = 0.847) or perceived breathiness (r = 0.098, *P* = 0.608). Additionally, computed pitch values were not significantly correlated with estimated pitch strength (r = 0.239, *P* = 0.204) or perceived breathiness (r = 0.327, *P* = 0.078). For the synthetic stimuli, different levels of breathiness were generated by varying AH and OQ while maintaining f_0. Thus, the presence of multiple perceived breathiness thresholds or pitch-strength estimates for a single f_0 (talker) violates the homoscedasticity assumption for linear regression. For these reasons, measures of f_0 were not included in the model.

The breathy model (based on natural stimuli) generalized very well to novel natural stimuli and poorly to synthetic stimuli. In Figure 3, the predicted breathiness of synthetic stimuli was lower than the perceived breathiness (by up to 7 dB) for most of the stimulus range (<22 dB), essentially underestimating the perceived breathiness of synthetic stimuli. For the stimuli judged to be the least breathy (>27 dB), the model overestimated the perceived breathiness of synthetic stimuli. The empirical basis for the model mismatch is evident in Figure 4, with different relationships between breathiness and estimates of pitch strength for natural and synthetic stimuli. For a given perceived breathiness (eg, 15 dB), the pitch strength of natural stimuli was about 0.20 lower than that of synthetic stimuli. The added jitter and shimmer present in the natural stimuli as well as differences in stimulus bandwidth, the LF model used in synthesis, and potential limitations in the Aud-SWIPE (P-NP) algorithm may contribute to the lower pitch-strength values for natural stimuli given a specific perceived breathiness value. For both AH and AO synthetic stimuli in Figure 4, estimates of pitch strength appeared to saturate near 0.7 for the least breathy stimuli although the range of perceived breathiness was below that for natural stimuli and may thus be of little practical concern.

CONCLUSIONS

In this study, a computational model was developed for predicting the perceived breathiness of acoustic stimuli using estimates of pitch strength based on the Aud-SWIPE (P-NP) model. Results indicate high correlations between perceived breathiness judgments from human observers and model predictions (R^2 = 0.67–0.94). This model is simpler and more accurate

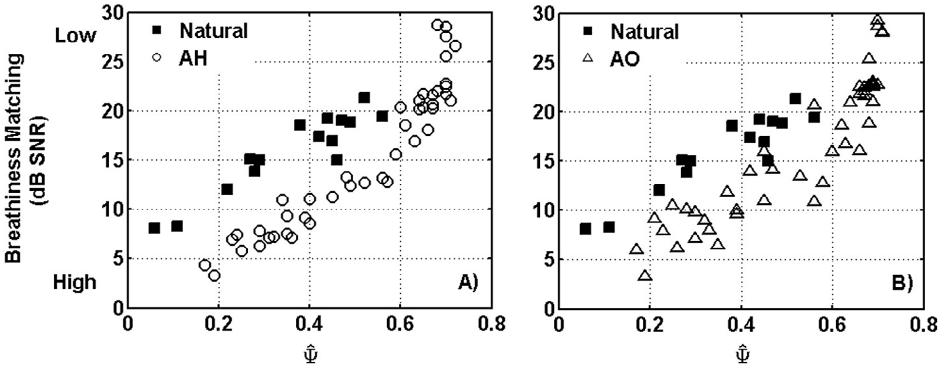

FIGURE 4. **A.** Pitch strength estimated from Aud-SWIPE (P-NP) $(\hat{\Psi})$ plotted against perceived breathiness obtained using SVMT for natural and synthetic AH stimuli. **B.** Pitch strength estimated from Aud-SWIPE (P-NP) $(\hat{\Psi})$ plotted against perceived breathiness obtained using SVMT for natural and synthetic AO stimuli.

than a loudness-based model described previously.[13,14] The current data indicate that the high precision predictions for natural stimuli can be extended to synthetic stimuli with a simple correction factor. Ultimately, the model utility will be determined by the potential to accurately predict perceived breathiness on novel data sets as those data are collected in the context of other investigations.

Acknowledgments

This research was supported by a grant from the National Institutes of Health (Grant No. R01 DC009029). The authors wish to thank Mark Skowronski for helpful comments on an earlier draft of the article.

REFERENCES

1. Hirano M. *Clinical Examination of Voice*. New York, NY: Springer-Verlag; 1981.
2. Kempster GB, Gerratt BR, Verdolini-Abbott K, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
3. 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.
4. de Krom G. Consistency and reliability of voice quality ratings for different types of speech fragments. *J Speech Hear Res*. 1994;37:985–1000.
5. de Krom G. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *J Speech Hear Res*. 1995;38:794–811.
6. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. *J Speech Hear Res*. 1994;37:769–778.
7. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J Speech Hear Res*. 1996;39:311–321.
8. Martin D, Fitch J, Wolfe V. Pathologic voice type and the acoustic prediction of severity. *J Speech Hear Res*. 1995;38:765–771.
9. Hanson HM. Glottal characteristics of female speakers: acoustic correlates. *J Acoust Soc Am*. 1997;101:466–481.
10. Kreiman J, Gerratt B. Measuring voice quality. In: Kent RD, Ball MJ, eds. *Voice Quality Measurement*. 1st ed. San Diego, CA: Singular Publishing Group; 2000:73–101.
11. Shrivastav R. Perceptual Structure of Breathy Voice Quality and Auditory Modeling of its Acoustic Cues. Unpublished Doctoral Dissertation, Indiana University, Bloomington; 2000.
12. Shrivastav R. The use of an auditory model in predicting perceptual ratings of breathy voice quality. *J Voice*. 2003;17:502–512.
13. 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.
14. Shrivastav R, Camacho A, Patel SA, et al. A model for prediction of breathiness in vowels. *J Acoust Soc Am*. 2011;125:1605–1615.
15. Shrivastav R, Eddins DA, Anand S. Pitch strength of normal and dysphonic voices. *J Acoust Soc Am*. 2012;131:2261–2269.
16. Moore BCJ, Glasberg BR, Baer T. A model for the prediction of thresholds, loudness and partial loudness. *J Audio Eng Soc*. 1997;45:224–239.
17. Shrivastav R, Sapienza C. Objective measures of breathy voice quality obtained using an auditory model. *J Acoust Soc Am*. 2003;114:2217–2224.
18. Patterson RD, Handel S, Yost WA, et al. The relative strength of tone and noise components of iterated rippled noise. *J Acoust Soc Am*. 1996;100:3286–3294.
19. Shofner WP, Selas G. Pitch strength and Stevens' power law. *Percept Psychophys*. 2002;64:437–450.
20. Yost WA. Pitch strength of iterated rippled noise. *J Acoust Soc Am*. 1996;100:3329–3335.
21. Zwicker E, Fastl H. Pitch and pitch strength. In: *Psychoacoustics: Facts and Models*. New York, NY: Springer-Verlag; 1990:103–132.
22. Hansen H. Tone-noise dichotomy: Investigating tonal content magnitude and pitch strength. Dissertation, University of Oldenburg, Germany; 2010.
23. Patel S, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: a comparison of psychophysical methods. *J Voice*. 2010;24:168–177.
24. 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.
25. Fant G, Liljencrants J, Lin Q. A four parameter model of glottal flow. Speech Transmission Laboratory Quarterly Report. 1985;1–3.
26. 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.
27. 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 Information Science, Signal Processing, and their Applications, 2012: 1107–1112.
28. Camacho A, Harris JG. A sawtooth waveform inspired pitch estimator for speech and music. *J Acoust Soc Am*. 2008;124:1638–1652.
29. Milenkovic PH. TF32 (Computer Software). Madison, WI; 2001.