Published in final edited form as:

J Speech Lang Hear Res. 2012 April ; 55(2): 639–647. doi:10.1044/1092-4388(2011/10-0337).

Developing a Single Comparison Stimulus for Matching Breathy Voice Quality

Sona Patel^a, Rahul Shrivastav^b, and David A. Eddins^c

^aUniversity of Florida, Gainesville

^bMalcom Randall VA Medical Center, Gainesville, FL

^cGlobal Center for Hearing and Speech Research, University of South Florida, Tampa, FL

Abstract

Purpose—In this experiment, a single comparison stimulus was developed as a reference in a perceptual matching task for the quantification of breathy voice quality. Perceptual judgments of a set of synthetic voice samples were compared to previous data obtained using multiple comparison stimuli “customized” for different voices (Patel, Shrivastav, & Eddins, 2010).

Method—Five male and 5 female samples of the vowel /a/ were selected from the Kay Elemetrics Disordered Voice Database and resynthesized using a Klatt synthesizer. Eleven samples were created for each base voice by manipulating the aspiration noise level. Five samples from each continuum were evaluated in a perceptual matching task in which a single sawtooth and noise comparison stimulus was used to obtain breathiness judgments. Linear regression was used to compare measurements obtained using the new comparison stimulus against the customized comparison stimuli.

Results—Results indicated that the noncustomized sawtooth comparison provides reliability and perceptual distances between stimuli similar to those obtained using customized comparison stimuli.

Conclusion—A single-variable matching task using a single comparison stimulus can be used to obtain perceptual estimates of breathiness across voices and experiments in a laboratory setting. This technique will help develop models of voice-quality perception.

Keywords

voice quality; matching; breathy; single-variable

The ability to quantify voice quality precisely and quickly is critical for a number of applications, such as for the clinical assessment of voice disorders as well as automatic detection of a speaker’s gender (e.g., Lugger & Yang, 2006), age (e.g., Harnsberger, Shrivastav, Brown, Rothman, & Hollien, 2008), and emotion (e.g., Gobl & Ní Chasaide, 2003; Scherer, 1986). Investigations of vocal quality commonly use rating scales, visual analog scales, or magnitude-estimation tasks. Although often useful, these approaches have several limitations. First, perceptual judgments obtained using traditional rating scales have been shown to be highly variable (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Second, rating-scale judgments are affected by a variety of contextual factors such as the range of the quality under study, the frequency at which stimuli are presented, and other

biases resulting from the assignment of numbers to stimuli (Poulton, 1989). Finally, the rating scale may only provide ordinal level measurement (Shrivastav, Sapienza, & Nandur, 2005). As a result, the perceptual distances between any two successive scores on a rating scale may not be equivalent. Ordinal measures may be useful for rank ordering voices by quality; however, these cannot be used to quantify the magnitude of the quality under measurement. For this purpose, interval or ratio-level measurement is needed and can be achieved using other tasks such as direct magnitude estimation (e.g., Toner & Emanuel, 1989). In a *direct magnitude-estimation task*, a given stimulus can be assigned any number greater than zero, while maintaining a ratio-level relationship between judgments. However, this task also has several limitations. For example, the values selected by participants are arbitrary and are not absolute measures of quality. Additionally, like the rating scale estimates, direct magnitude estimates are also biased by context (Poulton, 1989). For these reasons, neither data obtained from rating scales nor data obtained from direct magnitude-estimation tasks are suitable for generating computational models of voice quality. Instead, successful models of voice quality can only be generated if we obtain perceptual data that are relatively unbiased by context and have known measurement properties (i.e., ordinal, interval, or ratio level of measurement).

Rather than ask participants to assign numbers to quantify the magnitude of voice quality, an alternate approach is to use a *matching task*. This task provides ratio-level perceptual distances between voices and avoids many of the biases mentioned above. In the classical psychophysical matching task, two stimuli are presented during each trial: the “standard” and the “comparison.” The physical attributes of the standard stimulus (i.e., the test stimulus) are fixed within a block of trials, whereas the comparison stimulus is varied along one or more physical dimensions (characterized by one or more acoustic parameters) from trial to trial. The comparison stimulus then serves as a reference for measurement of the perceptual quality under study. Importantly, the variable dimension(s) of the comparison then provides a quantitative reference for the perceived quality of the standard. The matching task simply requires participants to judge whether or not the perceptual quality of the comparison matches that of the standard voice. This “classical” methodology (S. S. Stevens, 1934; S. S. Stevens & Volkman, 1940) is commonly used in all sensory systems and has served as the basis for the establishment and validation of quality scales and corresponding units (e.g., loudness), some of which have been adopted by national and international standards organizations.

An important consideration when choosing the comparison stimulus in a matching task is the nature and degree of physical similarity between the comparison and the standard stimuli. Similarity is not required, as is routinely demonstrated in cross-modality matching tasks, which produce reliable and valid matching data (e.g., matching the loudness of a tone to the brightness of a light; J. C. Stevens & Hall, 1966; J. C. Stevens & Marks, 1965). Although the stimuli in a cross-modality matching task may be considered to be at one similarity extreme, the stimuli used in a within-modality matching task also may differ substantially in their physical similarity. Humans are quite good at formulating and using such analogous perceptual concepts. Nevertheless, using a comparison stimulus that is physically similar to the standard has a secondary advantage in providing insight into the perception of the standard stimulus because the properties of the comparison stimulus are known. In fact, the physical units associated with the comparison can be used to quantify the relationship among different standard stimuli of interest.

By providing listeners with a comparison stimulus, the matching task has also been shown to reduce variability in judgments of voice quality (Kreiman & Gerratt, 2005; Kreiman, Gerratt, & Ito, 2007; Patel, Shrivastav, & Eddins, 2010). In the matching task used by Kreiman and Gerratt (2005), participants were asked to manipulate a number of synthesizer

voice parameters in a quasi-interleaved manner so that the comparison stimulus matched the standard in quality. The user-adjusted parameter settings served as a measure of voice quality. It was assumed that because vocal quality is multidimensional, listeners must be allowed to manipulate all of the relevant acoustic parameters for describing the quality in order to make a perceptual evaluation of the voice quality and obtain high agreement between listeners. Although this technique may be useful for obtaining high agreement across participants, the manipulation of multiple synthesizer parameters complicates the modeling of perceptual distances between voices because multiple combinations of synthesizer settings may result in the perception of equivalent breathiness. For example, two stimuli with different levels of intensity, fundamental frequency, and aspiration noise may be perceived as having comparable magnitudes of breathiness. To test whether accurate perceptual distances between stimuli have been obtained, subsequent perceptual tests to judge the similarity of these stimuli must be performed. Alternatively, the relative contributions of each of the synthesizer parameters to the perception of a certain vocal quality must be known. In addition, modeling of voice-quality perception would require accurate estimation of perceptual distances measured in a multidimensional space. These factors make it difficult to use a multiparameter matching technique to obtain perceptual data of voice quality that are suitable for generating predictive models of voice-quality perception. One approach to reducing this complexity is to obtain perceptual estimates of breathiness using a matching task that uses a comparison stimulus that can vary in only one dimension at a time.

It is important to note that the use of a one-dimensional metric does not presume that breathiness, or any other percept, is one-dimensional. Instead, this simply implies that changes in breathiness resulting from any and all acoustic cues can be measured against changes in the perception of the one-dimensional comparison stimulus. One such example of using a one-dimensional comparison stimulus is that for loudness. In many experiments, estimates of loudness are obtained by comparing it against the loudness of a 1-kHz sinusoid whose sound-pressure level (SPL) can be varied by the listener. This is a one-dimensional comparison stimulus because only one attribute of the stimulus (the SPL) can be manipulated by the listener. Although a number of acoustic characteristics of the test stimuli—for example, their duration, spectral shape, pitch, and so forth—affect their loudness, all of these changes can be evaluated by comparing them against changes in the loudness of the one-dimensional comparison stimulus (J. C. Stevens & Hall, 1966). Such mapping of multiple acoustic cues to a single comparison stimulus is further illustrated in Figure 1.

As shown in Figure 1, a given physical stimulus may give rise to a variety of perceived qualities, some of which are easily identified and quantified, whereas others are more subtle and less defined. Each perceptual quality arises from one or more underlying physical attributes. Because most sensory systems are highly nonlinear, the relationship between the perceived quality of the physical stimulus and its underlying physical attributes is likely to be nonlinear as well. One way to deal with such nonlinearities in sound-quality measurement is to use a matching task in which the quality under study, associated with a particular stimulus or stimulus class, can be mapped directly to the physical units of the comparison stimulus, thereby determining a quantitative description of the perceptual magnitude of that quality. The development of such a matching task for voice quality was undertaken by Patel et al. (2010) for the evaluation of breathiness in a set of synthetic voices. The standard stimuli included a set of /a/ vowels that were synthetically generated to resemble 10 voices from the Kay Elemetrics Disordered Voice Database (KEDVD; Kay PENTAX, Lincoln Park, NJ). Comparison stimuli consisted of a low-pass filtered sawtooth wave mixed with low-pass filtered broadband noise. Both the sawtooth wave and noise were customized to match the fundamental frequency (F0) and spectral slope of each base voice standard. Thus, a slightly different comparison stimulus was used for each of the 10 base

voice standards to facilitate matching breathiness across multiple stimuli. Participants adjusted the signal-to-noise ratio (SNR) of the reference signal (i.e., the comparison stimulus) until the breathiness of the comparison matched the breathiness of the standard. This was repeated for each of the standard /a/ stimuli. The SNR of the comparison stimulus at which listeners judged the comparison and the standard to be equally breathy was used as an index of breathiness, allowing breathiness to be measured in units of dB SNR. The matching thresholds compared favorably with magnitude estimation and rating-scale judgments with the same standard stimuli, indicating that the matching task can provide highly reliable estimates of voice quality at the ratio level. The matching task provided a measure of breathiness that was minimally biased by contextual factors when compared to rating-scale and magnitude-estimation data. In addition, the matching thresholds associated with each standard stimulus provided physical units (dB SNR) associated with the breathy quality of each standard.

Although Patel et al. (2010) demonstrated that a one-dimensional comparison stimulus was adequate to quantify changes in breathy voice quality, the customization of the comparison stimulus for each talker resulted in some difficulty when comparing data across multiple standards (or talkers). In other words, a matching threshold obtained for one voice could not be compared directly to that of another because each threshold was obtained using a different comparison stimulus. This limitation prevents more widespread application of the one-dimensional matching task to evaluate changes in breathiness.

There are two ways to address this limitation. One approach is to use multiple standard stimuli and subsequently compute a correction factor for changes in the fundamental frequency and spectral slope of the signal based on some calibration experiment. If such a correction factor were available, the SNR at which the breathiness of the standard and signals match could be transformed into standard units. This approach can help maintain comparison stimuli that are highly similar to the standard, while still allowing investigators to compare the perception of quality across multiple test items. For example, one may estimate the loudness of an alarm with a tonal comparison and the loudness of a motor with a noise-based comparison and then apply a correction factor that relates the loudness of tonal and noise stimuli.

A second and simpler approach is to make all measurements using a single comparison (e.g., 1000-Hz tone) and avoid the use of a correction factor altogether. In the context of measuring breathy voice quality, this would require using the same comparison stimulus (e.g., a sawtooth-plus-noise complex) to measure breathiness for all voice stimuli. It is possible, however, that the use of a common comparison stimulus may have the negative effect of making the matching task more difficult for the participants if the standard stimuli (vowel samples) differ along several dimensions such as pitch, loudness, or other qualitative dimensions.

Hence, the purpose of this experiment was to evaluate whether a single, noncustomized comparison stimulus could be used for matching breathiness across multiple voices. The comparison stimulus used here was a sawtooth wave-plus-noise complex. This stimulus was selected for two reasons. First, the sawtooth wave was chosen because of the similarity between its spectrum and the harmonics associated with the glottal source of natural speech. Second, it was possible to adjust the SNR of this complex by simply varying the intensity of the noise level prior to mixing the signals. Pilot work indicated that varying the SNR of a simple tone in noise also served as a viable comparison; however, participants found this task to be somewhat difficult and unnatural. At the other extreme, the use of a natural voice as a comparison stimulus would raise the issue of how similar the standard and comparison stimuli should be and be further complicated by inaccuracies in separating the aspiration

noise from the harmonic energy of the speech signal. In the present study, the perceptual judgments of breathiness obtained using a single, noncustomized comparison were examined in relation to judgments obtained using the customized sawtooth wave in noise reported by Patel et al. (2010). A single comparison stimulus across all test stimuli was used to allow direct comparisons of perceptual estimates across multiple voices and experiments. Such comparisons are necessary for facilitating related work to develop a computational model of voice-quality perception.

Method

Participants

Ten individuals (seven female, three male; mean age = 22 years) were recruited to participate in the matching experiment. All participants were native speakers of American English, had normal hearing bilaterally (air-conduction pure-tone threshold below 20 dB HL at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz; ANSI S3.21-2004), were students at the University of Florida, and were enrolled in either the Communication Sciences and Disorders or the Linguistics program. These students had a comparable amount of experience in listening to disordered voices and were paid for their participation in these experiments in accordance with the local institutional review board–approved informed consent.

Stimuli

The stimuli included a set of test stimuli (the standards) and a reference signal (the comparison). The *standards* refer to the voices to be evaluated in this experiment and consisted of fifty 500-ms samples of the vowel /a/. These included 10 base samples of the vowel /a/ that were synthetically generated to resemble 10 natural voices (five female and five male) from the KEDVD. These were generated using a Klatt synthesizer (HLSyn; Sensimetrics Corporation) with the LF model (Fant, Liljencrants, & Lin, 1985) as the voicing source. These 10 voices were selected based on a pilot experiment to determine a set of samples that includes a wide range of breathiness. A continuum of 11 tokens was created for each of the 10 base voice standards by systematically varying the level of aspiration noise (AH). The AH parameter was manipulated to create a range of breathiness for each base voice standard because it has been reported to be the most important cue in the perception of breathy vocal quality (e.g., Klatt & Klatt, 1990). The exact synthesizer parameter settings, including the range of AH values used (limited to include only those stimuli that were judged as natural sounding in a pilot experiment), are provided in Table 1.

Although 11 tokens were created from each base voice standard, only five tokens from each continuum were selected for this experiment. This decision was made for several reasons. First, Patel et al. (2010) reported that the matching task required greater test time than a rating-scale or magnitude-estimation task, suggesting that the test time for evaluating all 11 samples may be impractical. Second, Patel et al. (2010) showed that the first three samples in most voice continua demonstrated very little differences in perceived breathiness. This likely resulted from the large difference limens for breathiness seen for voices with high SNRs (e.g., Kreiman & Gerratt, 2000; Shrivastav & Sapienza, 2006). Thus, two of the three initial samples could be dropped without reducing the range of breathiness represented by the voice-standard continua. In order to reduce the total test time further, but maintain the change in AH level between samples, samples 3, 5, 7, 9, and 11 from each base voice standard continua were selected for the present experiment, resulting in 50 standards (10 base voice standards $\times$ 5 stimuli/standard). The use of these samples is consistent with that in Patel et al. (2010) to enable comparisons between studies. The stimuli were generated at a sampling rate of 11025 Hz and were then up-sampled to 12207 Hz to match the clock speed

of the hardware used for the listening test. The stimuli were ramped by a 10-ms cosine function to avoid artifacts during digital-to-analog (DA) conversion and were scaled to have equal RMS energy.

The *comparison* refers to the reference used for the evaluation of all standards. This comparison stimulus consisted of a sawtooth wave ($F_0 = 151$ Hz) mixed with speech-shaped noise and was generated at a sampling rate of 12207 Hz and with 16-bit quantization. Both the sawtooth and noise were low-pass filtered to have a slope of -7.262 dB/octave above 151 Hz. The characteristics of the comparison (F_0 and slope) were determined by averaging the F_0 and spectral slope of roughly 600 disordered voices from the KEDVD.

Procedure

Prior to testing, participants were required to undergo a short training task. The purpose of this exercise was to familiarize participants with the matching task and the speech samples and to reiterate that they should be judging only breathy vocal quality. Two stimuli were presented in each trial, the standard followed by the comparison stimulus. Participants were instructed to match the “breathiness” of the comparison stimulus to that of the standard by selecting a button labeled “Increase breathiness” or “Decrease breathiness.” These responses corresponded to a manipulation of the SNR of the comparison stimulus (e.g., to decrease breathiness, the SNR was increased). The SNR was always varied in 1-dB steps. When both voices were perceived as being equivalent in terms of breathiness, participants selected the “Equal breathiness” option. The breathiness of each standard was measured as the SNR at which the standard was perceived to be equal in breathiness to the comparison. Because this task resembled a “method of adjustment” design, this point of subjective equality was referred to as the *matching threshold*, or SNR_{TH} . Participants matched the comparison to six synthetic standard stimuli exclusively used for training. They were encouraged to increase and decrease the breathiness of the comparison stimulus to become accustomed to variations in the breathiness of the comparison. No feedback was provided during training. Once participants had matched the six standards of the training task, they began the experimental task.

In the main experiment, participants adjusted the breathiness of the comparison stimulus to match each of the 50 standards. Each standard was evaluated six times, for a total of six “runs.” Because the initial value of the comparison stimulus systematically influences the perceived threshold (for example, Driggers, 2003, pp. 2183–2184), it is common practice in psychophysics (e.g., the “method of limits” and the “method of adjustment”) to compute the thresholds as the mean of an equal number of “ascending” and “descending” runs to counter these effects (e.g., Appelle, 1971; Penner, 1984; Robin, Tranel, & Damasio, 1990). In three ascending runs, the initial SNR of the comparison stimulus was less than the predicted threshold. In the remaining three descending runs, the initial SNR of the comparison stimulus was greater than the predicted threshold. The SNR_{TH} for each standard was computed as the average threshold of three ascending and three descending runs. The standards were presented in random order across participants, but the three ascending and three descending runs for a standard were always presented successively before testing the next standard. A short (3–5 min) break was given after participants evaluated three sets of standards to maintain an optimum level of attention and to minimize fatigue. Stimulus generation, presentation, and response collection were under control of the SykofizX software application (Tucker-Davis Technologies, Inc., Alachua, FL), using an RP2 processor, PA5 attenuator, and HB7 headphone buffer (Tucker-Davis Technologies, Inc., Alachua, FL). Stimuli were delivered to the right ear of the participant via ER-2 ear insert (Etymotic, Inc., Elk Grove Village, IL) at 75 dB SPL. Although not critical to the present experiment, monaural presentation of stimuli was preferred to avoid potential binaural interaction effects that are not accounted for in loudness models such as those used to

predict breathiness in vowels (Shrivastav & Sapienza, 2003). Participants were tested in a single-walled, sound-treated booth for three to four sessions of maximally 1.5 hr each.

Results

The perceptual judgments of the 50 standards obtained in the present study were compared to the previous judgments of these standards obtained in Patel et al. (2010), in which customized comparisons were used. First, the reliability of judgments obtained using the comparison stimulus from the present study was computed and compared to the reliability of the customized comparison stimulus. Then, linear regressions were used to examine differences in the SNR_{TH} obtained from both noncustomized and customized comparisons.

Reliability

Within-participant (“intrajudge”) and between-participant (“interjudge”) reliability scores were computed using Pearson’s correlation coefficient. *Intrajudge reliability* was computed as the correlation between mean ascending and descending thresholds within a participant. *Interjudge reliability* was calculated as the mean correlation between each participant’s thresholds and the remaining participants’ thresholds. Results showed that judgments obtained using both noncustomized and customized comparison stimuli resulted in high inter- and intrajudge reliability (see Table 2). The mean interjudge score for the customized comparison stimuli (0.97, $SD = 0.01$) was greater than the score for the present comparison stimulus (0.90, $SD = 0.01$). In addition, participants were slightly more consistent within themselves when given customized comparisons (0.96, $SD = 0.03$) in contrast to the noncustomized comparison (0.91, $SD = 0.03$). These results were not surprising because the characteristics of the customized comparison stimuli were tailored to each base voice standard, thereby providing a perceptually similar comparison, in terms of F0 and spectral slope, for each standard.

Relating Thresholds From the Noncustomized and Customized Comparisons

To determine the suitability of a single comparison stimulus for obtaining perceptual judgments of breathiness, the relation between the SNR_{TH} obtained using the noncustomized comparison and the customized comparison stimuli was examined. The SNR_{TH} of the five stimuli (differing in AH levels) tested within each of the 10 base voice standard continua are shown in Figure 2. In this figure, high SNR_{TH} values indicate standards that were perceived as having lower magnitudes of breathiness. Linear regressions were performed to determine whether the perceptual distances obtained using the noncustomized and customized comparison stimuli were similar across base voice standards. A regression slope of 1 would indicate that the perceptual distances between standards are equivalent for both experiments. Results revealed high R^2 values for all linear regression functions (range = 0.95–1.00; see Table 3), suggesting a strong linear relation between the thresholds. The observed slopes of the regression functions were greater than 1 ($M = 1.47$, range = 1.07–2.10), indicating an expansion of the perceptual distances represented by the y -axis parameter. In this case the linear regressions were computed with the noncustomized comparison stimulus thresholds on the ordinate, which suggests that the noncustomized comparison stimulus resulted in a wider range of matching thresholds than customized comparisons for the same set of test stimuli.

To confirm these results, the standard deviation and range of the SNR_{TH} obtained using both noncustomized and customized comparisons were computed. The range was defined as the difference in SNR_{TH} between the most and least breathy stimulus in each base voice standard continuum. On an average across base voice standards, a larger range was observed when using the noncustomized comparison stimulus (13.8 dB vs. 9.7 dB). A similar pattern

was observed for the standard deviations (5.8 dB vs. 4.1 dB). The mean difference in the range of SNR_{TH} was 4.1 dB (range = 1.5–7.4 dB) across the 10 base voice standards, further supporting the observation that judgments made with the noncustomized comparison provide a wider range of matching thresholds than the customized comparisons.

Further statistical analysis was performed to investigate whether variations in the F0 across base voice standards resulted in systematic differences in the SNR_{TH} obtained using the noncustomized comparison. First, the correlation between the mean SNR_{TH} for each base voice standard and the F0 for that base voice standard was computed. The thresholds were poorly correlated with F0 ($r = .323$), indicating minimal systematic effect of F0 on thresholds. Second, a correlation between the slope of the regression function for each base voice standard and the F0 for that base voice standard was computed to determine whether F0 influenced the relationship between the SNR_{TH} obtained from both experiments. This correlation also was very low ($r = .218$), again indicating little systematic effect of F0 on the differences between SNR_{TH} . However, both of these results are limited by the small number of stimuli tested and the bimodal distribution of F0 values for male and female base voice standards. Hence, these observations should be considered as only preliminary in nature.

Discussion

Studies have shown that people are more consistent in their perceptual evaluation of voice quality when given a comparison stimulus or an “anchor” (e.g., Kreiman et al., 1993; Yiu, Chan, & Mok, 2007). The use of a synthetic comparison stimulus for the evaluation of voice quality has been successfully used previously by, for example, Yiu et al. (2007) in a paired comparison task and by Gerratt, Kreiman, Antonanzas-Barroso, and Berke (1993) in a 5-point rating-scale task. However, the matching approach has the additional advantage of minimizing errors arising from context such as the number and type of stimuli used in a particular experiment. Patel et al. (2010) reported that a matching task could be used to measure the magnitude of breathiness in vowels with better accuracy and reliability than rating-scale and magnitude-estimation tasks. Because the comparison was a novel stimulus, a sawtooth wave in noise, it was custom-designed to match the F0 and spectral slope of each of the 10 base voice standards. This procedure limits direct comparisons across base voice standards, because the differences resulting from the comparison stimuli were not accounted for when averaging judgments across voice standards. The present experiment addressed this limitation by testing whether a single comparison stimulus could be used to obtain reliable estimates of breathiness across multiple base voice standards. The experimental procedures and stimuli were identical to those of Patel et al. (2010), except for the comparison stimulus. Participants judged the breathiness of 50 vowel samples from 10 base voice standards relative to a 151-Hz sawtooth wave mixed with low-pass filtered noise. The use of a synthetic sawtooth-and-noise complex was intentionally selected to imitate naturally occurring vowels. This comparison stimulus also allowed easy isolation of the noise and harmonic components, which was necessary to modify the SNR of the comparison stimulus. Use of a natural stimulus would have introduced some error, because a noise-extraction step would have to be performed to separate the noise from the harmonic energy.

A comparison of the SNR_{TH} derived using the present comparison stimulus and customized comparison stimuli revealed a linear relationship with slopes between 1.07 and 2.10, indicating a slight expansion of the perceptual distances of the noncustomized comparison stimulus. Although both experimental procedures provided similar perceptual distances between standards, the expanded range of thresholds obtained using the noncustomized comparison is an advantage, because greater distances between samples indicate that they are more differentiable from one another. Therefore, a single sawtooth wave-in-noise comparison stimulus as described in the present article can be used as a reference for

estimating the magnitude of breathiness. Unlike arbitrary scales (e.g., n -point rating scales, visual analog scales, or magnitude estimation) a matching task using a single comparison stimulus provides a fairly unbiased and highly reliable method to quantify breathiness and can allow direct comparison of data across experiments, laboratories, and groups of participants.

Further experiments are ongoing to determine whether such high reliability is maintained for natural dysphonic standards and for other vocal qualities such as roughness and strain. In these experiments, the properties of the reference signal are slightly modified to address the percept being measured. For instance, experiments to quantify roughness use the same sawtooth and noise carrier as the comparison signal. However, instead of varying the SNR, listeners manipulate certain parameters of its amplitude modulation function, as these changes are closely related to the perception of roughness (Cummings et al., 2010; Zwicker & Fastl, 1999). The eventual goal is to refine and validate these matching techniques so as to allow quantification of voice-quality perception along its various subdimensions and for a wide range of stimuli. Obtaining bias-free estimates of perception is essential for developing models for voice-quality perception.

This approach for quantifying perceptual judgments of voice quality is recommended when an experimenter seeks to gather relatively unbiased perceptual distances across stimuli. The choice of the variable acoustic parameter is critical in a matching task. It is important that the variable parameter allow mapping of a multidimensional percept to a one-dimensional comparison stimulus. Although breathiness may be cued by multiple acoustic changes, the manipulation of a single parameter in the comparison stimulus simplifies the estimation of perceptual distances between stimuli. In contrast, matching tasks that allow manipulation of multiple acoustic parameters may result in several combinations of acoustic parameters to indicate the same final percept. This prevents accurate measurement of perceptual distances between stimuli without additional corrections. A potential downside to this approach is the extensive test time required for conducting matching experiments. The duration needed to complete a matching task is about three times greater than for a controlled rating-scale task (using procedures described in Shrivastav et al., 2005) with the same number of stimuli. Furthermore, this approach needed some automation of the test procedure, which was implemented using specialized hardware and software. However, the basic principles are very simple and can easily be replicated in most commercially available hardware or software. Nevertheless, because of these reasons, this approach is not suitable for everyday clinical practice. Rather, this approach has been developed with the intention of generating laboratory data that can be replicated across listeners or locations. Although further investigation is needed to apply the current technique to words and sentences, such data will help develop clinically useful tools for the measurement of voice quality. The data obtained here provide a basis for further expansion of this technique for estimating perceptual distances in voice quality.

Conclusion

Previous research has shown that a single-variable matching task could be used to obtain perceptual judgments of breathy voice quality and provide unbiased ratio-level measurements of breathy voice-quality perception (Patel et al., 2010). However, the use of a customized comparison stimulus for each base voice standard complicated the analysis of perceptual judgments across base voice standards. To avoid computing a conversion factor to compare judgments across standards in future experiments, the present study investigated whether a single comparison stimulus could be used for all base voice standards. Results showed that the reliability and perceptual distances between standards were similar for breathiness judgments obtained using the noncustomized and customized comparison

stimuli (mean SNR thresholds varied from 3.3 dB to 28.2 dB). Therefore, future research can use a matching task with a low-pass filtered 151-Hz sawtooth wave in noise as the comparison stimulus to obtain perceptual judgments of breathiness. The ability to obtain perceptual data that can be compared across voices and participants of different experiments will enable the development of acoustic models of voice quality.

Acknowledgments

This research was supported by National Institute on Deafness and Other Communication Disorders Grant R01 DC009029.

References

- American National Standards Institute. Methods for manual pure-tone threshold audiometry (ANSI S3.21-2004). New York, NY: Author; 2004.
- Appelle S. Visual and haptic angle perception in the matching task. *The American Journal of Psychology*. 1971; 84(4):487–499. [PubMed: 5142061]
- Cummings, S.; Shrivastav, R.; Eddins, DA.; Patel, S. Comparison of psychophysical methods for evaluation of rough voice quality; Presentation at the 39th annual symposium of the Voice Foundation; Philadelphia, PA. 2010 Jun.
- Driggers, RG. Encyclopedia of optical engineering. New York, NY: Marcel Dekker; 2003.
- Fant G, Liljencrants J, Lin Q. A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*. 1985; 4:1–13.
- Gerratt B, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. *The Journal of the Acoustical Society of America*. 1993; 36:14–20.
- Gobl C, Ní Chasaide A. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*. 2003; 40:189–212.
- Harnsberger JD, Shrivastav R, Brown WS Jr, Rothman R, Hollien H. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*. 2008; 22(1):58–69. [PubMed: 16968663]
- Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*. 1990; 87:820–857. [PubMed: 2137837]
- Kreiman, J.; Gerratt, BR. Measuring voice quality. In: Kent, RD.; Bell, MJ., editors. *Voice quality measurement*. San Diego, CA: Singular; 2000. p. 73–101.
- Kreiman J, Gerratt BR. Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*. 2005; 117(4):2201–2211. [PubMed: 15898661]
- Kreiman J, Gerratt B, Ito M. When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*. 2007; 122(4):2354–2364. [PubMed: 17902870]
- Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*. 1993; 36:21–40. [PubMed: 8450660]
- Lugger, M.; Yang, B. Proceedings of ITG-Sprach-Kommunikation-2006. 2006. Classification of different speaking groups by means of voice quality parameters.
- Patel S, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: A comparison of psychophysical methods. *Journal of Voice*. 2010; 24(2):168–177. [PubMed: 19185451]
- Penner MJ. Equal-loudness contours using subjective tinnitus as the standard. *Journal of Speech and Hearing Research*. 1984; 27:274–279. [PubMed: 6738040]
- Poulton, EC. Bias in quantifying judgments. Hove, England: Erlbaum; 1989.
- Robin DA, Tranel D, Damasio H. Auditory perception of temporal and spectral events in patients with focal left and right cerebral lesions. *Brain and Language*. 1990; 39:539–555. [PubMed: 2076495]
- Scherer KR. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*. 1986; 99:143–165. [PubMed: 3515381]

- Shrivastav R, Sapienza C. Objective measures of breathy voice quality obtained using an auditory model. *The Journal of the Acoustical Society of America*. 2003; 114(4):2217–2224. [PubMed: 14587619]
- Shrivastav R, Sapienza CM. Some difference limens for the perception of breathiness. *The Journal of the Acoustical Society of America*. 2006; 120(1):416–423. [PubMed: 16875237]
- Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech and Hearing Research*. 2005; 48:323–335.
- Stevens JC, Hall JW. Brightness and loudness as functions of stimulus duration. *Perception & Psychophysics*. 1966; 1:319–327.
- Stevens JC, Marks LE. Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Science*. 1965; 54:407–411.
- Stevens SS. The volume and intensity of tones. *The American Journal of Psychology*. 1934; 46(3): 397–408.
- Stevens SS, Volkman J. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*. 1940; 53(3):329–353.
- Toner MA, Emanuel FW. Direct magnitude estimation and equal appearing interval scaling of vowel roughness. *Journal of Speech and Hearing Research*. 1989; 32:78–82. [PubMed: 2704204]
- Yiu EM-L, Chan KM-K, Mok RS-M. Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clinical Linguistics & Phonetics*. 2007; 21(2): 129–145. [PubMed: 17364621]
- Zwicker, E.; Fastl, H. *Psychoacoustics: Facts and models*. New York, NY: Springer; 1999.

Figure 1.

A physiologically and behaviorally relevant model to predict the mapping of physical to perceptual attributes.

Figure 2.

The matching signal-to-noise ratio thresholds (SNR_{TH}) obtained using the noncustomized comparison as a function of the SNR_{TH} obtained using the customized comparisons for the 10 talkers (left column: male talkers; right column: female talkers). Error bars represent the standard error of the mean.

Table 1

Klatt-synthesizer parameters used to synthesize the 10 talkers.

Parameter	Male talkers					Female talkers				
	1	2	3	4	5	1	2	3	4	5
F0	133.1	113.7	115.5	117	134.4	220.4	209	209.1	195.5	200.7
AV	60	60	60	60	60	60	60	60	60	60
OQ	40	55	65	75	85	40	55	65	75	85
SQ	200	200	200	200	200	200	150	350	200	200
TL	0	10	20	30	40	0	10	20	30	40
FL	10	10	10	10	10	10	10	10	10	10
AH range	0-75	0-80	0-75	0-80	55-80	0-80	0-80	0-75	0-80	55-80
AH step size (dB)	7.5	8.0	7.5	8.0	2.5	8.0	8.0	7.5	8.0	2.5
FNP	180	180	180	180	180	180	180	180	280	180
BNP	1000	1000	1000	1000	1000	1000	1000	40	90	30
F1	661	559	814	586	814	891	759	1050	977	957
B1	200	400	600	800	1000	200	400	600	800	1000
F2	1122	1214	1473	1187	1473	1587	1333	1410	1356	1619
B2	200	200	200	200	200	200	200	200	150	200
F3	2281	2340	2250	2463	2250	3083	2930	3000	2905	2877
B3	300	300	300	200	250	300	300	300	200	250
F4	4198	3383	3701	3405	3701	3870	4232	4000	4651	4274
B4	400	400	400	250	300	400	400	400	250	300
F5	4415	4396	4990	4194	4990	4761	4736	4990	4990	4883
B5	500	500	500	300	350	500	500	500	300	350

Note. This table is reproduced from Patel et al. (2010). The level of the aspiration noise was varied for each talker. The ranges and step sizes (shown in bold) were determined through a pilot study. F0 = fundamental frequency; AV = amplitude of voicing; OQ = open quotient; SQ = speed quotient; TL = tilt of the voicing source spectrum; FL = flutter; AH = amplitude of aspiration noise; FNP = frequency of nasal pole; BNP = bandwidth of nasal pole; F1-F5 = formant frequencies; B1-B5 = bandwidths.

Table 2

Inter- and intrajudge reliability described by Pearson's correlation coefficient using the single and customized signals. The *M*, *SD*, and range of correlations are shown.

	Intrajudge		Interjudge	
	Single signal	Customized signals	Single signal	Customized signals
<i>M</i>	0.91	0.96	0.90	0.97
<i>SD</i>	0.03	0.03	0.01	0.01
Range	0.87–0.95	0.92–0.99	0.88–0.92	0.96–0.98

Table 3

Slope, intercept, and goodness of fit (R^2) of the linear regressions between the SNR_{TH} for each talker determined using the single and customized signals.

Talker	R^2	Slope	Intercept
Female 1	1.00	1.50	-16.99
Female 2	0.99	2.10	-25.80
Female 3	0.95	1.07	-5.25
Female 4	0.99	1.73	-4.36
Female 5	0.96	1.32	4.35
Male 1	0.99	1.15	-2.02
Male 2	1.00	1.80	-0.64
Male 3	0.98	1.26	5.42
Male 4	0.97	1.14	6.94
Male 5	0.98	1.60	2.05
<i>M</i>	0.98	1.47	-3.63