The perception of breathiness: Acoustic correlates and the influence of methodological factors

Ilse Bernadette Labuschagne* and Valter Ciocca†

The University of British Columbia,
Friedman Building, 2117 Wesbrook Mall, Vancouver, Canada, V6T 1Z3

Abstract: Research on the acoustic correlates of breathiness has been plagued by a lack of consistent findings across studies and low intra- and inter-rater agreement. Sources of variability can arise from different sources including: differences in stimulus types (recorded or synthesized); differences in speaker groups (for recorded stimuli) or in synthesis parameters (for synthesized stimuli); differences in experimental methodologies (task type, number of repetitions, listener backgrounds and experience). This review discussed these sources of variability, and described solutions that have the potential to address the variability and the inconsistencies often reported in the literature. A critical appraisal of the evidence about the relative importance of various acoustic measures resulted in the identification of measures of periodicity, noise content, and high-to-low frequency energy as the most likely acoustic correlates of breathiness.

Keywords: Breathiness, Acoustic correlates, Voice quality, Acoustic analysis

PACS number: 43.70, 43.71 [doi:10.1250/ast.37.191]

1. CHALLENGES IN VOICE QUALITY RESEARCH

Breathiness is the voice quality that is associated with an audible escape of air through the glottis during phonation, due to voice pathology or habitual speaking patterns [1–3]. In other words, breathiness is a perceptual quality of the voice that is produced by physiological processes that take place in the larynx during phonation. As a voice quality, breathiness is one of the many dimensions that allows listeners to discriminate different voices from one another [4].

Because it is the result of physiological processes that occur during phonation, breathiness is primarily investigated by focusing on the properties of voiced speech. Although one could observe the effects of breathy voice production on any voiced sound [5], most researchers have focused on the acoustic and perceptual characteristics of vowels. For this reason, this article will review primarily findings of investigations of breathy vowels. To discover the acoustic correlates of the perception of breathiness, researchers initially conducted experiments using a variety of experimental tasks and stimuli [1,4,6,7]. Perceptual data from these studies were plagued by large amounts of variability, resulting in inconsistent findings. Many researchers then attempted to identify the sources of variability in the findings, such as the type of tasks used, and the differences among individuals and listener group among studies, in order to attempt to improve predictive models of the perception of breathiness [8–13]. For example, Kreiman et al. [10] suggested that sources of variability can be introduced into perceptual data due to listener differences and task factors. They provided a framework for explaining and controlling these sources of variance. For example, they suggested that rating protocols include a reference stimulus in order to reduce inter-listener variability (see Sect. 3 for a comprehensive discussion of this recommendation).

Because a researcher can make several choices with regard to psychophysical experiments (choices among specific stimuli, tasks and procedures, listener groups and data analysis procedures), it is possible that sources of variability related to each of these choices affect the experimental outcomes. Decisions about the experimental setup should therefore be motivated by the specific research questions, since each of these choices potentially has advantages and disadvantages. Consider, for example, a researcher who intends to identify acoustic correlates of breathiness in an exploratory study. She/he may decide to use all the recordings from a pathological database as stimuli for an experiment. This choice has the advantage...
that the stimuli are representative of the population of pathological voices found in a clinical setting. However, using these stimuli may have the disadvantage of introducing sources of variability that can be difficult to identify or control, such as multiple acoustic correlates of various speech disorders that often co-occur with breathiness. Alternatively, a researcher may choose to synthesize stimuli in order to independently control single acoustic variables associated with breathiness, at the cost of sacrificing the naturalness of the stimuli and of weakening the external validity of the outcomes.

The current literature review will summarize the most common types of stimuli, experimental tasks, and listener characteristics that have been used to investigate breathiness, as well as other voice qualities. A summary of the results and of the conclusions that have been drawn from these studies will be presented in the following sections.

## 2. SELECTION OF THE STIMULI

In perceptual and acoustic studies on breathiness, stimuli can be selected from recordings produced by speakers with or without voice disorders, or synthesized to conform to a specific set of criteria suited to the research question.

### 2.1. Recordings of Natural Utterances

The multidimensional nature of different voice qualities is entirely present in recordings of natural utterances, which contain all the acoustic features of both normal and pathological voices. Such stimuli are useful when conducting exploratory studies to identify either the acoustic correlates of breathiness, or the perceptual dimensions that are important for breathiness (by carrying out, for example, multidimensional scaling (MDS) studies). Another advantage of using natural utterances is that the perception of breathiness can be investigated in the presence of other voice and speech characteristics, as they would be present in real life stimuli.

Natural voice recordings have been selected for research studies by using various selection criteria. For example, Castillo-Guerra and Ruiz [14] used a comprehensive set of stimuli from two pathological databases to investigate the acoustic correlates of breathiness. All the stimuli from the databases were included in their study. Although the authors did not give an explanation for this selection, they presumably made this choice in order to represent the whole range of pathological voices. By contrast, other researchers adopted inclusion or exclusion criteria for their investigations. Some researchers selected utterances that encompassed a full range of breathiness (from mild to severe), while reducing the influence of other qualities not under investigation [6,11,15]. For example, Shrivastav [16] chose a set of stimuli from the Sataloff/Heman-Ackah voice database. These stimuli were selected in order to maximize the variance in breathiness ratings and minimize the variance in roughness and strain; this selection was based on “grade, roughness, breathiness, asthenia, strain” (GRBAS) ratings [17]. Additionally, they also excluded voices with “bifurcations and chaos” due to difficulties with measuring acoustic characteristics for these voice types. Although these selection criteria were consistent with the goals of the study, one should be aware that the use of inclusion/exclusion criteria affects the external validity of a study.

Another consequence of using specific selection criteria may be a bias in the data that are obtained from perceptual studies on voice quality. In an MDS study, Kreiman et al. [9] found that acoustic characteristics that varied more within a set of voice recordings were more likely to be used as perceptually salient cues by listeners. They also found that listeners used different perceptual strategies for evaluating pathological and normal voices, showing that the choice of stimuli has an influence on perceptual data.

Some researchers have chosen to apply a gender criterion in their research by choosing stimuli from a single gender (male or female) to control the amount of variability in breathiness ratings [16,18,19]. Studies that have directly investigated the effect of gender on breathiness have found contrasting results. For example, Hillenbrand [1] showed that male speakers had higher breathiness ratings compared to female speakers for simulated, very breathy utterances — but not for mildly breathy and non-breathy vowels. By contrast, Klatt and Klatt found that females produced breathier utterances than males, on the basis of both acoustic and perceptual measures; they also observed that there was large variability within each gender group [7]. Klatt and Klatt’s findings suggest that female speakers may be inherently more breathy than males, but more extensive investigations need to be carried out to provide strong evidence for a gender effect, particularly for speakers with voice disorders. As with the previously discussed stimulus selection criteria, gender is a factor that needs to be taken into consideration for the selection of stimuli in studies of breathiness.

Another important factor that may have an effect on the external validity of studies on breathiness is whether sustained vowels or vowels isolated from continuous speech are used as stimuli. The few studies that investigated breathiness using continuous speech or entire words, rather than focusing only on vowel sounds, limited their acoustic analyses to long-term average spectra measures [4,20]. Although the acoustic measures normally used to characterize sustained vowels can be obtained for vowels extracted from continuous speech, the vast majority of studies on breathiness employed sustained vowels [1,6,8–10,15,16,18–20]. Hillenbrand and Houde [6] com-
pared the acoustic correlates of breathiness for sustained vowels and for vowels isolated from continuous speech. They found that correlations between breathiness ratings and acoustic measures were largely similar for both types of stimuli. Although it may be argued that continuous speech has more external validity, there is little evidence that, as a general rule, it should be preferred to sustained vowels in breathiness research. Furthermore, speech pathologists often use sustained vowels in clinical settings for diagnostic and treatment purposes.

Another consideration in the selection of vowel stimuli concerns vowel quality. Most studies used the vowel /a/ [6–11,13,15,16,18,20–23]; in the few cases when the effect of different vowels was investigated, results differed between perceptual and acoustic measurements. Hillenbrand et al. [1] investigated the effect of vowel quality (/a/, /i/, /æ/ and /o/) and found no significant effects of vowel quality on breathiness ratings. They did not compare differences in acoustic measures associated with breathiness among vowels. By contrast, Hanson [24] measured the acoustic characteristics of different vowels (/æ/, /ʌ/ and /ə/), and found that vowel type had a significant effect on two acoustic measures often investigated in breathiness research: 1. the amplitude difference between the estimated first and second harmonic of the glottal waveform (H1* – H2*), and 2. the amplitude difference between the first harmonic of the glottal waveform and the third formant amplitude, corrected for the effects of formants 1 and 2 (H1* – A3*). With various studies attempting to quantify sources of variability, it is surprising that the effects of vowel quality on breathiness ratings have not been investigated more extensively.

Signal processing characteristics of the stimuli, such as low-pass filtering (LPF) and sampling frequency selection, may at first seem trivial when discussing stimuli with various degrees of breathiness. However, these properties have important consequences for the acoustic characteristics associated with breathiness. For example, Hammarberg et al. [4] showed that the amount of high frequency energy between 5–8 kHz was approximately equal to the energy between 2–5 kHz for breathy utterances; non-breathy voices exhibited lower energy between 5–8 kHz compared to 2–5 kHz. Studies that use sampling frequencies as low as 10 kHz [21,25], or LPF cut-offs as low as 4.5 kHz [24], eliminate or reduce the energy at 5–8 kHz, thereby removing potentially important cues to breathiness. Therefore, the use of LPF or low sampling frequencies should either be justified or avoided to ensure that there is adequate representation of high-frequency cues to the perception of breathiness.

Recordings have their place in voice quality research as naturally-occurring, multidimensional, but non-deterministic stimuli. They present acoustic characteristics as they would appear in real life, but have the disadvantage that specific cues cannot be manipulated independently of other cues.

2.2. Synthesized Sounds

The advantage of using synthesized stimuli for research on voice quality is that it is possible to precisely control either acoustic or articulatory parameters that are associated with the perception of breathiness. Systematic changes to the synthesis parameters of acoustic/articulatory models of voice production have measurable effects on the acoustic properties of the stimuli. Such stimuli can then be used to further develop models of the articulatory or perceptual processes associated with breathiness. A related advantage of these models is that they are quasi-deterministic, and will effectively yield the same utterance given the same set of model parameters.

There are two main types of voice production models, mathematical and articulatory/kinematic models. Mathematical models, such as the KLatt synthesizer [26] or the Lijencrants-Fant (LF) glottis model [27], use equations to indirectly represent the functionality of the vocal tract and the larynx. If a researcher wanted to synthesize a breathy vowel, he/she would have to select the parameters of the larynx model that would generate a glottal waveform that represents a breathy voice. For example, varying the LF model parameter “t0” will result in a glottal waveform that has a more or less gradual closing phase, thus generating a glottal waveform that is typical of breathy or non-breathy voices, respectively (see Fig. 1). In other words, using these models a researcher can adjust the properties of the glottal waveform to simulate the vocal fold movement associated with a breathy utterance.

Articulatory or kinematic models directly represent the physical characteristics of the larynx and of the vocal tract in order to simulate the anatomical-physiological processes of voice production [28,29]. Examples of such models are Titze’s glottis model [30], and the finite element vocal tract model in VocalTractLab [31]. These models can be used to synthesize breathy utterances by manipulating the modelled physiological processes directly, unlike the more abstract parameters manipulated in mathematical models. For example, one of the model parameters in Titze’s glottis model [30] is the abduction quotient, Qa (the distance between the top portion of the vocal fold and the glottis midline—see Fig. 2). An increase in this articulatory parameter leads to a vocal fold vibration with a longer open phase, resulting in breathier stimuli. Unlike mathematical models, articulatory/kinematic models manipulate characteristics of the larynx and of the vocal tract; these changes can drive a synthesis model to generate acoustic waveforms that simulate normal or disordered voice qualities [31]. One of the advantages of using articulatory/kinematic...
models is that systematic manipulations of model parameters may elucidate the threefold relationship between anatomical/physiological processes, the acoustic characteristics of breathy signals, and perceptual ratings of breathiness [23].

3. EFFECTS OF TASK AND LISTENER CHARACTERISTICS ON VOICE QUALITY JUDGMENTS

Decisions about the selection of the methodology of a study depend on the questions that researchers intend to answer; such decisions also have important consequences for the outcomes of the research. In voice quality research, typical methodological decisions include: 1. the selection of the experimental task (for example, a rating or a matching task); 2. whether an external reference stimulus should be included within a trial; 3. how many times each trial should be repeated within an experiment. Task-related decisions are not made independently of the research questions, of the stimuli used, and of the analysis of the data. For example, synthesized stimuli used in rating tasks may require larger acoustic differences compared to two-alternative forced-choice (2AFC) tasks for breathiness discrimination. As another example, correlations can be calculated between rating data and acoustic measurements, while data from a same-different (SD) task may be analyzed using a MDS analysis. Rating tasks are time efficient and better reflect judgments made in a clinical setting. SD tasks can reveal cue salience without explicitly defining perceptual qualities such as “breathiness,” which may be difficult to define and explain to some groups of naïve listeners. Methodological decisions can also affect intra- and inter-rater agreement and reliability in studies of voice quality. Maximizing agreement and reliability of judgments within and across listeners is important if research outcomes are to be valid and applicable to clinical settings. For example, one would want to ensure that judgments of the same voice samples by one clinician are: 1. stable over time, and 2. in agreement with judgments made by other clinicians. These methodological factors, as well as listener characteristics and selection criteria, will be discussed in the following paragraphs.

The most commonly used task in voice quality research is the perceptual rating task using equal-appearing interval (EAI) scales. A review by Kreiman et al. [10] showed that 82.5% of the selected studies used this type of rating scale. Other tasks include direct magnitude estimation (DME), visual analog (VA) rating scales, matching tasks, 2AFC tasks and SD tasks. Comparing data across studies that use different tasks is not straightforward. For example,
Patel et al. [11] found evidence that a matching task had higher inter-rater reliability than EAI and DME tasks for the perceptual evaluation of breathy vowels. However, these results should be interpreted with caution because of the small dataset employed in this study. Kreiman et al. [10] found that perceptual roughness ratings differed across experimental blocks for an EAI task, but not when listeners used a VA scale. These findings show that VA scales might be preferable to EAI scales for judging roughness. These direct comparisons among different tasks used for the evaluation of the same stimuli imply that attempts to compare results across studies using different task types are beset with serious difficulties, especially given that different stimuli are typically used across studies.

Kreiman et al. [10] remarked that perceptual rating tasks require listeners to compare the stimuli with an "internal reference" that is retrieved from memory (the internal reference is a mental representation of a perceptual quality). They observed that listeners’ internal references vary over time and differ among individual listeners. Therefore, in their conceptual framework of voice quality, Kreiman et al. proposed that the internal reference should be replaced with an external reference stimulus. Subsequent studies supported the idea that providing a reference stimulus, or using a matching task (which by its nature includes a reference or standard stimulus) improves inter- and intra-rating reliability compared to rating tasks without an external reference [11, 22]. It is important to consider that using tasks with an external reference in a clinical setting may not be feasible because such tasks are more time-consuming than rating tasks. Moreover, the selection of an appropriate external reference to be used in clinical settings is far from trivial.

As explained above, the use of a reference stimulus or of a matching task can result in improvements in inter- and intra-rater variability. There are other ways to minimize variability in perceptual judgments of voice quality without using a reference stimulus. Shrivastav et al. [13] noted that prior studies had typically employed measurements of individual stimuli, and used raw (non-standardized) rating scores for data analysis (see, for example, [8, 9]). They found that the amount of inter-rater reliability and agreement improved if ratings made by each listener for each stimulus were averaged across several blocks, and if they used standardized, instead of raw, rating scores. On the basis of these results, Shrivastav et al. concluded that individual perceptual spaces may differ less than previously suggested, and that it is possible to minimize variability and increase agreement among participants through the selection of appropriate experimental tasks and data analysis procedures.

In addition to task-related factors, Kreiman et al. [10] discussed the effect of listener characteristics on perceptual judgments of voice quality. They concluded that there is considerable inter-judge variability in perceptual judgments of voice quality, especially among expert listeners. In a previous MDS study, Kreiman et al. asked expert and naïve listeners to judge the dissimilarity of normal and pathological stimulus pairs that differed in several voice qualities [8, 9]. They found that expert and naïve listeners used different perceptual strategies for making their judgments. For example, both listener groups perceived pathological voices along three dimensions, but the specific acoustic cues that correlated with each dimension differed between groups. For both groups, pathological and normal voices were judged according to different perceptual spaces. Specifically, the fundamental frequency ($F_0$) correlated with the first dimensions of the perceptual spaces for both voice types, but the perceptual space for pathological voices included three dimensions, while the space for normal voices comprised only two dimensions. Finally, experts showed more varied individual perceptual strategies compared to the naïve listeners, who were more uniform in the way they judged the voice qualities of the utterances. These somewhat surprising findings mean that, for reasons that are not well understood, larger amounts of training and experience in voice quality judgments result in larger idiosyncratic listener bias.

Some researchers have attempted to reduce the variability introduced by listeners with different levels of experience with pathological voices, and with different language backgrounds, by using a uniform listener group. For example, Shrivastav and colleagues selected native speakers of American English from among university students in the same voice training program as their participants [11–13, 15, 18, 22]. It is unclear whether different language backgrounds affect variability in perceptual ratings of voice quality [32, 33]. It is possible that using listeners with similar levels and types of experience with pathological voices may reduce variability in perceptual outcomes. However, to our knowledge this hypothesis has not been tested empirically. While uniformity in participant characteristics may reduce variability in perceptual outcomes, the selection of homogenous groups also reduces the external validity of the findings as experimental outcomes may only apply to specific listener groups.

After having reviewed the methodological aspects of stimuli, tasks and listener characteristics that affect voice quality judgments, we now turn to the findings about the acoustic cues that are associated with perceived breathiness.

### 4. ACOUSTIC CORRELATES OF BREATHINESS

Acoustic correlates of breathiness can be loosely categorized according to the main effect they have on the following acoustic properties of breathy signals: spectral
shape, noise content, periodicity of the signal, and amplitude perturbation (dynamic amplitude changes). While there is some clear overlap among these categories (for example, the amount of noise will also affect the periodicity of the signal), the individual measures have been categorized according to properties of the signal that each correlate affects most directly. Each of the categories of acoustic correlates will be discussed in the following subsections.

4.1. Measures of the Spectral Shape of Breathy Vowels

Measures of the spectral shape of the signal include: measures of the difference in amplitude between the first and second harmonic \((H1 - H2)\) \cite{3,6,9,15,18,19,23,34}; comparisons of the spectral energy in different frequency bands \cite{4,20,23,24}; measures of spectral tilt \cite{1,6}; measures related to formant magnitudes and bandwidths \cite{3,7,15,18,24}.

Perhaps the most often investigated spectral characteristic is \(H1 - H2\). Fischer-Jørgensen \cite{35} and Ladefoged \cite{36} investigated the acoustic properties of breathy vowels in Gujarati and the Khoisan language, !Xóõ, respectively. They were the first researchers who observed that breathy vowels have larger amplitude of the first harmonic \((H1)\) than non-breathy vowels. Subsequent studies measured this correlate of breathiness as the difference in dB) between the amplitudes of the first and second harmonics, \(H1 - H2\). Klatt and Klatt \cite{7} measured \(H1 - H2\) values at isolated vowel midpoints from continuous speech utterances, and found a strong correlation \((r = 0.83)\) between \(H1 - H2\) and breathiness ratings made on a 7-point EAI rating scale. Hillenbrand et al. \cite{1} found a lower correlation coefficient \((r = 0.66)\) between \(H1 - H2\) measured at sustained vowel midpoints and breathiness ratings using a DME scale. Therefore, both studies found strong correlations between \(H1 - H2\) and breathiness, in spite of differences in a number of methodological characteristics including the number of listeners (Klatt and Klatt tested three listeners compared to Hillenbrand et al. who used 20 listeners) and the type of task (EAI vs. DME). A subsequent study by Hillenbrand and Houde \cite{6} also investigated the strength of the relationship between \(H1 - H2\) and breathiness. Rather than calculating \(H1 - H2\) at the vowel midpoints, they used an automatic algorithm that estimated \(H1\) and \(H2\) every 10 ms over the duration of each utterance (the sustained vowel /a/). They then averaged the \(H1 - H2\) values thus obtained for each utterance. Correlations of 0.7 and 0.66 were obtained between \(H1 - H2\) and breathiness ratings for samples of pathological and normal voices, respectively. Hillenbrand and Houde \cite{6} also calculated \(H1 - H2\) for four different vowels (/a/, /i/, /e/ and /o/) excised from a reading of the “rainbow passage” by speakers with voice pathologies. \(H1 - H2\) for these vowels was calculated from hand-selected midpoints for each of the vowels, and then averaged across vowels. For these vowels, the correlation between \(H1 - H2\) and breathiness ratings was lower \((r = 0.52)\) than the correlations obtained for sustained vowels.

Using an average of \(H1 - H2\) values from different vowels is problematic because the specific frequency and the amplitude of the formants of each vowel have an influence on the amplitudes of \(H1\) and \(H2\). In order to circumvent this problem, some authors have modified values of \(H1 - H2\) by effectively removing the effect of the vocal tract configuration. The modified \(H1 - H2\) measure \((H1^* - H2^*)\) is calculated from the spectrum of the estimated glottal waveform itself, rather than the radiated speech sound \cite{19,23,24}. The asterisks are used to indicate that these harmonics pertain to the glottal waveform. \(H1^* - H2^*\) can therefore also be calculated directly from the glottal waveform in synthesis studies or electroglottographic measurements. Samlan and Story \cite{23} provided evidence about the relationship between \(H1^* - H2^*\) and ratings of breathiness. They used a finite element model of the vocal folds, in combination with a vocal tract model, to synthesize /a/ vowels with various amounts of vocal fold separation during phonation. They found that, as vocal fold separation increased, the perception of breathiness also increased. Vocal fold separation accounted for 61.4% of the variance in breathiness ratings. However, \(H1^* - H2^*\) did not increase monotonically as a function of vocal fold separation. Instead, as vocal fold separation was increased, \(H1^* - H2^*\) first increased (for moderately breathy signals), but then decreased with further vocal fold separation (for severely breathy signals). This finding suggests that \(H1^* - H2^*\) is not monotonically related to breathiness ratings.

The studies reviewed so far reported that measures based on \(H1 - H2\) were found to be strongly correlated with breathiness (but see findings by Hartl et al. \cite{20} for contrasting findings). It is worth noting that using \(H1^* - H2^*\) as an acoustic correlate of breathiness may be unwarranted for two reasons. First, listeners hear the waveform that has been filtered by the vocal tract and radiated at the lips, not the glottal waveforms. Secondly, several authors have questioned the accuracy of glottal inverse filtering used to recover glottal waveforms from the recorded utterances (see Alku \cite{37} for a review). Nonetheless, \(H1^* - H2^*\) has proved to be useful in other ways. For example, preliminary evidence has shown that this measure may be used to predict a speaker’s glottal configuration during an vowel utterance (for example, an incomplete glottal closure) \cite{24}.
frequency region. Another class of acoustic correlates of breathiness reflects the ratio of the energy present within low and high frequency regions. Hillenbrand and Houde [6] defined the high-to-low frequency energy ratio (H/L) as the ratio of the energy above 4kHz to the energy below 4kHz. Significant moderate to strong correlations were found between breathiness ratings and the H/L for sustained vowels (r = 0.64), for vowels isolated from continuous speech of speakers with voice pathology (r = 0.84), and for sustained vowels from normal speakers (r = 0.51). These findings indicate that an increase in high frequency energy (higher H/L ratio) was associated with higher breathiness. Hartl et al. [20] found that the difference in energy between a low frequency spectral band (B0, 0–0.4kHz) and three high frequency spectral bands (B1, 0.4–2kHz; B2, 2–5kHz; B3, 5–8kHz) was significantly lower for a group of speakers with unilateral vocal fold paralysis, compared to speakers without voice pathologies. They also found that the group with voice pathologies had significantly higher breathiness ratings compared to the control group. Other authors have compared the energy in high and low frequency bands using various frequency boundaries, and replicated the above results [1,4,20,24]. Samlan and Story [23] calculated the acoustic measure B0 – B2 for various amounts of vocal fold separation in their finite element glottis model. They found that B0 – B2 significantly contributed to the amount of variance explained in breathiness ratings, although it was not a major factor. Similar to their findings on H1* – H2*, Samlan and Story also found that B0 – B2 was non-monotonically related to breathiness ratings. In summary, findings from these spectral ratio measures show that high-frequency energy generally increases as voices become breathier, but that evidence has not always been consistent with this conclusion [23].

Measurements of spectral shape that compare energy at high and low frequency bands require arbitrarily defined frequency boundaries. It is possible to compute a measure of spectral slope that does not require such boundary definitions. The breathiness index (BRI) is a measure of spectral slope that is calculated as the energy of the second derivative of a pre-emphasized speech signal, compared to the energy of the underived pre-emphasized signal [2]. As the high-frequency energy increases, the value of BRI will increase. Hillenbrand and colleagues found moderate to strong correlations between breathiness ratings and the BRI [1,6]. However, not all findings on spectral slope measures are in agreement. The term “spectral tilt” is used as a parameter in synthesis studies, and directly reflects the overall spectral slope of the speech signal. In two synthesis studies a higher spectral tilt (lower high frequency energy) was perceived as an increase in breathiness [7,25]. These contradictory findings may be the result of two opposing acoustic consequences of breathy voice production. First, spectral tilt will increase, because the less abrupt opening and closing of the glottis during breathy phonation results in the attenuation of high frequency harmonics [24,25]. Second, the increased amount of turbulent airflow during the glottal open phase will result in increased high-frequency noise energy [18,20,24]. Klatt and Klatt showed that simultaneously increasing spectral tilt and adding high frequency noise resulted in increased breathiness [7]. Samlan and Story suggested that the salience of these cues may change depending on the severity of breathiness. When vocal fold separation is small, a listener may first attend primarily to spectral tilt. As vocal fold separation increases, a listener may pay more attention to the amount of high-frequency noise [23].

The spectral shape measures reviewed in this section were affected by the presence of aspiration noise at high frequencies. The next section will focus on acoustic correlates that directly estimate the amount of noise present in breathy signals.

### 4.2. Amount of Noise in Breathy Signals

The presence of high-frequency noise was observed in the earliest investigations of the spectral characteristics of breathy vowels [35,36]. Using stimuli synthesized with a Klatt synthesizer, researchers showed that higher amounts of (high frequency) aspiration noise produced an increase in breathiness [7,21]. Several studies have attempted to measure the amount of noise as an acoustic cue that differentiates breathy from non-breathy vowels. The amount of noise has often been calculated as the ratio of the energy at the harmonic frequencies, relative to the amount of energy at inharmonic frequencies (“harmonics-to-noise ratio,” HNR; “signal-to-noise ratio,” SNR [1,6,15,20,23,34]). Various software programs and algorithms have been used to calculate HNR [8,9,15,20,23]. Given its widespread use, one may expect that HNR would be a reliable and strong acoustic correlate of breathiness. While there is evidence suggesting that HNR is an acoustic correlate of breathiness [14,15,23,38], there are also contrasting findings [8,9,20]. Hillenbrand and colleagues [1,6] argued that measuring HNR is prone to errors when pitch (F0) tracking is not sufficiently accurate. The lack of accuracy of pitch tracking algorithms may have contributed to the inconsistent findings about the role of HNR as an acoustic correlate of breathy voice. The following section will review measures of periodicity, some of which also rely on accurate estimates of the F0 of phonation.

### 4.3. Periodicity of the Speech Signal

Researchers have investigated periodicity measures as acoustic correlates of breathy voice because these measures are sensitive to the amount of noise in speech signals.
Measures that represent periodicity include: cepstral peak prominence (CPP) [1,6,15,20,23,34]; pitch perturbation measures of jitter [9,10,14,15,20]; and glottal-to-noise excitation ratio (GNE) [3,14].

CPP uses “cepstral analysis,” which is a signal processing technique used to estimate the fundamental frequency of an amplitude waveform. Performing a cepstral analysis on a periodic signal yields a cepstrum (in the time domain) that will show a prominent peak at a specific location (“quefrency”) in the cepstrum. This quefrency is used as an estimate of the period of the signal. The “prominence” of the peak can be used to estimate the periodicity of a signal: signals with higher noise energy will have a less prominent peak (lower periodicity). CPP is calculated as the difference between the level of the cepstral peak and the level of the regression line fitted to the cepstrum at the quefrency of the peak [1].

CPP has consistently been found to have strong negative correlations with breathiness (lower periodicity corresponds to higher breathiness). Hillenbrand and Houde [6] found strong correlations between breathiness and CPP for vowels isolated from continuous speech (r = -0.88), and for sustained vowels produced by speakers with voice pathologies (r = -0.89). For control speakers, the correlation was equally strong (r = -0.92). Both Shrivastav and Sapienza [15] and Samlan and Story [23] replicated these results; they reported that CPP was the single best predictor of breathiness, accounting for 75.1% and 86.7% of the variance in breathiness ratings, respectively. Hartl et al. [20] found significant differences in CPP between speakers with and without voice pathology. They did not perform a multiple regression analysis in order to determine the amount of variance in breathiness ratings explained by CPP; instead, Hartl et al. calculated Spearman’s rank order correlations between CPP and breathiness ratings. They reported low correlations between CPP and breathiness ratings for both speaker groups.

In summary, the findings of the majority of studies indicate that CPP accounts for a large portion of the variance in breathiness ratings. Therefore, CPP is likely to be a strong acoustic correlate of breathiness. It is unclear why Hartl et al. [20] found low correlations between CPP and breathiness ratings. However, it is worth noting that they had relatively small sample sizes, and that the group of speakers with voice pathology consisted of only speakers with unilateral vocal fold paralysis.

Jitter is another commonly used measure of periodicity that represents the amount of variability in the duration of successive pitch periods. Although there are different formulas to calculate the amount of jitter in a signal, this measure is usually calculated as the average difference in duration between successive periods of a waveform divided by the average duration of the periods. Jitter is often expressed as a percentage [39]. Shrivastav and Sapienza [15] found that the percentage of jitter correlated strongly with breathiness ratings (r = 0.863). By contrast, Hartl et al. [20] reported low correlations between breathiness ratings and jitter. It is possible that contrasting findings about jitter are a result of the dependence of this measure on accurate pitch tracking, which may be difficult to achieve. For this reason, some researchers have chosen to avoid investigating this measure as an acoustic correlate of breathiness [1,6,18].

An acoustic measure that can be defined as a periodicity measure is the GNE [3] (N.B.: GNE can also be categorized as a noise measure; see for example, [40]). This measure estimates the amount of periodic-to-aperiodic energy in the waveform by first estimating the glottal waveform through inverse filtering techniques. The glottal waveform is band-pass filtered, using narrow-band filters. The temporal envelopes of the filtered waveforms are cross-correlated to obtain the correlation functions between each pair of temporal waveforms. The maximum of each correlation function is found; the GNE is the highest value from this set of maxima. This measure is based on the assumption that glottal closure will excite all frequency channels simultaneously—in which case the temporal envelopes of pairs of band-filtered signals will be similar. By contrast, noise due to turbulent airflow through the glottis will not excite all frequency channels simultaneously—leading to low correlations between pairs of band-filtered waveforms. Fröhlich et al. [3] found that the GNE was significantly different between a control group without voice disorders and a group of people diagnosed with laryngeal cancer (the latter group produced post-operative phonation samples). In this study, several other acoustic measures of the spectral shape of the signal (H1 – H2, B1 – B0 and A1 – A3) were not significantly different between groups. GNE is not used often, because of the previously mentioned difficulty with accurately estimating the glottal waveform from speech signals [37].

In summary, CPP is probably the most often investigated and strongest acoustic correlate of breathiness among measures of periodicity. CPP is a more robust measure than both jitter (which relies on pitch tracking), and GNE (which relies on glottal inverse filtering techniques).

4.4. Amplitude Perturbation Measures

Amplitude perturbation measures represent amplitude changes from one cycle of a periodic signal to the next. These measures are strongly affected by the noise content of the signal in a similar way as periodicity measures. An incomplete glottal closure and a long glottal open phase, both of which are characteristics of breathy voice, cause an increase in turbulent airflow at the glottis. The turbulent airflow associated with a breathy signal may result in
random amplitude changes over successive cycles of the waveform. Amplitude perturbation measures include shimmer [15,20], and measures based on correlations of successive waveform periods (Pearson’s r at the autocorrelation peak, RPK; mean waveform matching coefficient, MWC) [1,6,14].

Shimmer represents the amount of amplitude variability across n successive glottal pulses. It is calculated as the average of n − 1 differences between the amplitudes of successive periods, divided by the average amplitude of n periods. For example, Praat has built in functions that can calculate shimmer for two or more successive periods of a waveform [39]. Shrivastav and Sapienza [15] reported that shimmer significantly contributed to the explained variance in breathiness ratings, although its contribution was smaller than that made by CPP and HNR. Hartl et al. [20] found that their two speaker groups (pathological and controls) were significantly different in terms of shimmer. They found that shimmer and breathiness ratings were strongly correlated (r = 0.71) for the pathological group of speakers, but that no correlation was found for the group of normal speakers.

RPK is a measure of the similarity of the amplitude waveforms of adjacent periods of a vowel [1]. It is calculated as follows. First, the fundamental frequency of a 30 ms portion of the signal is estimated by searching for the peak of the autocorrelation function. A correlation coefficient is then calculated between the 30 ms signal portion and the same portion that has been delayed by the period of the fundamental frequency calculated in the previous step. An increase in noise in the vowel sound will decrease the similarity of adjacent waveforms, and decrease RPK [1]. Hillenbrand and colleagues found moderate to strong correlations between breathiness ratings and the average of RPK measures obtained every 10 ms from unfiltered and band-pass filtered vowel stimuli (r = −0.54 to −0.89), across the range of stimuli and speakers used in their studies [1,6]. Castillo-Guerra and Ruiz [14] measured the mean waveform matching coefficient (MWC), which is similar to RPK. MWC differs from RPK in that the correlation coefficient is calculated for a pair of successive wave periods [41]. Castillo-Guerra and Ruiz found that MWC made no significant contribution in explaining the variance in breathiness ratings. By contrast, in their study shimmer made a small but significant contribution to the amount of variance in breathiness ratings (adjusted R value increased by 4.2%).

Amplitude perturbation measures have not been investigated often in the literature on breathiness, because they usually account for a small proportion of the variance when other acoustic measures (spectral shape, noise and periodicity measures) are included in regression models. Shrivastav [18] excluded shimmer from his set of acoustic correlates for the reason that shimmer’s high correlation with SNR complicates the interpretation of findings on the perception of breathiness.

4.5. Discussion

This review of the acoustic correlates of breathiness shows that several acoustic features correlate with the perception of breathiness and with one another. Fischer-Jørgensen emphasized the multidimensional nature of breathiness by stating that: “We are thus faced with a situation where a large number of unstable acoustic cues correspond to a simple physiological difference and to one functional feature” [35]. Because the large number of acoustic cues that contribute to the perception of breathiness co-vary with physiological processes associated with breathy voice, manipulating any single cue by itself can result in unexpected perceptual consequences. For example, Klatt and Klatt found that increasing the amplitude of H1, led to increased perceptions of nasality as well as breathiness [7]. An increase in perceived nasality was also observed by Arai as a result of the increase in the value of a synthesis parameter that represented the open phase of the glottis during phonation [21].

Given the multidimensional nature of breathiness, one may ask which acoustic measures are the most important acoustic correlates of breathy voice. This question has been studied either by comparing the strength of the correlations between breathiness ratings and a number of acoustic measures, or by using multiple regression models. Researchers that used the strength-of-correlation analyses have reported that CPP and high-to-low frequency energy showed moderate to strong correlations with breathiness ratings [1,6]. Although this type of analysis provides useful information about which acoustic measures strongly correlate with breathiness, it is difficult to draw conclusion based on the relative strengths of the correlations coefficients alone, especially since the various acoustic measures are correlated with one another [1,6,7,23,24]. Multiple regression analyses have the advantage of providing information about the amount of variance that is explained by each variable independent of other variables. Using multiple regression models, researchers found that CPP [15,23], various measures of noise [7,15,21,23–25] and high-to-low frequency energy [4,23] were the most consistently salient cues to breathiness. H1 − H2 (or H1” − H2”) appeared to be an important cue in some studies [1,6,7,14], but contrasting results have also been reported [23]. To conclude, the measures that have been found to be important across studies are CPP, measures of noise, and high-to-low frequency energy. These results should be interpreted with caution because not all studies included all types of acoustic measures (spectral shape, noise, periodicity and amplitude perturbation measures) in
their analyses. It is important to consider that breathiness is a multidimensional quality. The failure to include an acoustic measure (A) in a study may result in another correlated measure (B) to appear as a strong correlate of breathiness, even though B might have explained a much smaller portion of the variance in breathiness ratings had A been included in the study. It is possible that some of the inconsistent findings reported above may be due to the inclusion of different sets of measures in different studies. Another reason for inconsistent findings is that some acoustic correlates may not be monotonically related to breathy voice. For example, Samlan and Story [23] found that $H1^* - H2^*$ and $B0 - B2$ are non-monotonically related to breathiness ratings. Their results suggest that a set of stimuli including a wide range of severity of breathy utterances might result in a poor correlation between these measures and breathiness ratings. By contrast, the inclusion of only mild to moderately breathy stimuli might result in a stronger correlation between $H1^* - H2^*$ and perceptual ratings.

5. SUMMARY

The aim of this review was two-fold. The first goal was to discuss the methodological choices that affect the outcome of investigations on breathiness. Each of these choices has the potential to introduce a source of variability in the data. An attempt to increase agreement in findings would require minimizing or controlling variability. In terms of the stimulus type used, recordings have the potential to introduce acoustic characteristics over which the researcher has little or no control. Using synthesized stimuli means that it is easier to control (mathematical, acoustic or articulatory) model parameters. However, the manipulation of these parameters can at times result in unexpected consequences. For example, the manipulation of $H1$ may cause listeners to perceive this change as an increase in first formant amplitude and/or bandwidth, if the first formant and $H1$ are close in frequency. This review has shown that including external reference stimuli in tasks [11,22], using average ratings across several stimulus repetitions and standardized scores [13] can reduce the effects of unstable internal listener references. The differences in tasks, acoustic measures and listener characteristics across studies have prevented the emergence of a clear consensus regarding the importance of various acoustic correlates of breathy voice (discussed in Sect. 4.5). The second goal was to review the evidence about the acoustic measures that have been associated with breathiness. This review identified the acoustic measures that are likely to be the most important acoustic correlates of breathiness. We hope that this article will prove to be a useful guide for selecting methods and acoustic measures that are most likely to result in valid and reliable findings in future investigations on the acoustic correlates of breathiness.

REFERENCES


