NIH Public Access
Author Manuscript
J Speech Lang Hear Res. Author manuscript; available in PMC 2012 June 1.

Published in final edited form as:
J Speech Lang Hear Res. 2011 June ; 54(3): 803–812. doi:10.1044/1092-4388(2010/10-0083).

Comparing Two Methods for Reducing Variability in Voice Quality Measurements

Jody Kreiman and **Bruce R. Gerratt**
University of California, Los Angeles

Abstract

Purpose—Interrater disagreements in ratings of quality plague the study of voice. This study compared two methods for handling this variability.

Method—Listeners provided multiple breathiness ratings for two set of pathological voices, one including 20 male and 20 female voices unselected for quality and one including 20 breathy female voices. Ratings for each listener were averaged together, mean ratings were z-transformed, and the likelihood that two listeners would agree exactly in their ratings was calculated as a function of averaging and standardizing condition. Data were also multidimensionally scaled to examine similarities among listeners in perceptual strategy. Results were compared to parallel analyses of existing breathiness ratings of the same voices gathered using a method-of-adjustment task.

Results—Three-way interactions between the mean rating for a voice, standardization condition, and the number of voices averaged together were observed, but no main effect of averaging condition emerged. Multidimensional scaling revealed significant residual differences in perceptual strategy across listeners after averaging and standardizing. Ratings from the method-of-adjustment task showed both high agreement levels and consistent perceptual strategies across listeners, as theoretically predicted.

Conclusion—Averaging multiple ratings and standardizing the mean are inadequate in addressing variations in voice quality perception.

Keywords

Voice quality; breathiness; scaling

Interrater disagreements in ratings of voice quality have plagued the scientific study of voice perception virtually since its inception. Given the magnitude and persistence of this problem and its implications regarding both rating validity and clinical practice, it is not surprising that authors have proposed a number of methods for handling it. One common approach assumes that significant variability in ratings derives from uncontrollable random error, which can be eliminated by averaging repeated ratings together, either across multiple raters (e.g., Eadie & Doyle, 2005; cf. Emanuel, Lively, & McCoy, 1973, who use the median of many listeners' ratings) and/or across multiple ratings from a single rater (e.g., Shrivastav & Sapienza, 2003). In the traditional version of these techniques, authors typically argue that averaging multiple ratings together allows the investigator to derive the "true" rating for a voice.

Correspondence concerning this article should be addressed to Jody Kreiman, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, CA 90095-1794; jkreiman@ucla.edu.
Jody Kreiman and Bruce R. Gerratt, Division of Head and Neck Surgery, UCLA School of Medicine.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

More recently, Shrivastav, Sapienza, and Nandur (2005) have applied similar logic to model ratings from a single listener. They attribute variability in voice quality ratings to two sources: random errors (caused by lapses in attention, fatigue, and so on) and "criterion" errors corresponding to systematic response biases that affect all stimuli equally (for example, the tendency of one rater to rate all voices towards the higher end of the scale relative to another rater). They then argue that random errors can be corrected by averaging together a number of ratings from a single rater, and that criterion errors can be corrected by standardizing scores from each rater, leading to better estimates of the true rating for each voice and hence better listener agreement as the estimate of the "true" rating improves. In a test of this hypothesis, the probability of exact inter-listener agreement in ratings of breathiness increased significantly (by about 0.47, to 0.90) once such vagaries were corrected. Shrivastav and colleagues interpreted these results as indicating that listeners agree better in these tasks than others have argued. Because ratings of specific voice qualities like breathiness cannot be valid in the face of very low interrater agreement (Kreiman & Gerratt, 1998), these results in turn suggest that such ratings may be valid after all, and that "the idea of a common perceptual space for voice quality across listeners should not be discarded" (Shrivastav et al., 2005, p. 334).

An alternative approach to the problem of interrater unreliability argues that most variability in listener ratings derives from the difficulty listeners experience in performing rating tasks, and not from random factors or inherent differences among listeners (Kreiman, Gerratt, & Ito, 2007). In this view, four factors contribute to variability in listener ratings: instability of listeners' internal standards for different qualities (corresponding experimentally to the presence or absence of "anchor" stimuli; e.g., Chan & Yiu, 2002), difficulties isolating individual attributes in complex voice patterns (corresponding to "generic" anchor stimuli versus custom anchors that matched each of the test voices), scale resolution (an EAI scale versus a visual analog scale; e.g., Yiu & Ng, 2004; see e.g. Toner & Emanuel, 1989, for discussion of other scaling approaches), and the extent to which a voice possesses the attribute being measured. These four factors accounted for over 84% of the observed variance in listener agreement levels in that study, indicating a good match between the psychometric model and patterns of listener agreement/disagreement. Further, listeners agreed exactly in their ratings 96% of the time when performing a task that controlled these sources of error, compared to 28% exact agreement for a traditional 6-point rating scale (which controls none of these sources of error).

These two approaches to the issue of listener agreement in voice quality ratings imply rather different methods of assessing voice quality. Because Shrivastav and colleagues do not attribute random and criterion errors to any particular sources, an experimenter applying this approach simply records each rating response, eliminates random error by averaging, and finally converts the mean value to a z-score, thus correcting errors after listeners have already chosen their responses. In contrast, Kreiman et al. (2007) proposed controlling error (which they view as resulting predominantly from non-random task-related factors) during the measurement process by applying a method-of-adjustment task in which voice quality is quantified as the settings of a speech synthesizer a listener selects to produce a voice sample whose quality matches the natural target voice under assessment. The use of a matching task eliminates listeners' dependence on unstable internal standards for the quality being measured, and also minimizes the tendency of listeners to agree better about stimuli near the endpoints of a scale. In addition, in this task listeners manipulate each synthesizer parameter in the context of the entire voice pattern, a process that makes it easier for the listener to isolate that dimension in the overall pattern (Kreiman & Gerratt, 2000). Finally, the method-of-adjustment approach uses scales whose resolution can be optimized based on listeners' sensitivity to each parameter under consideration. Because this measurement process controls previously-documented sources of rating variability, in this framework averaging

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and standardizing can at best provide only a small additional increment in overall agreement levels.

These differences derive in part from the rather different purposes of the two lines of research. The approach applied by Shrivastav and colleagues assumes that listener disagreements are always error of one type or the other, because the true rating for each voice in this view is entirely a function of the voice, with the listener serving as a virtual acoustic analysis system. In contrast, the approach applied by Kreiman and colleagues treats disagreements as data about the cognitive processes involved in perceiving complex auditory patterns, not as error, and examines disagreements as evidence about the processes involved in mapping from a signal to a response. By analogy, two people in the same room may differ significantly in their assessment of the temperature, with one individual perceiving the room as unpleasantly hot, while the other complains bitterly of the cold and puts on an extra sweater. If the measurement goal is to gauge the true temperature of the room (in other words, using humans as virtual thermometers), these differences in perception are error, and the best estimate of the true temperature can indeed be obtained by averaging together a number of such judgments. However, if the question concerns how people perceive temperature (the interaction between the physical world and the sensation evoked), both responses about the temperature are valid, and this kind of variability is data about perceptual processes that must be modeled, and not error.

The issue of perceptual processes is especially important in the case of complex, multidimensional voice stimuli. Although perceived quality has its origins in a physical sound wave that can be objectively measured (as perceived temperature has its origins in the physical temperature of the room), recent models of voice quality derived from behavioral and neuropsychological data (e.g., Van Lancker, Kreiman, & Emmorey, 1985a; Van Lancker, Kreiman, & Wickens, 1985b; Li & Pastore, 1995; Schweinberger, Herholz, & Stieff, 1997; Andics et al., 2010; Kreiman & Sidtis, 2011) provide converging evidence that listeners do not perceive voice quality as the sum of a number of separate features, but instead as an integral pattern (cf. Melara & Marks, 1990). These data are also consistent with the fact that the largest source of measurement error in quality assessment protocols is listeners' inability to isolate individual features in complex voice patterns (Kreiman et al., 2007). In the case of temperature, it is straightforward to compare listeners' judgments to objective (and unidimensional) measures of temperature; but the complexity of voice patterns necessarily introduces cognitive processes involved in identifying, isolating, and assessing different facets of the stimuli during quality assessment, all of which offer opportunities for the introduction of non-random variability in listeners' quality judgments.

Despite these differences in method and theory, however, it remains possible that the two measurement approaches still provide equivalent information about voice quality, because both seek to control measurement error (one a priori, and one post hoc). To assess this possibility, we undertook three experiments. The first two sought to replicate the findings reported in Shrivastav et al. (2005), and to derive a set of breathiness ratings for use in Experiment 3. Experiment 3 applied multidimensional scaling to these ratings and to ratings from Kreiman et al. (2007), to determine the extent to which the two measurement techniques produce comparable assessments of the breathiness of the stimulus voices across listeners.

Experiment 1

Method

Stimuli—Stimuli were identical to those used in our previous study (Kreiman et al., 2007), and are described in detail in that paper. Briefly, 40 voices (20 male speakers, 20 female

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

speakers) were selected at random from a library of pathological voices recorded under identical conditions. No attempt was made to select stimuli that possessed any particular quality, nor did we attempt to create a continuum from mild to severe breathiness, although voices did span the range from near-normal to severely deviant in terms of overall level of pathology. A 1-s sample was excerpted from the middle of a sustained /a/. Each sample was copied using a custom formant synthesizer optimized for precisely modeling pathologic voice quality (Kreiman, Antoñanzas-Barroso, & Gerratt, 2010). A listening pretest confirmed that listeners could not distinguish the synthetic stimuli from the original natural tokens at above-chance levels. Stimuli were normalized for peak amplitude (by setting the largest value in the file to 0 dB and adjusting other values relative to that) prior to presentation.

Listeners and Procedure—Twenty naïve listeners (UCLA students and staff members; 15 females, 5 males; mean age = 28.3 years; sd = 13.1 years) participated in this experiment. None had participated in the pretest. All reported normal hearing. All provided informed consent according to procedures approved by the UCLA Institutional Review Board. Listeners heard the complete set of 40 stimulus voices 10 times, for a total of 400 trials/subject. Stimuli were completely rerandomized for each presentation and listener. Listeners heard the stimuli one at a time in a double-walled sound suite, in free field at a comfortable constant listening level, and rated each voice on a 5 point scale for breathiness (where 1 indicated “not breathy at all” and 5 meant “severely breathy”). They were able to play each voice as often as they wished before responding. Testing lasted about one hour.

Table 1 summarizes the experimental details in comparison to those used by Shrivastav et al. (2005). Some differences in method are apparent. First, the present experiment used a larger, more varied set of voices, to determine if the averaging/standardizing approach generalizes across voice qualities and fundamental frequencies. If errors are truly random, agreement levels should not vary with this change in the listening context. Stimuli in the present study are also longer in duration than those used by Shrivastav et al., but providing listeners with longer voice samples should enhance agreement, not lower it, by providing listeners with more information on which to base their ratings. Finally, the present study used naïve listeners, not speech pathology students. This was unavoidable given the lack of a speech training program at UCLA, but differences between groups in experience are rather small, and previous studies suggest that expertise increases rather than reduces rating variability (Kreiman, Gerratt, & Precoda, 1990). On the whole, then, differences in experimental protocols should favor replication of previous findings.

Results

Learning effects and rating variability—To determine whether learning/practice effects occurred across the repeated trials of this protocol, following the procedures described by Shrivastav et al., we first calculated the probability that two listeners would agree exactly in their ratings of a given voice, across all voices for all possible pairs of listeners. Contrary to their findings, one-way repeated measures ANOVA indicated a significant increase in the probability of exact agreement (from an average of 0.32 to 0.37, where chance equals 0.2) from the first two to the last two trials ($F(1, 79) = 27.12$, $p < .05$), indicating that practice effects did accrue cumulatively across the entire test. Note that these effects, although significant, are relatively small, and differences between adjacent blocks of trials were not significant.

Learning effects may also produce a gradual reduction in rating variability across trials. To test this possibility, we calculated the standard deviations (sd) for the first three and for the last three ratings of each voice for each listener, and then compared these values using a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

matched samples t-test. We found a significant difference in response variability from the beginning to the end of the test (mean sd for the first three repetitions = 0.94; mean sd for the last three repetitions = 0.80; matched samples $t(39) = 8.23$, $p < .05$). Two explanations for this finding are possible. Listeners may remember the ratings given to individual voices, in which case changes in variability across the test should be independent of mean ratings. Alternatively, listeners may not remember individual stimuli, but instead may normalize their ratings for the overall range of stimuli presented over the course of the experiment. In this case, variability should improve most near the endpoints of the scale, and least in the mid-range. Regression analysis relating the overall mean rating for each voice to changes in the coefficient of variation (the standard deviation normalized for the mean) from early in the test to late in the test showed no significant relationship between these variables [$F(1,38) = 0.05$, n.s.], consistent with the hypothesis that learning individual voices contributed to decreased rating variability across the test.

Finally, rating variability was lowest overall for the most severely breathy voices, and highest overall for the least breathy voices, at both the beginning and the end of the test [$F(1, 73) = 18.57$, $p < .05$; Figure 1]. This indicates that listener agreement levels depend on the severity of the quality being rated, as previously reported (Kreiman & Gerratt, 1998;Kreiman et al., 2007).

Interrater agreement—Shrivastav et al. (2005) reported that the probability that two listeners would agree exactly increased from an average of .43 for single ratings to .90 when 10 ratings of a voice were averaged together and the mean ratings standardized. The effects of averaging and standardizing in that study were approximately equal: combining the two improved agreement levels by an additional 10–14%. This result did not replicate, either in the pattern or in the extent of improvement in listener agreement (Table 2). A two-way repeated measures ANCOVA (within-subject factors = the number of ratings averaged together and the presence vs. absence of standardization; covariate = the mean rating for each voice) showed a main effect of standardization on overall agreement levels [$F(1, 38) = 18.21$, $p < .01$] and a significant effect of the covariant [$F(1, 38) = 6.86$, $p < .01$], but no significant main effect of the number of ratings averaged [$F(9, 342) = 3.58$, n.s.]. In addition, all two-way interactions were significant [number of ratings averaged by standardization condition: $F(9, 342) = 13.68$, $p < .01$; number averaged by mean rating: $F(9, 342) = 3.89$, $p < .01$; standardization condition by mean rating: $F(1, 38) = 10.24$, $p < .01$], as was the three-way interaction between the number of ratings averaged, standardization condition, and the mean rating [$F(9, 342) = 12.92$, $p < .01$].

Figure 2 shows these effects. Each frame in this figure shows the change in the probability of exact rater agreement for two averaging conditions, plotted as a function of the overall mean rating for the voice. In panels A and D the y axis represents the difference in the probability of exact agreement for the average of two ratings versus a single rating; in panels B and E the y axis represents the difference in the probability of exact agreement for the average of five ratings versus a single rating; and in panels C and F the y axis represents the difference between the average of 10 ratings and a single rating. When a plotted difference is positive, listener agreement improved with the increase in the number of voices averaged; when the difference is negative, agreement decreased with increased averaging. The horizontal line in each figure represents zero difference, or no change in agreement with increased averaging; the oblique line represents the best linear fit to the data. Panels A-C show unstandardized ratings, and panels D-F show standardized ratings.

As this figure shows, the relationship between the effects of averaging and the overall mean rating (which in this model represents the estimated true value of breathiness for each voice) depends on the standardization condition. When ratings were unstandardized, increased

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

averaging improved agreement when voices were moderately or severely breathy, but decreased agreement when voices were mildly breathy. In contrast, when ratings were standardized, listener agreement increased with increased averaging for mildly and moderately breathy voices, but remained relatively constant across the remainder of the scale. These patterns cancel each other out, resulting in no main effect of the number of voices averaged together, but significant effects of standardization and mean rating were observed, as described above.

Discussion

These results represent a partial failure to replicate the findings of Shrivastav et al. (2005). Although we did find a significant effect of standardization condition on interrater agreement, no main effect of the number of ratings averaged together occurred (and post-hoc tests confirmed that no two conditions differed significantly; $p > .05$). In addition, the present experiment produced a main effect of the mean rating on listener agreement, along with an array of significant interaction effects. Some of these differences can be attributed to differences in the statistical approach applied. Shrivastav and colleagues conducted a series of one-way ANOVAs without repeated measures, which cannot reveal interaction effects, and did not include a covariate. We note that the F value reported for the effect of averaging in their paper (9.27, $p < 0.001$), while significant, is rather small relative to the number of degrees of freedom (9, 260), suggesting that this statistically-reliable result still accounted for relatively little variance in the underlying data. It is possible that when this variance was partitioned among an additional independent variable and a covariate in the present analysis that no main effect of averaging remains. In any case, differences in statistical approach do not explain differences in the magnitude of the effects of averaging and standardizing between the two experiments (Table 2). Even if a different analysis reveals a significant effect of averaging on agreement, the effect is much smaller in the present data set, as is the effect of standardization on agreement.

Another possible explanation for the apparent failure to replicate is that the smaller (27 vs. 40 speakers) and more homogeneous set of voices (all female, selected for breathiness) used by Shrivastav et al. helped listeners focus better on the target dimension, leading to better agreement. To test this hypothesis, we conducted a second experiment using a smaller set of stimuli that included only breathy female voices.

Experiment 2

Method

Stimuli—Listeners in this experiment heard the same 20 synthetic female voice samples used in Experiment 1. However, these stimuli were manipulated slightly to remove features not related to breathiness. The voice source was altered in 6 cases to increase sinusoidality and consequently decrease the level of high-frequency harmonic excitation; bandwidths of the higher formants were broadened in 6 cases to further reduce high frequency harmonic energy; a vocal tremor was removed from one sample; and the noise-to-harmonics ratio was increased (4 voices) or decreased (4 voices) as needed to create a set ranging from −29.8 dB to −9.3 dB, corresponding to a range from mild to very severe levels of turbulent noise. No changes were made to nine of the voices. The final stimulus voices were all primarily breathy, and ranged from mild to severe, in the opinion of the authors.

Listeners and procedure—Twenty naïve listeners (13 females, 7 males) participated in this experiment after providing informed consent as described above. They averaged 29.2 years of age (sd = 11.4) years). Listeners heard the complete set of 20 stimuli 10 times, for a total of 200 trials/subject. The stimulus set was re-randomized prior to each presentation. All

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

other methods were identical to those used in Experiment 1. The complete task took about 20 min to complete.

Results and Discussion

A one-way repeated measures ANOVA showed no significant changes in the probability of exact agreement between pairs of raters from the first two ratings [mean p(exact) = 0.30] to the last two ratings [mean p(exact) = 0.31; $F(1,39) = 0.65$, n.s.]. In addition, a comparison of the coefficients of variation for the first three ratings to those for the last three ratings showed no changes from the beginning to the end of the test [matched samples $t(19) = 0.03$, n.s.] These results indicate that learning effects did not occur in this experiment as they did in Experiment 1, and suggest that learning effects are a function of the homogeneity of the set of stimuli. This further suggests that listening context (a non-random factor) contributes to variability in listeners' judgments of voice quality, as we have argued previously (e.g., Gerratt, Kreiman, Antoñanzas-Barroso, & Berke, 1993).

Table 3 shows the mean probability of exact rater agreement for the different experimental conditions. A two-way repeated measures ANCOVA (number of ratings averaged together by standardization condition; covariate = mean rating for each voice) revealed no significant main effects of the number of ratings averaged together [$F(9, 162) = 1.86$, n.s.] or of standardization [$F(1, 18) = 5.02$, n.s.] on listener agreement rates. (Post-hoc comparisons confirmed that no averaging level differed from the unaveraged ratings; $p > .01$.) A significant effect of the covariate on agreement levels occurred [$F(1, 18) = 10.66$, $p < .01$], as did a significant interaction between standardization condition and the number of ratings averaged together [$F(9, 162) = 4.85$, $p < .01$] and a significant three-way interaction [$F(9, 162) = 2.76$, $p < .01$]. As Figure 3 shows, the relationships among variables in this experiment are essentially identical to those found in Experiment 1, despite the smaller, more homogeneous set of voices studied here. Our failure to replicate the findings of Shrivastav et al. (2005) therefore cannot be attributed to the set of voices studied.

Experiment 3

The results of Experiments 1 and 2 suggest that averaging multiple ratings of a voice and/or standardizing the mean rating are less effective as a means of dealing with interrater disagreements about breathiness than previously reported (Shrivastav et al., 2005). Despite this fact, it remains possible that averaging-and-standardizing methods and the method-of-adjustment approach to quality assessment (Kreiman et al., 2007) still produce equivalent measures of the perceived breathiness of the voices. That is, if in fact variability in listeners' breathiness ratings is noise (Patel, Shrivastav, & Eddins, 2010), then ratings gathered using the average-and-standardize approach should yield the same perceptual distances between stimuli for all the listeners in a study (because averaging and standardizing ratings has hypothetically corrected all measurement errors, leaving only the true score for each stimulus voice). To investigate this issue, we used multidimensional scaling (MDS: Kruskal & Wish, 1978; Murry, Singh, & Sargent, 1977; Kreiman, Gerratt, & Berke, 1994) to derive a "breathiness" space for each listener in Experiment 1. In these analyses, differences between individual stimuli in the ratings they received were treated as measures of the "distance" between stimuli in their level of breathiness. MDS takes these distances and generates coordinates for each stimulus in an n-dimensional space, such that stimuli that are similar in their level of perceived breathiness are close together, and stimuli that differ in breathiness are farther apart. Dimensions of this space are often interpreted with reference to characteristics of the stimuli, and are considered to reflect the perceptual structure of the stimulus set (see e.g. Schiffman, Reynolds, & Young, 1981, for more discussion). Thus, by comparing stimulus coordinates in the spaces derived by these analyses, it becomes possible to determine how similar listeners are in the perceptual strategies that underlie their ratings.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

To the extent that stimuli pattern similarly in the spaces for different listeners, the set of ratings may be considered reliable and valid. Note that MDS analyses like these are descriptive: They reveal the structure of a given data set, but conclusions can only be generalized to other sets of data when the stimulus set has been chosen to permit such generalization (which is not the case here; but see Kreiman & Gerratt, 1996, for an example of this kind of study).

Method

Two sets of multidimensional scaling analyses were undertaken. The first used the standardized average of the 10 ratings gathered in Experiment 1. For comparison purposes, we also performed parallel analyses of the data gathered in Kreiman et al. (2007) using a method-of-adjustment task to assess the breathiness of the same stimuli. In that task, 20 listeners adjusted the noise-to-harmonics ratio (NHR; one commonly-proposed acoustic correlate of breathiness; Hillenbrand, Cleveland, & Erickson, 1994; Klatt & Klatt, 1990; Yiu & Ng, 2004; Shrivastav & Sapienza, 2003; Patel et al., 2010) until the synthetic stimuli exactly matched the target voices. For both sets of data, we calculated the difference between each pair of stimuli in the ratings (or NHR values) they received from each listener, and used these "distances" as input into non-metric individual differences (three-way) multidimensional scaling analyses using Systat software (version 12.02; Systat Software, Inc., Chicago, IL). Each analysis included 20 lower-half matrices (one from each listener) containing the distances between each pair of stimuli for that listener. In addition, separate two-way scaling analyses were run for each individual listener. Because the input distances represent the difference between pairs of stimuli in breathiness, solutions were calculated in one dimension only.

Results and Discussion

Every solution for the two-way analyses of data from individual listeners in both experiments accounted for 100% of the variance in the original breathiness/NHR data, indicating perfect mappings between the original ratings and the coordinates in the one-dimensional spaces. However, when listener data were combined in three-way group analyses, the one-dimensional scaling solution for Experiment 1 data accounted for 15% of the variance in the underlying data, indicating that the perceptual strategies for individual listeners had little in common. In contrast, the three-way group scaling solution for the method-of-adjustment data from Kreiman et al. (2007) accounted for 94.6% of the variance in the underlying dissimilarity data, indicating overall agreement among listeners about the perceptual relationships among the different stimuli. To estimate the inter-rater similarities in these individual perceptual spaces, for all possible pairs of listeners in each group we calculated the correlations between stimulus coordinates in their personal perceptual spaces. For the Experiment 1 data, correlations ranged from 0.03 to 0.96, with a mean of 0.66 (sd = 0.20). In contrast, for the data from Kreiman et al. (2007), correlations were significantly higher [range = 0.72 to 0.97; mean = 0.89; sd = 0.06; $F(1, 378) = 11.22$, $p < .01$].

These results indicate that individual listeners in Experiment 1 did not consistently use comparable perceptual strategies, despite the fact that data from each individual listener could be perfectly modeled in a single dimension. In contrast, listeners from our previous study agreed much better in their perceptual strategies. We conclude that, despite any increases in interrater agreement that may be gained by averaging and standardizing procedures, these procedures do not generate rating scores with the same informational content across listeners. In contrast, ratings gathered with the method of adjustment task are characterized by high levels of interrater agreement (96%), but also by homogeneous perceptual strategies. We conclude that differences in perceptual strategies exist in

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

breathiness ratings beyond random and criterion errors addressed by averaging and standardizing, as the model proposed in our previous study describes.

General Discussion

Results of these experiments demonstrate that averaging many ratings from a single rater and standardizing the mean value may not provide desired increases in interrater agreement levels, and at the same time leave behind significant variations among listeners in their perceptions of breathiness. In contrast, near-perfect concord between listeners was achieved, both in the ratings given and in the perceptual strategy behind those ratings, when the factors underlying rater variability were controlled through use of a method-of-adjustment task (Kreiman et al., 2007). These findings are consistent with the view that cognitive and perceptual factors, in addition to random and criterion errors, are in play when listeners judge the quality of complex auditory stimuli like voices. We interpret these effects in the context of a cognitive model in which voice quality is perceived as an integral pattern, and not as a set of distinctive features (Kreiman & Sidtis, 2011), although much more research is needed regarding the specific processes involved.

Traditional average-and-standardize approaches to modeling perception of complex voice stimuli cannot control or correct for such higher-level factors, because the underlying psychometric theory was developed largely to account for listeners' judgments of acoustically simple sounds, like sine waves. In those studies, the opportunities for variation in perceptual strategy were quite limited, most variability *is* random, and psychometric modeling need not account for the manner in which listeners cope with multiple acoustic attributes, because stimuli are low-dimensional. When stimuli are complex, however, factors beyond random errors and consistent biases (including how attention is allocated to different facets of the stimuli and interactions among dimensions; see e.g. Neuhoff, 2004, for discussion) affect the precise response a listener gives in a quality assessment task. The present data confirm that listeners' responses to complex stimuli cannot be completely described and understood without attention to such higher-level, cognitive factors.

Thus, it appears that the psychometric models proposed by Shrivastav et al. (2005) and by Kreiman et al. (2007) account for different aspects of the perception of voice quality. The model proposed by Kreiman et al. (2007) specifies a set of higher-level determinants of listeners' responses (difficulties isolating individual attributes in complex voice patterns, instability of listeners' internal standards for different qualities, scale resolution, and the extent to which a voice possesses the attribute being measured as the first stage in accounting for agreement) as the first stage in accounting for agreement. Controlling these factors results in high levels of agreement among listeners because it allows them to converge on highly similar perceptual strategies. Once such factors are controlled, a small amount of interrater variability (on the order of 5%) remains. If this level of error is of concern, average-and-standardize methods can reasonably be applied at this point. However, as these data demonstrate, averaging and standardizing do not control differences in perceptual strategy among listeners, and are unlikely to prove adequate as the sole means of addressing variations in voice quality perception.

Acknowledgments

This research was supported by NIH/NIDCD grant DC01797. We thank Norma Antoñanzas-Barroso for significant programming support. We also thank our anonymous reviewers for their thoughtful and very helpful comments on a previous version of this paper.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

References

Andics A, McQueen JM, Petersson KM, Gál V, Rudas G, Vidnyánszky Z. Neural mechanisms for voice recognition. To appear in Neuroimage. 2010 in press.

Chan KMK, Yiu EML. The effect of anchors and training on the reliability of perceptual voice evaluation. Journal of Speech, Language, and Hearing Research. 2002; 45:111–126.

Eadie TL, Doyle PC. Classification of dysphonic voice: Acoustic and auditory-perceptual measures. Journal of Voice. 2005; 19:1–14. [PubMed: 15766846]

Emanuel FW, Lively MA, McCoy J. Spectral noise levels and roughness ratings for vowels produced by males and females. Folia Phoniatrica. 1973; 25:110–120. [PubMed: 4700754]

Gerratt BR, Kreiman J, Antoñanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. Journal of Speech and Hearing Research. 1993; 36:14–20. [PubMed: 8450655]

Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. Journal of Speech and Hearing Research. 1994; 37:769–778. [PubMed: 7967562]

Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America. 1990; 87:820–857. [PubMed: 2137837]

Kreiman J, Antoñanzas-Barroso N, Gerratt BR. Integrated soTo appear. Behavior Research Methods. 2010 in press.

Kreiman J, Gerratt BR. The perceptual structure of pathologic voice quality. Journal of the Acoustical Society of America. 1996; 100:1787–1795. [PubMed: 8817904]

Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. Journal of the Acoustical Society of America. 1998; 104:1598–1608. [PubMed: 9745743]

Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. Journal of the Acoustic Society of America. 2000; 108:1867–1876.

Kreiman J, Gerratt BR, Berke GS. The multidimensional nature of pathological voice quality. Journal of the Acoustical Society of America. 1994; 96:1291–1301. [PubMed: 7962996]

Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. Journal of the Acoustic Society of America. 2007; 122:2354–2364.

Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. Journal of Speech and Hearing Research. 1990; 33:103–115. [PubMed: 2314068]

Kreiman, J.; Sidtis, D. Foundations of Voice Studies. Boston: Wiley-Blackwell; 2011.

Melara RD, Marks LE. Interaction among auditory dimensions: Timbre, pitch, and loudness. Perception and Psychophysics. 1990; 48:169–178. [PubMed: 2385491]

Kruskal, JB.; Wish, M. Sage University Paper series on Quantitative Applications in the Social Sciences, number 07-011. Sage Publications; Newbury Park, CA: 1978. Multidimensional Scaling.

Murry T, Singh S, Sargent M. Multidimensional classification of abnormal voice qualities. Journal of the Acoustical Society of America. 1977; 61:1630–1635. [PubMed: 893810]

Neuhoff, JG. Ecological psychoacoustics: Introduction and history. In: Neuhoff, JG., editor. Ecological Psychoacoustics. Amsterdam: Elsevier; 2004. p. 1-13.

Patel S, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: A comparison of psychophysical methods. Journal of Voice. 2010; 24:168–177. [PubMed: 19185451]

Schiffman, SS.; Reynolds, ML.; Young, FW. Introduction to Multidimensional Scaling: Theory, Methods, and Applications. New York: Academic; 1981.

Schweinberger SR, Herholz A, Stief V. Auditory long-term memory: Repetition priming of voice recognition. Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology. 1997; 50:498–517.

Shrivastav R, Sapienza CM. Objective measures of breathy voice quality obtained using an auditory model. Journal of the Acoustical Society of America. 2003; 114:2217–2224. [PubMed: 14587619]

Shrivastav R, Sapienza C, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. Journal of Speech, Language, and Hearing Research. 2005; 48:323–335.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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]

Van Lancker D, Kreiman J, Emmorey K. Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. Journal of Phonetics. 1985a; 13:19–38.

Van Lancker D, Kreiman J, Wickens TD. Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. Journal of Phonetics. 1985b; 13:39–52.

Yiu E, Ng CY. Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. Clinical Linguistics and Phonetics. 2004; 18:211–229. [PubMed: 15151192]

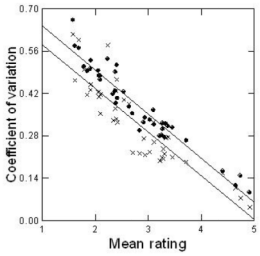

Figure 1.
Rating variability (measured as the coefficient of variation) for the first three and last three ratings each voice received, both plotted as a function of the overall mean rating for the same voice. Coefficients of variation for the first 3 ratings are plotted with filled circles; those for the last three ratings are plotted with stars. All values are averaged across listeners.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

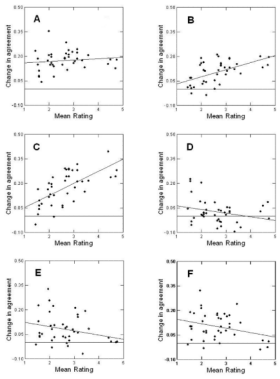

Figure 2.

Changes in the probability of exact rater agreement as a function of the overall mean rating for the voice for the data in Experiment 1. In panels A and D, the y axis represents the difference in the probability of exact agreement for the average of two ratings versus a single rating; in panels B and E, the y axis represents the difference in the probability of exact agreement for the average of five ratings versus a single rating; and in panels C and F, the y axis represents the difference between the average of 10 ratings and a single rating. Panels A, B, and C show unstandardized ratings, and panels D, E, and F show standardized ratings. Positive values on the y axis represent improved listener agreement with the increase in the number of voices averaged. The horizontal line in each figure represents zero difference, or no change in agreement with increased averaging. The best linear fit to the data is also shown in each panel as an oblique line.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

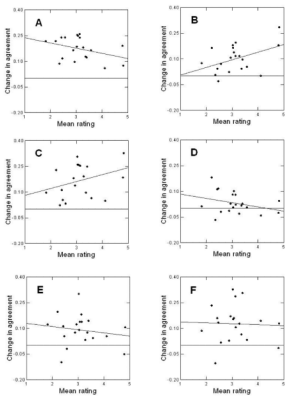

Figure 3.

Changes in the probability of exact rater agreement as a function of the overall mean rating for the voice for the data in Experiment 2. In panels A and D, the y axis represents the difference in the probability of exact agreement for the average of two ratings versus a single rating; in panels B and E, the y axis represents the difference in the probability of exact agreement for the average of five ratings versus a single rating; and in panels C and F the y axis represents the difference between the average of 10 ratings and a single rating. Panels A-C show unstandardized ratings, and panels D-F show standardized ratings. Positive values on the y axis represent improved listener agreement with the increase in the number of voices averaged. The horizontal line in each figure represents zero difference, or no change in agreement with increased averaging. The best linear fit to the data is also shown in each panel as an oblique line.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Details of the experimental methods used in the present experiment and in Shrivastav et al. (2005)

The present experiment	**Shrivastav et al. (2005)**
20 male, 20 female speakers; randomly selected	27 female speakers; samples preselected for breathiness
1 sec /a/	500 ms /a/
5 male, 15 female listeners (naïve)	10 female listeners (1 class in voice)
Stimuli presented in free field	Stimuli presented via headphones
1 listening session; 10 repetitions (complete list in 10 random orders; stimuli completely rerandomized for each presentation/listener) = 400 trials	1 listening session; 10 repetitions (complete list in 10 random orders; list order randomized) = 270 trials

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

The probability of exact agreement between two listeners, for different conditions of averaging and standardizing in Experiment 1. Data from Shrivastav et al. (2005) are included for comparison.

	Mean p(exact) across all voices	
Kind of rating	**The present research**	**Shrivastav et al. (2005)**
A single rating	0.31	0.43
Average of 2 ratings	0.49	0.74
Average of 5 ratings	0.42	0.67
Average of 10 ratings	0.49	0.70
A single standardized rating	0.43	0.76
Standardized average of 2 ratings	0.45	0.82
Standardized average of 5 ratings	0.51	0.87
Standardized average of 10 ratings	0.53	0.90

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

The probability of exact agreement between two listeners, for different conditions of averaging and standardizing in Experiment 2.

Kind of rating	p(exact)
A single rating	.31
Average of 2 ratings	.49
Average of 5 ratings	.41
Average of 10 ratings	.47
A single standardized rating	.36
Standardized average of 2 ratings	.39
Standardized average of 5 ratings	.45
Standardized average of 10 ratings	.49