Published in final edited form as:
*J Voice*. 2010 July ; 24(4): 395–405. doi:10.1016/j.jvoice.2008.12.001.

# A computational model to predict changes in breathiness resulting from variations in aspiration noise level

**Rahul Shrivastav** and
Department of Communication Sciences and Disorders, University of Florida, Gainesville, Florida, 32611 and Brain Rehabilitation Research Center, Malcom Randall VA Medical Center, Gainesville, FL

**Arturo Camacho**
Department of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, 32611

Rahul Shrivastav: rahul@ufl.edu; Arturo Camacho: acamacho@cise.ufl.edu

## Abstract

Perception of breathy voice quality is cued by a number of acoustic changes including an increase in aspiration noise level (AH) and spectral slope[1]. Changes in AH in a vowel may be evaluated through measures such as the harmonic-to-noise ratio (HNR), cepstral peak prominence (CPP) or via auditory measures such as the partial loudness of harmonic energy (PL) and loudness of aspiration noise (NL). Although a number of experiments have reported high correlation between such measures and ratings of perceived breathiness, a formal model to predict breathiness of a vowel has not been proposed. This research describes two computational models to predict changes in breathiness resulting from variations in AH. One model uses auditory measures while the other uses CPP as independent variables to predict breathiness. For both cases, a translated and truncated power function is required to predict breathiness. Some parameters in both of these models were observed to be pitch-dependent. The "unified" model based on auditory measures was observed to be more accurate than one based on CPP.

## 1. Introduction

An increase in the level of aspiration noise (AH) is one of the primary acoustic cues for the perception of breathy voice quality [1,2]. Several algorithms have been proposed to quantify the relative level of aspiration noise in voices [(see 3, for a review)]. These algorithms vary in their underlying assumptions, but generally attempt to separate a vowel into a periodic and an aperiodic component. The ratio of the level of these components is then used to quantify the level of "breathiness" in vowels. More recently, Shrivastav and his colleagues [4,5] have applied an additional transformation to these two components. They used an auditory processing model as a signal processing front-end to estimate (1) the loudness of the harmonic component in a vowel when it is masked by its aspiration noise, and (2) the loudness of the aspiration noise itself. The resulting measures were called the "partial loudness" (PL) of the harmonic energy and the aspiration "noise loudness" (NL), respectively. PL and NL, both measured in ratio-level units called "Sones", were observed to account for greater variance in

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

perceptual judgments of breathiness than conventional measures that estimated the ratio of aperiodic and periodic component levels. In these experiments, it was observed that breathiness was negatively correlated with PL and positively correlated with NL. It was also found that the breathiness in some of the vowels, particularly those with very high levels of AH, was better predicted by NL rather than PL[4].

The auditory transformation of the acoustic signal provides certain advantages over many conventional signal processing methods. First, it accounts for some of the non-linear transformations that are inherent to the auditory-perceptual process, thereby improving the fit to perceptual data. Second, the use of an auditory processing model also accounts for a part of the multidimensionality observed between various acoustic cues for breathiness and its perception. For example, it is known that in addition to AH, breathiness is also correlated with acoustic features such as the amplitude of the first harmonic, spectral slope and formant bandwidths [1]. Measures such as the PL and NL not only vary with the overall levels of AH and harmonic energy, but are also affected by changes in the spectral shape of these signals. Therefore, these auditory measures may capture, at least partially, the effects of multiple acoustic cues for breathiness.

Another approach that has been successful in quantifying breathiness is the cepstral peak prominence[2] (CPP). As the name describes, CPP is the normalized amplitude of the cepstral peak of a vowel segment and this has been found to be negatively correlated to the vowels' perceived breathiness. A direct comparison of the auditory based measures and CPP showed that PL and NL were slightly better than CPP in accounting for the variance in perceptual ratings of breathiness[4]. In contrast, other conventionally used measures of vowel acoustic signals such as measures of frequency and intensity perturbation, relative noise levels and spectral slope show relatively lower and often inconsistent correlation with perceptual judgments of breathiness [6–10]. A direct comparison of the auditory measures, CPP and many of these conventional acoustic measures (signal-to-noise ratio, jitter, shimmer, H1–H2, H1–A1 and H1–A3) showed that both auditory based measures and CPP resulted in a stronger correlation with perceptual data than the conventional acoustic measures [4]. Based on these findings, CPP and the auditory measures PL and NL may be considered to be the most sensitive measures to quantify changes in breathiness in vowels. Further, it must be noted that most of the conventionally used acoustic measures were originally intended either as descriptors of the voice acoustic signal (e.g. short-term perturbation measures) or were designed to identify some information regarding vocal fold dynamics (for example see [11] for the physiological correlates of H1–H2 or H1–A3). In contrast, some of the recently developed objective measures of voice, such as the PL and NL, have specifically been designed to quantify the perception of voice quality.

These differences in underlying bases for various objective measures may explain why a large number of experiments have reported the correlation between specific acoustic measures and perceptual judgments of breathiness but few have attempted to develop a formal model for predicting breathiness in vowels. There are two limitations to using correlation data alone for generating such a model. First, correlation does not indicate cause-and-effect relationships, and thus, correlational evidence alone is not sufficient to completely understand how listeners perceive breathiness. Second, correlation values do not show the nature of the psychophysical relationship. Thus, for example, a high correlation between aspiration noise level (AH) and perceived breathiness for a set of natural voices cannot establish how the magnitude of perceived breathiness might change if all factors, except AH, were held constant. Without a clear understanding of such psychophysical relationships, it is difficult to develop a computational model for voice quality perception.

One experiment to determine the psychometric functions for breathiness with varying signal-to-noise ratio (SNR) was reported by Hillenbrand [12]. In this experiment, a set of synthetic vowels varying in their spectral slope were evaluated by a panel of listeners using a direct magnitude estimation task. A nonlinear but monotonic increase in perceived breathiness was observed as the SNR was decreased. Interestingly, the effects of SNR on breathiness were found to be independent of spectral slope, a finding that is in contrast to others that have reported a positive correlation between spectral slope and breathiness [e.g., 13]. However, since recent research has demonstrated that both auditory measures and CPP are better correlated with perceived breathiness[4], it is advantageous to predict perceived breathiness using either of these measures instead of SNR. For this reason, the present experiment examined the psychometric functions for breathiness associated with increasing AH. However, unlike Hillenbrand [12], these changes were modeled as a function of varying NL/PL as well as CPP.

In other words, the goal of this study was to determine how perceived breathiness varied with NL/PL and CPP, and to develop a mathematical model to describe this relationship. This information is necessary to understand how listeners perceive dysphonic voice quality and to develop tools to quantify or predict changes in voice quality. In order to develop such tools, it is essential to generate computational models for dysphonic voice quality that not only discriminate vowels across specific voice quality dimensions, but also predict differences in magnitude within each dimension.

However, it is essential to remember that although changes in AH are likely to be one of the most important acoustic cues for breathiness, additional factors such as the vowel spectral slope, fundamental frequency, etc. may also affect perception of breathiness. It is likely that all of these factors are not adequately represented by the few acoustic and/or auditory- measures evaluated in this experiment. Therefore, the computational model developed in the present experiment should not be expected to provide a comprehensive description of breathiness. Instead, the functions developed in this experiment are merely the first step in generating a comprehensive model for the perception of dysphonic voice quality.

## 2. Methods

### A. Stimuli

Ten samples of the vowel /a/ (5 male and 5 female) were synthesized with a Klatt synthesizer using the LF model [14] as the source. These samples were based on natural speakers selected from a large database of disordered voices (Kay Elemetrics Disordered Voice Database). Speakers were selected to represent voices exhibiting a wide range of breathiness as judged by a panel of four listeners in a pilot experiment. Two of these listeners were very familiar with dysphonic voice quality (having worked with patients having dysphonic voices for > 5 years) whereas two others were graduate students in speech-language pathology. All listeners initially rated the breathiness in each voice using a five-point rating scale. Voices for which listeners showed the greatest agreement in ratings were selected for this experiment. The average fundamental frequency and the first three formants frequencies for each speaker were determined manually and used to create a synthetic copy of each sample. Other relevant synthesis parameters (open quotient, speed quotient, flutter, formant bandwidths) were subjectively adjusted to obtain a close match to the target speaker. The goal of the synthesis was not to obtain an exact match to the target speaker; rather, synthetic vowel stimuli were generated so as to obtain the same range of breathiness as that observed in the natural voices. The parameters used to generate these vowels are shown in Table 1. All stimuli were 500 ms in duration.

Next, a pilot listening test was conducted to determine the range of AH that resulted in perceptually “natural” vowels for each of these ten voice samples. Each synthesized vowel was used to generate a stimulus continuum that varied in AH from 0 dB to 80 dB in 5 dB steps. Three listeners rated these stimuli as either “natural” or “synthetic.” The range of AH levels that produced perceptually “natural” tokens for all three listeners was determined. Finally, this range was linearly divided to produce a continuum of 11 stimuli for each synthesized vowel sample resulting in a total of 110 stimuli for the main experiment (10 synthetic vowel continua X 11 stimuli/continuum). Most vowel samples were judged as “natural” for the entire range of AH (0 dB to 80 dB), except for the two samples that were based on speakers with the greatest breathiness. In these two synthetic vowels, reducing the AH levels below 55 dB resulted in stimuli that were described as “nasal” and/or “buzzy” by the listeners.

### B. Listeners

Ten young-adult listeners from the student body at the University of Florida were recruited to participate in this experiment. All listeners were native speakers of American English and had normal hearing as confirmed by a hearing screening (hearing thresholds below 20 dB HL at octave frequencies between 250 Hz and 8 kHz). Only listeners who had taken at least one class on voice disorders were recruited for this experiment. This step was taken to ensure that all listeners were familiar with breathy voice quality. However, one listener withdrew from the study without completing all test sessions and that listener’s data were discarded. This listener did not provide any further information about the reasons for discontinuing. Therefore, the data reported here represent the perceptual judgments made by nine listeners. All listeners were paid for participating in this experiment.

### C. Procedures

Listeners were tested in a sound-treated room in three 1-hour sessions over a two-week period. The stimuli were presented monaurally in the right ear at 75 dB SPL using ER2 ear inserts (Etymotic Research Inc.). Monoaural presentation was preferred to avoid complications related to binaural integration in the calculation of PL and NL. Ten blocks of stimuli, each consisting of three repetitions of a stimulus from a single talker, were presented to the listeners using an RP2 processor (TDT-System III; Tucker Davis Technology, Inc.). The order of the stimuli within each block and the order of the blocks were randomized across listeners. Stimulus presentation and listener responses were collected automatically using Sykofizx (Tucker Davis Technologies, Inc.).

Listeners were asked to estimate the breathiness of each stimulus in a direct magnitude estimation task. In this task, listeners assigned each voice stimulus a number that reflected the magnitude of breathiness of that stimulus. Listeners were instructed that a stimulus perceived to be twice as breathy as another should be assigned double the score as the first one. A value of zero was not permitted, but listeners were free to use any other number, including fractions, to estimate breathiness. Listeners were not allowed to repeat any stimuli and no anchor was provided for making these ratings. Listeners made their judgments by typing the desired numbers using a computer keyboard. Listeners were provided a maximum of 10 seconds to respond; however, almost all judgments were made in the first couple of seconds and none took longer than the maximum allotted time.

For statistical and modeling purposes, the geometric mean of the magnitude estimates for each stimulus, across listeners and across ratings, was determined. The geometric mean was preferred over the arithmetic mean since it is a better estimate of central tendency for ratio-level data as obtained in the magnitude estimation task. For ease of computation, these values were then transformed to a base-10 logarithmic scale. Since all the data obtained were between 100 and 1000, this transformation resulted in scores between 2 and 3. Log-transformed mean estimations were then translated or shifted so as to obtain values between 0 and 1 which was more convenient for determining the psychometric functions as described below. This translation was done merely to simplify the curve-fitting operations and it does not influence the general form of the resulting functions. The absolute values of various constants in the regression functions derived through this data were not critical because the data obtained in a magnitude estimation task are highly context dependent [15]. Rather, our interest was in the form of the psychometric function and how well it fit the observed perceptual data.

### D. Computation of Auditory and Acoustic Measures

The PL and NL for all voice stimuli were estimated using a loudness model described by Moore, Glasberg and Baer [16] and used in previous experiments to study breathy voice quality[4,5]. This model takes the spectrum of the periodic and aperiodic (i.e. aspiration noise) components as input to calculate the PL and NL values for that vowel. Both PL and NL are measures of loudness and are measured in units called “Sones”. Any sound perceived to have the same loudness as that of a 1 kHz tone at 40 dB SPL is defined to have a value of 1 Sone. A sound that is twice as loud would be assigned a loudness of 2 Sones whereas one that is half as loud would have a loudness of 0.5 Sones. Thus, this unit of measurement provides information about perceived loudness in ratio units of measurement.

The periodic and aperiodic components for each stimulus were isolated using the Klatt synthesizer as follows. The periodic component was determined by re-synthesizing each stimulus with the amplitude of aspiration noise set to zero but with the amplitude of voicing left at its original value for each synthetic stimulus. Likewise, the aspiration noise was isolated by re-synthesizing a copy of each stimulus with the amplitude of voicing set to zero, but with the aspiration noise level left at its original value. The ratio of NL to PL (NL/PL) obtained for each vowel stimulus was used as the independent variable in predicting the breathiness of that vowel.

CPP was computed for each stimulus using the method described by Hillenbrand [2]. Briefly, CPP is determined by first computing the cepstrum over a window of the signal. Next, a linear regression line is fit between the cepstrum level (dB) and quefrency (ms). The CPP is calculated as the difference between the level of the first cepstral peak and the level of the regression line at the same quefrency as the first cepstral peak. As with the auditory measures, a second set of functions using CPP for each vowel to predict the breathiness for each vowel continua were derived.

## 3. Results

### A. Reliability

Inter-rater reliability was estimated by calculating the average Pearson’s correlation between each listener’s mean ratings. The average correlation was found to be 0.73 (range: 0.30 – 0.94). Similarly, intra-rater reliability was determined by calculating the average Pearson’s correlation among the three ratings of each stimulus. This was found to average 0.87 and ranged from 0.77 to 0.97. Overall, these numbers suggest that listeners were able to complete the magnitude estimation task consistently, both individually and as a group. Therefore, data from all listeners was averaged and used to determine the general form of the functions relating changes in perceived breathiness resulting from changes in AH. No additional corrections or normalizations were applied to the magnitude estimates of breathiness prior to computing the predictive functions described below.

### B. Predicting breathiness from auditory measures: Modeling each vowel continuum independently

A series of curve fitting operations were completed to determine the relationship between NL/PL and breathiness for each vowel continuum. The decision to use the ratio of NL to PL as a predictor of breathiness was taken because a previous experiment had demonstrated a negative correlation between breathiness and PL and a positive correlation between breathiness and NL [4]. To obtain more precise regression functions, rather than trying to predict the whole dataset with a single function, we first attempted to fit the data for each synthetic vowel continuum individually. The curve fitting operations used and the rationale for each are described below.

The relationship between NL/PL and perceived breathiness for each of the ten synthetic vowel continua are shown in Figure 1. It was observed that all vowel continua demonstrated a monotonically increasing relationship between NL/PL and breathiness except for some stimuli that were perceived to have very low breathiness. The magnitude estimates of breathiness at very low NL/PL values showed random variation and were poorly correlated with changes in NL/PL as well as with changes in AH. To account for the non-monotonic relationship between NL/PL and breathiness at very low values of NL/PL, breathiness was modeled as a truncated and translated function of NL/PL. In this function, a threshold breathiness value was first determined for each vowel continuum. Stimuli with breathiness values at or below this threshold (henceforth referred to as $b_{\mathrm{TH}}$) were set to have the NL/PL value as zero and their breathiness values equal to $b_{\mathrm{TH}}$. The breathiness for the remaining stimuli (i.e. stimuli with breathiness greater than $b_{\mathrm{TH}}$) was modeled as a power function of NL/PL. Given this constraint, the following model was obtained:

$$\eta=\begin{cases} k(b-b_{\mathrm{TH}})^{p}, & b>b_{\mathrm{TH}} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where, $\eta$ is the value of NL/PL for the test stimulus, $b$ is the magnitude estimate of breathiness for the same test stimulus, $b_{\mathrm{TH}}$ is the magnitude estimate of breathiness at threshold NL/PL, $p$ is the power of the function relating breathiness to NL/PL, and $k$ is a constant. The values for these parameters were calculated for each of the ten vowel continua using an iterative function that determined the parameters of the regression equation that resulted in minimum error for that continuum. However, $b_{\mathrm{TH}}$ was not computed for FEML5 and MALE5 because these vowel continua did not contain any stimuli with NL/PL values close to zero as these were generated with a minimum AH level of 55 dB only.

Finally, to predict breathiness from NL/PL, a pseudoinverse of equation 1 was derived. A breathiness value of $b_{\mathrm{TH}}$ was assigned to NL/PL values of zero, and the values of $k$ and $p$ were used to predict breathiness for positive NL/PL values. The values of $k$, $p$, and $b_{\mathrm{TH}}$ that resulted in the smallest mean absolute error (MAE) were retained as the best set of parameters for that particular stimulus sample. The minimum, maximum, and average values obtained for these parameters over all the 10 samples are shown in Table 2. Based upon these data, the breathiness in a vowel varying in aspiration noise level may be predicted using a function of the following form:

$$b(\eta)=\left(\frac{\eta}{k}\right)^{1/p}+b_{TH} \tag{2}$$

Table 3 shows the MAE for the best fitting curve for each vowel continuum. The MAE were relatively small compared to the range of variation in breathiness for all ten talkers (approximately 5% of the range), suggesting that the truncated and translated power model was a good predictor for changes in breathiness resulting from variations in AH. In developing this model, the predictor variable (NL/PL) was not converted to a logarithmic scale because several stimuli demonstrated NL value of zero. Incidentally, even if the stimuli with NL=0 were discarded from this computation, the form of the function predicting breathiness from NL/PL retained a similar form.

### C. Predicting breathiness from auditory measures: Modeling all vowel continua with a single function

The long-term goal of this research is to create a model to predict breathiness for all voices. The following section describes the steps taken to derive a single function that might be successful in predicting breathiness for all the vowel continua tested in the present experiment. This model is hereafter referred to as the “unified” model.

The relatively large dispersion of various parameters ($b_{TH,}$ *k* and *p)* observed across the ten vowel continua suggested that additional parameters may be necessary to generate a satisfactory model for predicting breathiness from NL/PL. As shown in Figure 1, the psychometric functions for the male vowels are characterized by higher $b_{TH}$ and shallower slopes than for female vowels. This suggests that differences in pitch may contribute to the breathiness psychometric functions. Based on this observation, the computational model derived previously was modified to include fundamental frequency (F0) as an independent variable.

Since F0 was correlated with pitch in our stimuli and because pitch is better expressed in a logarithmic or quasi-logarithmic scale like the mel, ERB, or Bark rather than Hertz, we transformed F0 from Hertz to each of these three scales. It was observed that all three of these non-linear scales produced better results than the use of the Hertz scale. Of these three non-linear scales, the ERB provided the best results, though the difference in performance between the three was fairly small. Note that the bimodal distribution of F0 in our stimuli (because of the two genders) makes it difficult to clearly determine the benefit of using one scale over the others. Although more research is necessary to determine the ideal method and scale for computing pitch in dysphonic voices, the ERB scale has been used to report all measures of F0 in the present study.

A set of linear regressions was computed to determine the relationship between F0 ($\varphi$, in ERB) and the parameters $b_{TH}$, *k* and *p*. F0 was observed to account for a fairly large amount of variance in $b_{TH}$ (R2 = 0.63, p = 0.0185), and *p* (R2 = 0.53, p = 0.0414), but it did not show a significant correlation with *k* ($R^2 = 0.13$, p = 0.3804). However, the limited number of stimuli tested and the bimodal distribution of F0 may have reduced the accuracy of these results. Therefore, these findings should only be considered as an approximation at this the present time. The scatter plots for these regressions are shown in Figure 2.

Based on these findings, the $b_{TH}$, and *p* values for each talker were recomputed using the linear regression function based on F0. Since *k* was not found to be strongly F0-dependent, the mean value of *k* for all 10 stimulus continua ($k = 4.59$) was used on the unified model for breathiness. Thus, the unified computational model for breathiness is:

$$b(\eta,\varphi)=\left(\frac{\eta}{k}\right)^{1/p(\varphi)}+b_{TH}(\varphi) \tag{3}$$

where $b_{TH}$, and *p* are functions of F0($\varphi$) measured in ERB according to the formulae obtained using linear regressions.

Finally, to evaluate the success of this unified model, these equations were used to predict breathiness for stimuli in each of the ten vowel continua, and the MAE for each continuum was calculated. The results are shown in Table 3 and in Figure 3. The average MAE for the unified model (0.0598) is greater than that obtained using parameters based on individual curves (0.0244). However, despite this increase the MAE remains significantly smaller than the effective range of variation in breathiness reported by listeners (approximately 8% of the range).

### D. Predicting breathiness from CPP: Modeling each vowel continuum independently

The procedures used for predicting breathiness from NL/PL were also followed to develop a model for predicting breathiness based on CPP. Figure 4 shows the relationship between CPP and magnitude estimates of breathiness for each of the ten vowel continua. Note that the data for CPP has been plotted in an inverse scale (-CPP) to facilitate visual comparison with psychometric functions based on NL/PL (Figure 1). Figure 4 shows that unlike the NL/PL curves, the CPP curves do not demonstrate a common zero point. In other words, the minimum CPP level was different for each vowel continuum. Therefore, an additional parameter was required to model the CPP curves - a threshold ($c_{\mathrm{TH}}$) that determines the CPP value at which the power model starts to apply. The form of the equation to model CPP as a function of breathiness is:

$$-c=\begin{cases} k(b-b_{TH})^{p}-c_{TH}, & b>b_{TH} \\ -c_{TH}, & \text{otherwise,} \end{cases} \tag{4}$$

where $c$ is the CPP of the stimuli, $c_{TH}$ is the threshold value of CPP, and the other parameters are the same as those in equation (1).

As described previously, an iterative procedure was used to determine the values for the parameters $p$, $k$, $b_{\mathrm{TH}}$, and $c_{\mathrm{TH}}$ that resulted in minimum error for each vowel continuum. The minimum, maximum, and average values of these parameters as obtained from the ten vowel continua are shown in Table 4. Finally, the pseudoinverse function was computed to predict breathiness from CPP values:

$$b(c)=\begin{cases} \left[\frac{1}{k}(c_{\mathrm{TH}}-c)\right]^{1/p}+b_{\mathrm{TH}}, & c<c_{TH} \\ b_{TH}, & \text{otherwise,} \end{cases} \tag{5}$$

The set of parameters that produced the least MAE for each vowel continuum were retained as parameters for that model. Table 5 shows the MAE for the best fitting curve for each vowel continuum. A comparison between the MAE of the individual models in Tables 3 and 5 indicates that on average CPP was marginally better at predicting breathiness than NL/PL for each individual vowel continuum.

### E. Predicting breathiness from CPP: Modeling all vowel continua with a single function

The same procedures as used to generate the unified model using NL/PL were also used to generate a unified model using CPP. Figure 5 shows the relations between F0 (in ERB) and the parameters of the model. Only two of these parameters were observed to show a modest correlation with F0, though neither approached statistical significance: $b_{\mathrm{TH}}$ ($R^2 = 0.20$, $p = 0.2723$) and $k$ ($R^2 = 0.2062$, $p = 0.2583$). However, estimating these parameters in a F0-dependent manner provided the best overall results (i.e. lower overall MAE) and hence F0 was included in the computational model. This also permits easy comparison of performance for models based on CPP and auditory measures. The unified model using CPP may be described as:

$$b(c)=\begin{cases} \left[\frac{1}{k(\varphi)}(c_{\mathrm{TH}}-c)\right]^{1/p}+b_{\mathrm{TH}}(\varphi), & c<c_{TH} \\ b_{TH}(\varphi), & \text{otherwise}, \end{cases} \tag{6}$$

where $k$ and $b_{\mathrm{TH}}$ are now functions of pitch ($\varphi$). The predicted breathiness for all vowel continua using this equation is shown in Figure 6. The MAE of the unified model is shown in Table 5. A comparison with the unified model using NL/PL in Table 3 indicates that, once F0 is introduced into the model, the unified model using NL/PL is a better predictor of breathiness (average MAE = 0.0598) than the unified model using CPP (average MAE = 0.0822).

## 4. Discussion

The goal of the present experiment was to understand how breathiness in a vowel changes as a function of increasing aspiration noise levels in synthetic vowels and to determine how auditory measures (NL/PL) and a cepstral-domain measure (CPP) may be used to predict perceived breathiness. Aspiration noise levels were manipulated because a number of experiments have suggested this to be the primary cue for breathiness [1]. Listeners estimated the magnitude of breathiness for several stimuli and the resulting data were predicted either using NL/PL or using CPP. These two measures were selected because prior research found these to show the highest correlation with perceptual judgments of breathiness [4].

The auditory measure, NL/PL, is computed using a loudness model based upon psychoacoustic data [16]. Estimating partial loudness involves computation of masked loudness whereas that for noise loudness does not account for auditory masking. In both cases, the loudness model represents specific processes that are believed to occur during the auditory transduction process in an average listener. In contrast, the computation of CPP does not take any auditory-perceptual processes into account. Nevertheless, the high correlation between CPP and breathiness judgments in prior research [2,4,17] suggests that it may be a good candidate for predicting breathiness in speech. Therefore, two different models for breathiness – one using NL/PL and another using CPP – were computed.

For vowels that vary in AH only, breathiness was observed to increase with an increase in NL/PL except when NL/PL was close to zero (as seen in stimuli that had very low levels of AH). At higher values of NL/PL, the breathiness of a vowel is a power function of its NL/PL. For the ten voice continua studied in the present experiment this function has an average power value of 0.46 (inverse of 2.166 in Table 2) suggesting a compressive relationship between NL/PL and perceived breathiness. This compressive relationship holds even if the stimuli with NL=0 were discarded and the NL/PL for the remaining data were converted to a logarithmic scale. In this regard, much like other psychophysical continua, the relationship between aspiration noise level and perceived breathiness appears to follow Steven's Law [18]. Further, these effects appear to be pitch dependent since both, the threshold NL/PL and the power, were observed to differ for male and female voices. In general, male voices (stimuli with lower pitch) had a higher power and lower NL/PL thresholds. Therefore, when all other factors were held constant, voices with lower pitch tend to show a greater increase in breathiness for the same amount of change in NL/PL (which is related to changes in AH).

Breathiness in vowels could also be predicted by the CPP of that vowel. Breathiness is inversely related to the CPP and a power relationship appears to be the best fitting function. However, for the ten voices studied in this experiment, the average power was 1.1, a value that makes the relationship very close to a linear function. As with the auditory measures, stimuli with very low magnitudes of breathiness showed poor correspondence with CPP and the power function applied only to those stimuli that were above a threshold CPP. However, unlike the auditory measures, the threshold CPP values were highly variable across stimuli and neither the threshold CPP nor the power were found to have a clear pitch dependency. The unified model for predicting breathiness from CPP did not fit the perceptual data as well as that based on auditory measures. While CPP is significantly simpler to compute (relative to the auditory measures), a single function predicting breathiness from CPP did not generalize as well to all voice stimuli tested in this experiment.

Modeling the change in breathiness that occurs with increasing AH levels required defining a threshold value for all stimulus continua. Listener judgments of breathiness below this threshold (i.e. at very low values of NL/PL or CPP) remained highly variable, suggesting that listeners were unable to estimate the magnitude of breathiness consistently. This inconsistency in perceptual judgments may be explained by recent observation that the difference limens for aspiration noise are fairly large at low AH levels [19,20]. These experiments found that stimuli with low AH levels may need as much as a 15–20 dB change in AH before listeners can discriminate its breathiness. In contrast, the stimulus continua tested in this experiment varied the AH level only a maximum of 8 dB between successive stimuli. Since this change is less than the difference limens at low AH levels, the first few stimuli in several vowel continua may have perceptually equal breathiness.

The observation that the unified model based on auditory measures accounted for more variance in perceptual data than one based on CPP follows previous findings based on correlation data [4]. It is hypothesized that using an auditory processing model as a signal-processing front-end helps account for some of the non-linear processes inherent to the auditory perceptual process. Similar front-end processing is commonly employed in a number of applications that require mapping a physical acoustic signal to its percept, such as in automatic speech recognition, MP3 compression, etc. It is evident that the use of such a front-end is also advantageous for the study of voice quality perception and in the development of tools for its quantification.

While every attempt was made to obtain stable perceptual data for computing the psychometric functions, a number of factors may have adversely affected the MAE. First, listeners rated stimuli from each talker in a different listening “block”. Thus, all voices in a single continuum were compared against each other, but were never directly compared to stimuli from other voice continua. Since the absolute scores assigned in a magnitude estimation task are context dependent [15], the scores for one talker may not be directly comparable to that of another. Unfortunately, the unified model described here does not account for such context dependencies in perceptual judgments, and these differences may have contributed to an increase in the overall MAE. One solution to this problem is to obtain perceptual data using a matching technique that is relatively unbiased by context. A matching technique for assessment of breathiness has recently been described by Patel and colleagues [21] and may help improve the MAE further.

Second, even though the data clearly show a pitch dependency, the methods used to determine these relationships are only preliminary. This is because of two reasons – (1) the small number of stimuli tested and (2) the use of the ERB scale to estimate pitch for these stimuli. The small number of stimuli tested resulted in two clusters of low and high pitch (male and female speakers, respectively). Unfortunately, there is little systematic variation of pitch within each group of speakers. Therefore, although we obtained an overall positive correlation between pitch and various model parameters, it is not clear how pitch affects these parameters within a single gender. Further, it is possible that using a better method to compute the actual pitch (instead of F0) for each voice may further improve the results of the unified models. Finally, it is possible that there are additional parameters that affect the perception of breathiness that have not been incorporated in the models reported here.

The eventual success of any computational model can only be determined by testing its performance for novel stimuli. More specifically, the success of the computational model needs to be tested with different vowels, multiple complex synthetic stimuli and natural stimuli. However, the fact that natural stimuli co-vary in multiple acoustic features makes it difficult to isolate the contributions of one specific acoustic change to the overall voice quality percept. The synthetic stimuli tested in the present experiment allow the development of a simple model for the perception of breathiness in vowels that vary only in AH. Subsequent experiments will build upon this initial finding and will help generate a model for the perception of breathiness for natural vowels and running speech.

## 5. Conclusions

A model to predict changes in breathiness that result from variations in aspiration noise levels is proposed. A truncated and translated power model predicting perceived breathiness from NL/PL provided the best fit to the data. A high correlation between pitch and translation parameters of the model was observed. This finding led to the inclusion of pitch into the model. The resulting model suggests that breathiness ratings on a free magnitude estimation task are related to the noise-to-partial loudness ratio. The relationship between this ratio and breathiness beyond a specific threshold is best described by a power function. The power of this function is pitch dependent, but is generally less than one. In sum, a compressive relationship between NL/PL and perceived breathiness was observed.

## Acknowledgments

The authors would like to thank David Eddins for his valuable contributions and Sona Patel for the perceptual data. We would also like to acknowledge two anonymous reviewers for their help in improving this manuscript. This research was supported by a grant from NIDCD (DC006690).

## References

1. 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. [PubMed: 2137837]
2. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. J Speech Hear Res 1994;37:769–778. [PubMed: 7967562]
3. Buder, EH. Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990. In: Kent, RD.; Ball, MJ., editors. Voice quality measurement. San Diego, CA: Singular Publishing Group; 2000. p. 119-244.
4. Shrivastav R, Sapienza C. Objective measures of breathy voice quality obtained using an auditory model. J Acoust Soc Am 2003;114:2217–2224. [PubMed: 14587619]
5. Shrivastav R. The use of an auditory model in predicting perceptual ratings of breathy voice quality. J Voice 2003;17:502–512. [PubMed: 14740932]
6. Wolfe V, Martin D. Acoustic correlates of dysphonia: Type and severity. J Commun Disord 1997;30:403–415. quiz 415–406. [PubMed: 9309531]
7. Shoji K, Regenbogen E, Yu JD, Blaugrund SM. High-frequency power ratio of breathy voice. Laryngoscope 1992;102:267–271. [PubMed: 1545654]
8. Martin D, Fitch J, Wolfe V. Pathologic voice type and the acoustic prediction of severity. J Speech Hear Res 1995;38:765–771. [PubMed: 7474970]
9. Fukazawa T, el-Assuooty A, Honjo I. A new index for evaluation of the turbulent noise in pathological voice. J Acoust Soc Am 1988;83:1189–1193. [PubMed: 3281987]
10. Childers DG, Lee CK. Vocal quality factors: Analysis, synthesis, and perception. J Acoust Soc Am 1991;90:2394–2410. [PubMed: 1837797]
11. Hanson HM. Glottal characteristics of female speakers: Acoustic correlates. J Acoust Soc Am 1997;101:466–481. [PubMed: 9000737]

12. Hillenbrand J. Perception of aperiodicities in synthetically generated voices. J Acoust Soc Am 1988;86:2361–2371. [PubMed: 2970486]

13. Yanagihara N. Significance of harmonic changes and noise components in hoarseness. J Speech Hear Res 1967;10:531–541. [PubMed: 6081935]

14. Fant G, Liljencrants J, Lin Q. A four parameter model of glottal flow. Speech Transmission Laboratory Quaterly Report 1985:1–3.

15. Poulton, EC. Bias in quantifying judgments. Hove, U.K: Lawrence Erlbaum Associates Ltd; 1989.

16. Moore BCJ, Glasberg BR, Baer T. A model for the prediction of thresholds, loudness and partial loudness. Journal of Audio Engineering Society 1997;45:224–239.

17. Heman-Ackah YD, Heuer RJ, Michael DD, Ostrowski R, Horman M, Baroody MM, Hillenbrand J, Sataloff RT. Cepstral peak prominence: A more reliable measure of dysphonia. Ann Otol Rhinol Laryngol 2003;112:324–333. [PubMed: 12731627]

18. Stevens SS. On the psychophysical law. Psychological Review 1957;64:153–181. [PubMed: 13441853]

19. Shrivastav R, Sapienza CM. Some difference limens for the perception of breathiness. J Acoust Soc Am 2006;120:416–423. [PubMed: 16875237]

20. Kreiman J, Gerratt B. Difference limens for vocal aperiodicities. J Acoust Soc Am 2003;113:2328.

21. Patel S, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: A comparion of psychophysical methods. J Voice. In Press.

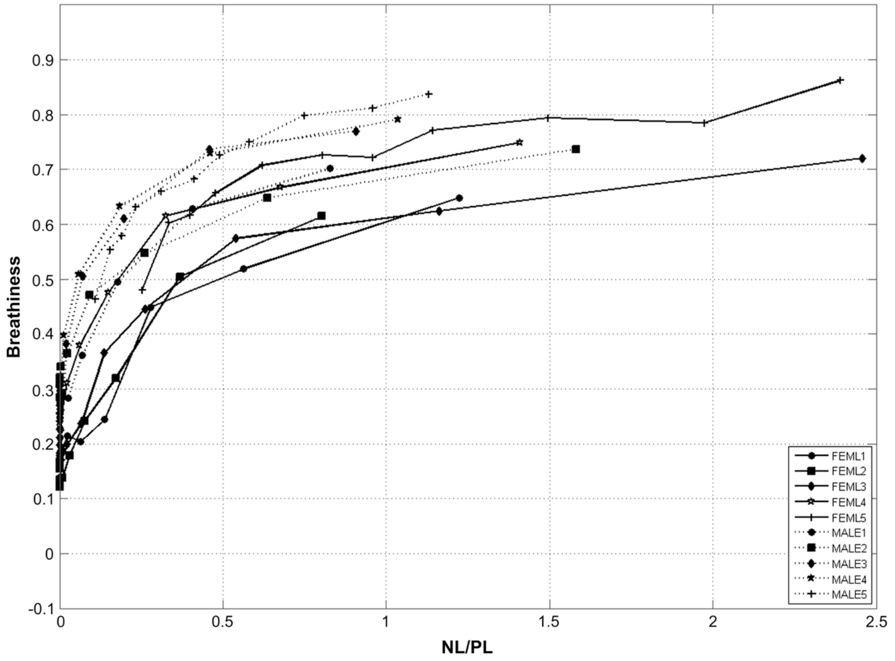

**Figure 1.**
Breathiness vs. NL/PL for each of the 10 vowel continua.

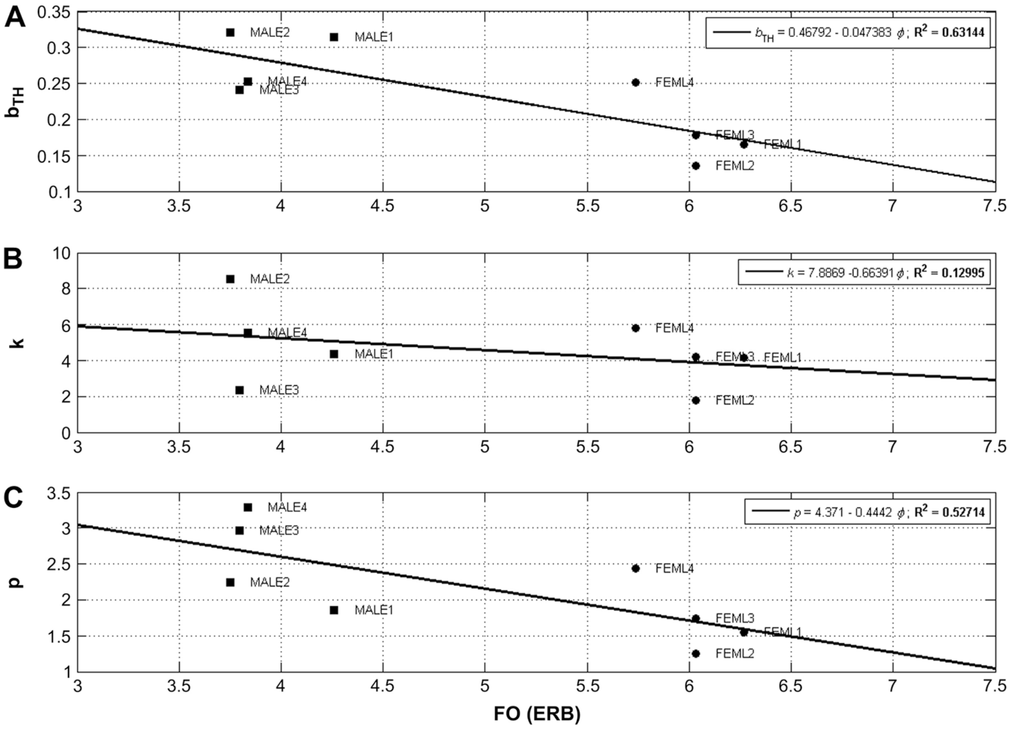

**Figure 2.**

Linear regression predicting (a) $b_{TH,}$ (b) *k*, and (c) *p* from pitch; used for generating the unified model predicting breathiness from auditory measures.

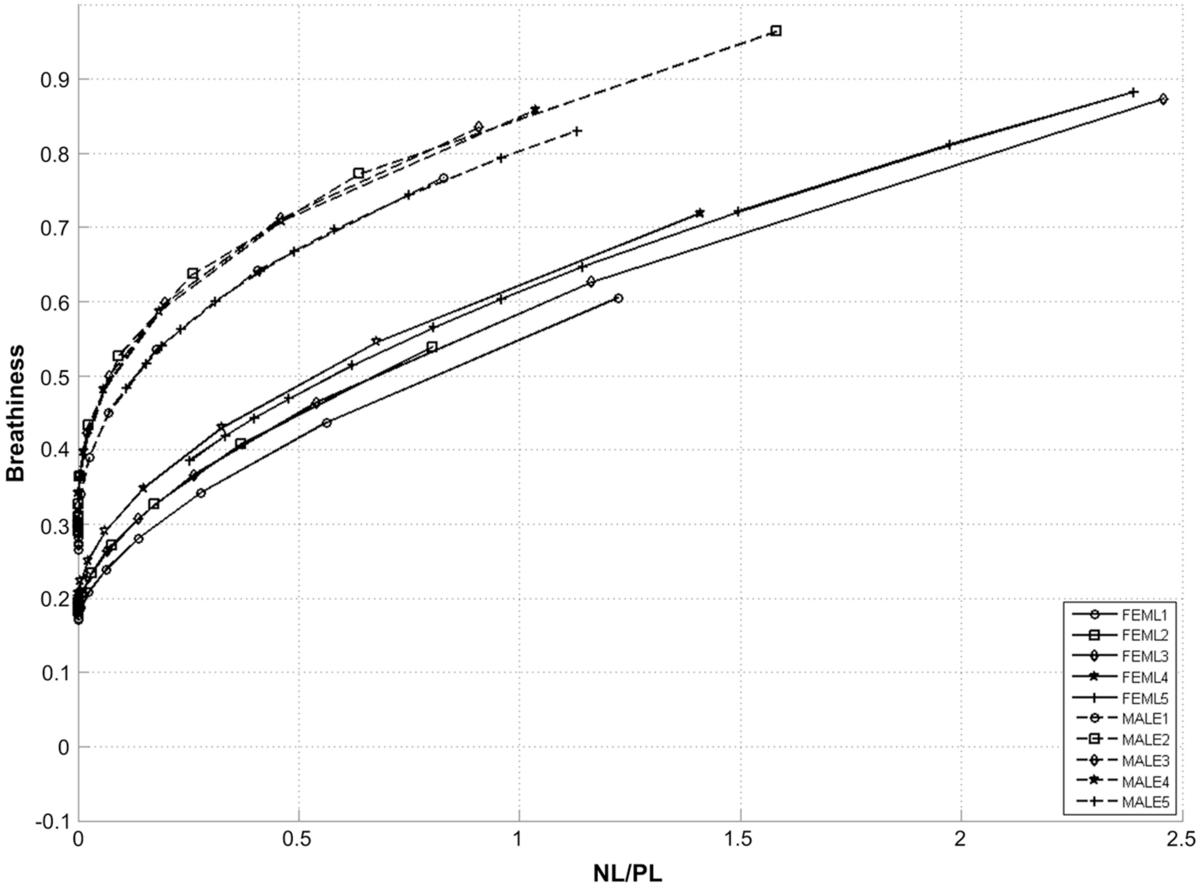

**Figure 3.**

Breathiness for all 10 vowel continua as predicted by the unified model based on auditory measures.

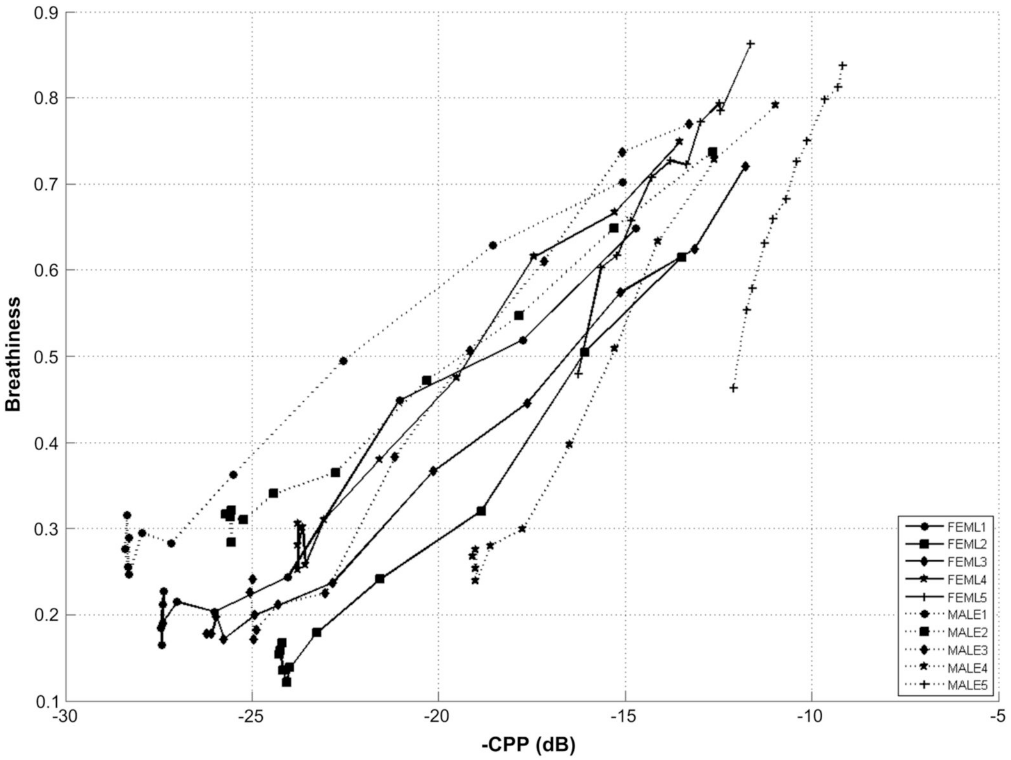

**Figure 4.**
Breathiness vs. CPP for each of the 10 vowel continua.

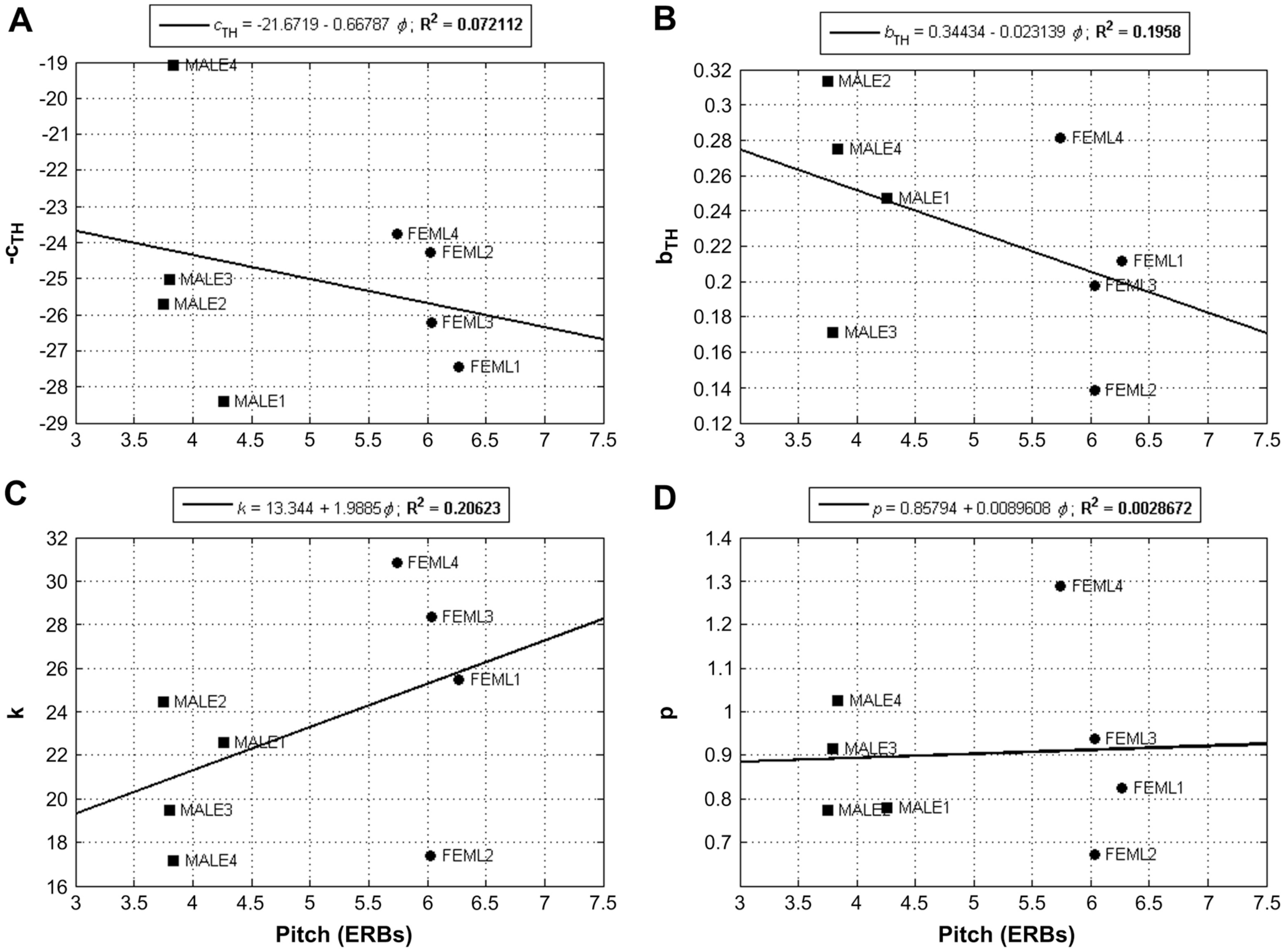

**Figure 5.**

Linear regression predicting (a) $c_{TH,}$ (b) $b_{TH,}$ (c) $k$ and (d) $p$ from pitch used in the unified model from CPP.

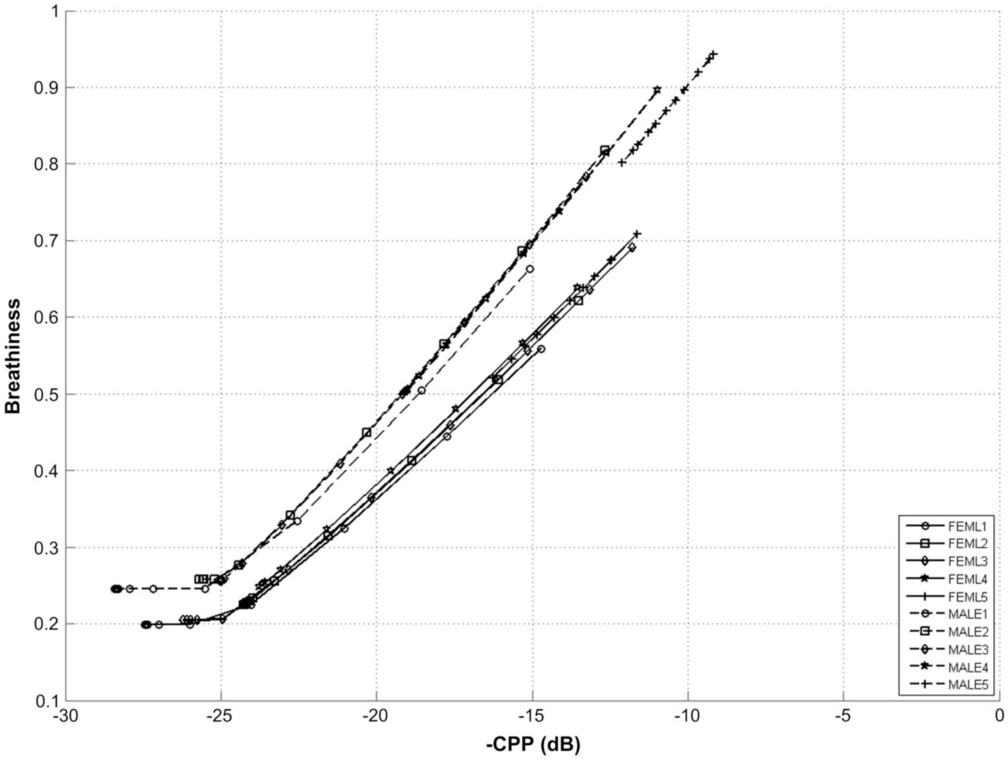

**Figure 6.**

Breathiness for all 10 vowel continua as predicted by the unified model based on CPP.

## Table 1

Klatt synthesizer parameters used to generate the 10 vowel continua.

| Parameter | Females | | | | | Males | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | 4 | 3 | 2 | 1 | 5 | 4 | 3 | 2 | 1 |
| F0 (Hz) | 200.7 | 195.5 | 209.1 | 209.0 | 220.4 | 134.4 | 117 | 115.5 | 113.7 | 133.1 |
| AV (dB) | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| OQ (%) | 85 | 75 | 65 | 55 | 40 | 85 | 75 | 65 | 55 | 40 |
| SQ (%) | 200 | 200 | 350 | 150 | 200 | 200 | 200 | 200 | 200 | 200 |
| TL (dB) | 40 | 30 | 20 | 10 | 0 | 40 | 30 | 20 | 10 | 0 |
| FL (%) | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| AH (dB) | 80 | 60 | 50 | 40 | 30 | 80 | 60 | 50 | 40 | 30 |
| FNP (Hz) | 180 | 280 | 180 | 180 | 180 | 180 | 180 | 180 | 180 | 180 |
| BNP (Hz) | 30 | 90 | 40 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| F1 (Hz) | 957 | 977 | 1050 | 759 | 891 | 814 | 586 | 814 | 559 | 661 |
| B1 (Hz) | 1000 | 800 | 600 | 400 | 200 | 1000 | 800 | 600 | 400 | 200 |
| F2 (Hz) | 1619 | 1356 | 1410 | 1333 | 1587 | 1473 | 1187 | 1473 | 1214 | 1122 |
| B2 (Hz) | 200 | 150 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |
| F3 (Hz) | 2877 | 2905 | 3000 | 2930 | 3083 | 2250 | 2463 | 2250 | 2340 | 2281 |
| B3 (Hz) | 250 | 200 | 300 | 300 | 300 | 250 | 200 | 300 | 300 | 300 |
| F4 (Hz) | 4274 | 4651 | 4000 | 4232 | 3870 | 3701 | 3405 | 3701 | 3383 | 4198 |
| B4 (Hz) | 300 | 250 | 400 | 400 | 400 | 300 | 250 | 400 | 400 | 400 |
| F5 (Hz) | 4883 | 4990 | 4990 | 4736 | 4761 | 4990 | 4194 | 4990 | 4396 | 4415 |
| B5 (Hz) | 350 | 300 | 500 | 500 | 500 | 350 | 300 | 500 | 500 | 500 |

## Table 2

Range of values obtained for the parameters of the model based on NL/PL.

| | $k$ | $p$ | $b_{TH}$ |
|---|---|---|---|
| Minimum | 1.802 | 1.256 | 0.1352 |
| Average | 4.591 | 2.166 | 0.2327 |
| Maximum | 8.519 | 3.285 | 0.3215 |

## Table 3

Mean absolute error (MAE) for the individual and group data using NL/PL. Separated models were not computed for FEML5 and MALE5 because these continua did not demonstrate a threshold point.

| Vowel Continua | Mean Absolute Error | |
|---|---|---|
| | Separated Models | Unified Model |
| FEML1 | 0.0261 | 0.0387 |
| FEML2 | 0.0212 | 0.0475 |
| FEML3 | 0.0448 | 0.0462 |
| FEML4 | 0.0214 | 0.0895 |
| FEML5 | ---------- | 0.1234 |
| MALE1 | 0.0224 | 0.0456 |
| MALE2 | 0.0159 | 0.0621 |
| MALE3 | 0.0263 | 0.0699 |
| MALE4 | 0.0173 | 0.0336 |
| MALE5 | ---------- | 0.0416 |
| Average | 0.0244 | 0.0598 |

### Table 4

Range of values obtained for the parameters of the model based on CPP.

|  | $k$ | $p$ | $b_{TH}$ | $c_{TH}$ |
|---|---|---|---|---|
| Minimum | 17.15 | 0.6729 | 0.1382 | 18.98 |
| Average | 23.22 | 0.9024 | 0.2295 | 24.86 |
| Maximum | 30.84 | 1.289 | 0.3133 | 28.29 |

## Table 5

Mean absolute error (MAE) for the individual and group data using CPP.

| Talker | Mean Absolute Error | |
|---|---|---|
| | **Separated Models** | **Unified Model** |
| FEML1 | 0.0199 | 0.0386 |
| FEML2 | 0.0199 | 0.0684 |
| FEML3 | 0.0170 | 0.0189 |
| FEML4 | 0.0221 | 0.0609 |
| FEML5 | ---------- | 0.0940 |
| MALE1 | 0.0209 | 0.0619 |
| MALE2 | 0.0096 | 0.0458 |
| MALE3 | 0.0240 | 0.0441 |
| MALE4 | 0.0315 | 0.1986 |
| MALE5 | ---------- | 0.1903 |
| Average | 0.0206 | 0.0822 |