Perceptual Voice Qualities Database (PVQD): Database Characteristics

Patrick R. Walden, *Queens, New York*

Summary: Objectives. To develop a perceptual voice quality database for educational and research purposes.
Study design. Development of a database.
Methods. A total of 296 high quality audio file recordings consisting of sustained /a/ and /i/ vowels and sentences from Consensus Auditory-Perceptual Evaluation of Voice were made in clinical environments. Nineteen experienced voice clinicians rated the audio samples using voice qualities from the Consensus Auditory-Perceptual Evaluation of Voice (without visual anchors) and GRBAS scales.
Results. The database includes samples of a wide range of voice quality severities across a wide range of speaker age and sex. Both inter- and intrarater reliabilities were established to be good for the database overall.
Conclusions. The database is housed in the Mendeley Data online repository and is free for public use.
Key Words: Database−Auditory-perceptual evaluation−Voice, GRBAS−CAPE-V.

INTRODUCTION AND PURPOSE

Despite the importance of auditory-perceptual voice evaluation in clinical settings,[1] auditory-perceptual ratings are often seen as unreliable and subjective,[2–4] especially for inexperienced listeners. Luckily, training in auditory-perceptual evaluation of voice qualities has been shown to improve the consistency of listener rating of voice, even for inexperienced listeners.[5–12] A large database of voice samples exemplifying various voice qualities of various severities across ages and sex which have been rated by experienced voice professionals would provide educators with standardized materials to better train preservice clinical voice professionals

Unfortunately, a widely available mechanism to support listener training does not currently exist. To provide these training experiences, an extensive database of voice samples exemplifying a broad range of salient voice qualities at varying levels of severity across various ages and sexes is necessary. In addition, listeners who provide ratings of voice quality must be shown to be reliable raters (intra-rater reliability) and severity ratings should be similar across raters (inter-rater reliability). Although databases of normal and dysphonic voice samples exist, they are either no longer available for purchase (eg, Massachusetts Eye and Ear Infirmary Voice Database[13]), are not freely available to the public (eg, databases built for research such as that created by Awan and Roy[14]), are not built with voice quality evaluation as a prime variable of interest,[15,16] are not in English,[17] or do not contain enough samples to allow for an in-depth training experience across a range of severities, ages, and sexes (eg, Voice Disorders: Simulations and Games[18]). Further, none which focus on voice, save the simulation and games website,[18] provide quality ratings for the voice samples. The paucity of a publicly available, in-depth, expert-rated voice quality database is a significant barrier to improving the reliability of perceptual voice evaluation through formal training in voice quality perception.

The purpose of creating the *Perceptual Voice Qualities Database* (*PVQD*)[19] included providing free, public access to quality voice recordings in order to afford expanded exposure to a wide range of voice qualities across age, impairment levels, and sex. Access to the database will allow for instructors to design and deliver quality voice-related learning experiences for preservice voice clinicians. Further, researchers interested in exploring the acoustic bases of voice quality perception can also use the database. High quality recordings consisting of the same speech stimuli which were captured with equipment and using methods which allow acoustic analysis result in data of sufficient quality for research. Last, the expert ratings, given good inter-rater reliability, provide further data points for anyone to freely explore acoustic-voice quality connections.

The *PVQD* is made up of high quality, reliably rated, clinical recordings of voice samples elicited using the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)[20] protocol. The database is housed in the Mendeley Data[21] online data repository and is free to access and use. The *PVQD* is accessed by visiting the Mendeley Data website and searching for the database by name. Each sample in the database has been rated by experienced voice clinicians using a 100-point visual analog scale to mimic CAPE-V scoring as well as the GRBAS[22] scale. The 100-point visual analog scale involves listening to a voice and marking perceived severity on a 100-mm line. The tick mark is then measured and the location along the length of the line the tick is located is the score. Qualities rated using the 100-point scale were borrowed from the CAPE-V (Overall severity, Roughness, Breathiness, Strain, Pitch, and Loudness) but without visual anchors for severity. The GRBAS scale requires listening to an individual's voice and rating the qualities of Grade, Roughness, Breathiness, Asthenia, and Strain on a

Accepted for publication October 2, 2020.
From the St. John's University, Queens, New York.
Address correspondence and reprint requests to Patrick R. Walden, St. John's University, 8000 Utopia Parkway, Queens, NY 11439. E-mail: waldenp@stjohns.edu
Journal of Voice, Vol. 36, No. 6, pp. 875.e15−875.e23
0892-1997
© 2020 The Voice Foundation. Published by Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.jvoice.2020.10.001

four-point scale (0 = normal; 1 = minimal severity; 2 = moderate severity; 3 = severe severity). Both of these methods of auditory-perceptual assessment of voice are regularly used in clinical voice evaluation.[23]

This manuscript details the *PVQD*. The methods used to collect and rate the voice samples are described. Additionally, the characteristics of the voice samples included in the database are explained in detail.

QUESTIONS

The following questions were asked:

1. What are the database speakers' age, gender and diagnostic characteristics?
2. What is the database raters' level of experience rating voice quality?
3. What are the voice quality features and severities represented in the database?
4. What are the intra- and inter-rater reliabilities for listener ratings?

METHOD

Creation of the database was approved by St. John's University's Institutional Review Board and all speakers consented to having their recordings made available to the public. Voice samples from the community were collected and rated on a 100-point visual analog scale and the GRBAS scale. Speakers were recruited to provide voice samples from five sites. Four of the sites were voice clinics and one site was a university program that trains speech-language pathologists (SLPs) and audiologists. All sites were in the United States of America. Speakers recruited from the voice clinics were both patients and staff of the clinics. Speakers recruited from the university program were graduate and undergraduate students as well as their friends and family members. Speakers without voice complaint/dysphonia diagnosis were recruited exclusively by word of mouth in both the voice clinics and the university program. The intent of the sampling methods was to gather as broad a set of speaker voices (with and without dysphonia) as possible.

Twenty-five raters were recruited via word of mouth among colleagues, via the American Speech-Language-Hearing Association's Special Interest Group 3: Voice and Upper Airway Disorders' online community,[24] and the University of Iowa-hosted Voiceserve electronic mailing list.[25] Criteria to rate the samples included (1) voice professional with at least 2 years' experience working with voice and voice disorders on at least a monthly basis and (2) familiarity with both the CAPE-V and GRBAS rating scales and to currently use at least one of them on at least a monthly basis. Of the 25 recruited, 19 completed the ratings. The other six left the project. Only data from the 19 who completed the ratings are included in the database.

Recording procedures

All recordings were made in a quiet clinical environment using a head-mounted condenser microphone at a 6-cm distance from the corner of the mouth and the Computerized Speech Lab (Pentax Medical) using 16-bit quantization and a sampling rate of 44.1k and the microphone gain was adjusted to prevent signal clipping. Clipped samples were discarded and the recording repeated. Samples from the voice clinics were taken during a patient's voice evaluation by the SLP. Samples from the university were taken by trained graduate and undergraduate students in a quiet area in the communication sciences and disorders department. The sample for all sites included the sustained /a/ and /i/ vowels and the sentences from the CAPE-V.[20] Each recording was edited to, as effectively as possible, remove clinician instructions provided to the speaker. Clinician-client simultaneous speaking was sometimes present and could not be removed from the file. No further audio file editing (eg, normalization) was completed.

Rating procedures

Audio samples were randomly assigned to one of six presentation blocks. Any given audio sample could only be included in a single block. Within each of the six blocks, each included audio sample was presented twice in a random order. Presentation of audio samples was accomplished using six different surveys created using Qualtrics[26] survey software (one survey for each block). Audio samples were housed in the cloud using Soundcloud[27] online hosting. A link to the cloud-based file was inserted into the online survey for presentation during ratings. Ratings were accomplished within each web-based survey. Raters were randomly assigned to a single survey. Five surveys were rated by three listeners. One of the six surveys was rated by four listeners. This difference occurred due to rater attrition.

Raters used their own computer to access their assigned survey using a link emailed to them and to listen to the samples and rate voice quality via a web-based system that included custom-made electronic scales, including a 100-point visual analog scale and the GRBAS.[22] This method of audio sample presentation was chosen to mimic the eventual listeners targeted by this database (ie, users will most likely listen to the samples on a computer). Order of presentation of the scales was the same for each file rated (first was the 100-point scale and then the GRBAS scale) and both scales were rated on the same "page" as the other (only scrolling down the "page" was required to use both scales). The rater was free to rate using whichever scale he/she wanted first. Note that severity markers/visual anchors (MI, MO, SE) *were not included* on the 100-point visual analog scale to avoid influencing the concurrent rating using the GRBAS scale. Listeners rated approximately 50 files each and each file was rated twice, in random order, for reliability measurement (for a total of approximately 100 ratings per listener). Raters were urged to rate the samples over several days to avoid fatigue and one month was allotted for each

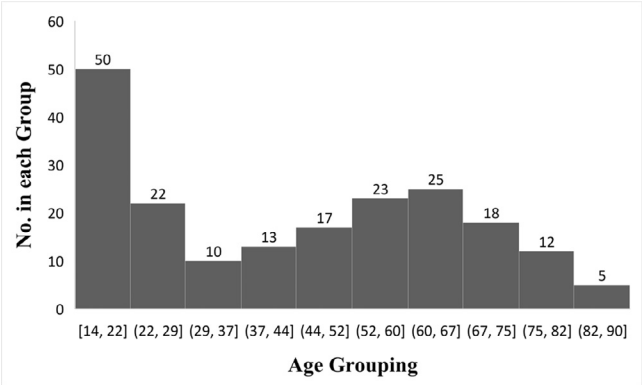


FIGURE 1. Distribution of female speaker ages. The grouping was spread out into 10 groupings based on minimum and maximum age of speakers (eg, [14,22] indicate an age grouping of 14-22 years of age).

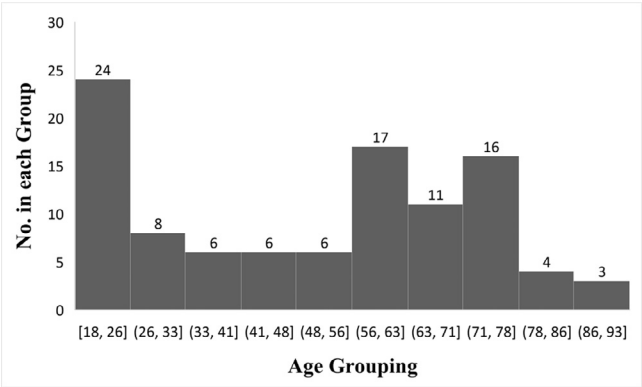


FIGURE 2. Distribution of male speaker ages. The grouping was spread out into 10 groupings based on minimum and maximum age of speakers (eg, [18,26] indicate an age grouping of 18-26 years of age).

listener to complete the ratings. Listeners could stop and close the online survey and start again at a later date/time directly from the point when the survey was last closed. Each audio sample could be rewound or fast forwarded, and listeners could listen to the samples as many times as they needed to provide a rating. Raters were instructed that accuracy was more important than speed in completing the ratings. Listener ratings were recorded and stored in a spreadsheet that was downloaded from the survey software.

Reliability procedures

Intraclass correlation (ICC) was used to assess the intra- and inter-rater reliabilities of the severity ratings on the 100-point visual analog and GRBAS scales. Intraclass correlation is not a unitary index of reliability such as Cronbach's alpha. Instead, its computation varies with the nature of the rating data and the manner in which the multiple ratings are to be used. The rating data can consist of all raters rating all targets or of different randomly selected raters rating different subsets of the targets. In the former case, it is possible to account for differences in the rating levels characteristic of different raters and to thereby assess not only consistency (ie, degree of ordinal similarity of ratings between raters) but also agreement (ie, degree of similarity in levels of ratings between raters). In the case of different raters rating different subsets of targets, it is only possible to assess reliability in terms of consistency between (or within) raters. This situation represents the case for the rating procedures for the *PVQD*, thereby limiting the reliability assessment to ascertaining only the degree of consistency between (or within) raters with the intent of generalizing to the universe of raters. This calls for the application of Shrout and Fleiss' (1979)[28] ICC(1, *m*) model, which requires the use of the mean square components from a one-way random ANOVA to compute the intraclass correlation.

The *m* in the ICC(1, *m*) designation refers to the assumption that the final ratings will be based on the means of ratings by *m* raters per target. This stands in contrast to computations that assume that the reliability assessment is intended to apply to any single rater drawn from the universe of raters. It was assumed in this database that the final ratings would be based on three raters in computing the inter-rater reliabilities, and on two repeated within-rater ratings in computing the intra-rater reliabilities. The final value of the ICC was computed by using the value of *m* (ie, 3 or 2) in the Spearman-Brown formula to correct the ICC upward to account for the higher number of ratings per target. Because some files were rated by four listeners rather than three, three raters for those files were randomly selected for inclusion in the analysis because the number of raters had to be consistent.

RESULTS

Speaker demographics

The database consists of 296 different speakers, including speakers with a voice complaint/diagnosis of dysphonia and

TABLE 1.
Overall Database Characteristics by Quality: 100-Point Visual Analog Scale

Quality Rated	Mean	Median	Mode	Minimum	Maximum
Severity	29.4	19.5	19.3	0.3	98.6
Roughness	20.7	13.7	9.7	0.1	84.8
Breathiness	19.8	12.2	5	0	99.5
Strain	21.1	12.2	4.5	0.1	96.8
Pitch	16.3	9.3	0.5	0	99.2
Loudness	18.7	8.8	0.7	0	99.2

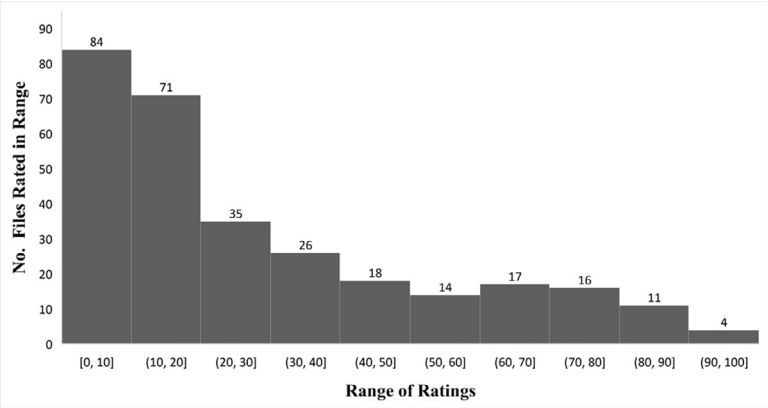


FIGURE 3. Average ratings for overall severity on the 100-point visual analog scale. The range of ratings was broken up into 10-point categories (eg, [0,10] indicated average ratings from 0 to 10).

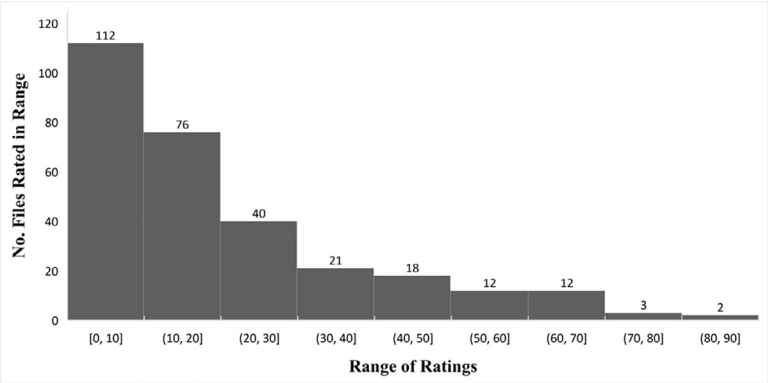


FIGURE 4. Average ratings for roughness on the 100-point visual analog scale. The range of ratings was broken up into 10-point categories (eg, [0,10] indicated average ratings from 0 to 10).

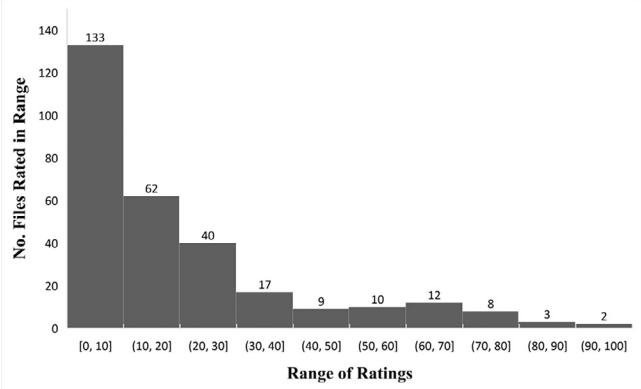


FIGURE 5. Average ratings for breathiness on the 100-point visual analog scale. The range of ratings was broken up into 10-point categories (eg, [0,10] indicated average ratings from 0 to 10).

those without complaints. Of the 296 speakers, 195 are female and 101 are male. The mean age of speaker (female or male) is 46.6 year, the median age is 48 years with a minimum age of 14 years and a maximum age of 93 years (34-year age range).

Of the speakers, 89 had no voice complaints. Sixty-eight of the speakers with no voice complaint are female (mean age of 29.1 years, median age of 21 years, minimum age of 17 years, and maximum age of 83 years). The other 21 speakers with no voice complaint are male (mean age of 29.4 years, median age of 24 years, minimum age of 18 years, and maximum age of 77 years). There are no data points regarding voice complaint/diagnosis for 20 of the speakers. Of these 20, 15 are female (mean age of 57.6 years, median age of 65 years, minimum age of 21 years, and maximum age of 88 years) and five are male (mean age of 54.6 years, median age of 55 years, minimum age of 34 years, and maximum age of 70 years).

The remaining187 speakers reported either a voice complaint or had a confirmed diagnosis of dysphonia. Of these, 112 are female (mean age of 52.4 years, median age of 55 years, minimum age of 14 years, and maximum age of 90 years) and 75 are male (mean age of 55.8 years, median age of 59 years, minimum age of 19 years, and maximum age of 93 years). Each file is labeled in the database regarding voice complaint or confirmed diagnosis status. When a specific dysphonia etiology was available, it was listed. Otherwise, each file in the database indicates no complaint or is blank when no data were available for complaint/diagnosis status. Figure 1 depicts the female speaker age distribution calculated by using the range of ages for female speakers and breaking them into 10 categories. Figure 2 depicts the same for the male speakers.

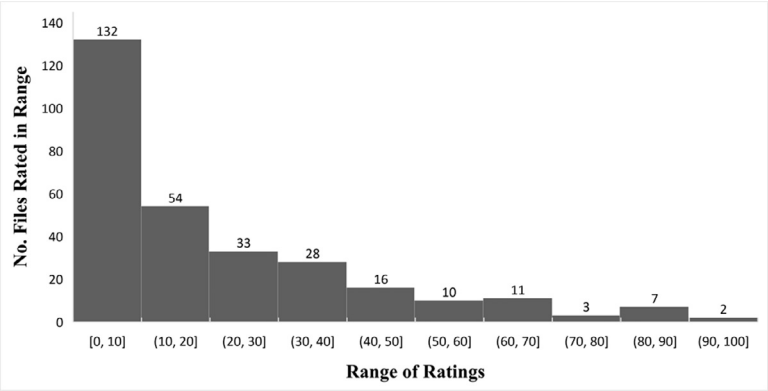


FIGURE 6. Average ratings for strain on the 100-point visual analog scale. The range of ratings was broken up into 10-point categories (eg, [0,10] indicated average ratings from 0 to 10).

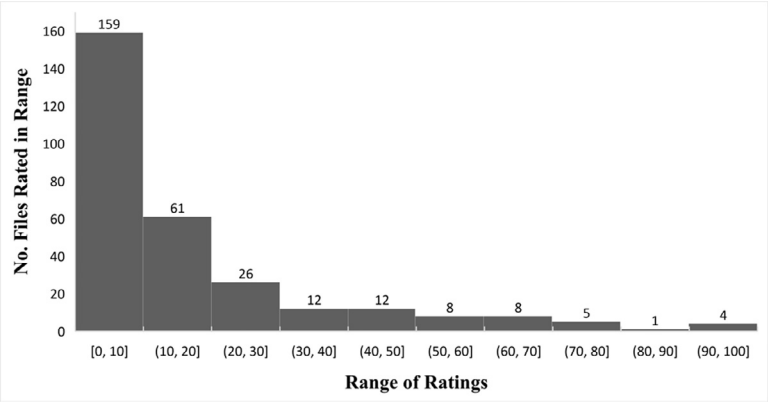


FIGURE 7. Average ratings for pitch on the 100-point visual analog scale. The range of ratings was broken up into 10-point categories (eg, [0,10] indicated average ratings from 0 to 10).

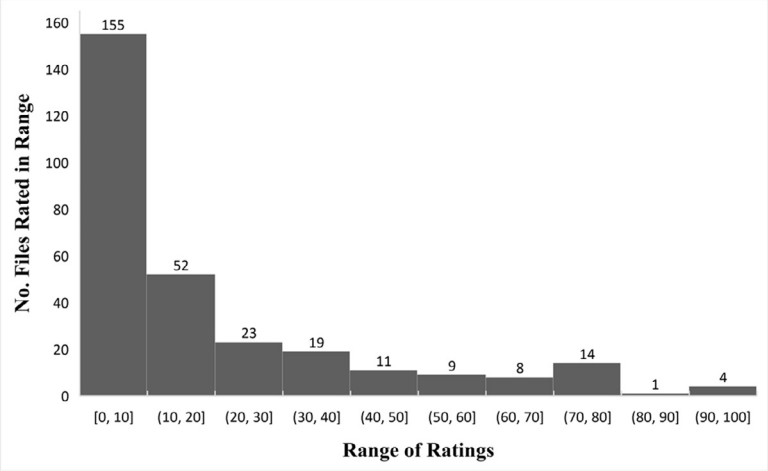


FIGURE 8. Average ratings for loudness on the 100-point visual analog scale. The range of ratings was broken up into 10-point categories (eg, [0,10] indicated average ratings from 0 to 10).

Rater demographics

A total of 19 experienced listeners rated the audio files on the 100-point visual analog scale and the GRBAS scale. All were SLPs. The SLPs reported an average of 13.6 years working as an SLP with an average of 12.5 years working with voice disorders. The years of experience ranged from 2 to 37 for both working as an SLP and working with voice disorders. The median years of experience working as an SLP was 13 years and the median for working specifically with voice disorders was 9 years.

TABLE 2.
Overall Database Characteristics by Quality: GRBAS Scale

Quality Rated	Mean	Median	Mode	Minimum	Maximum
Grade	1	0.8	0	0	3
Roughness	0.8	0.7	0	0	3
Breathiness	0.7	0.4	0	0	3
Asthenia	0.6	0.2	0	0	3
Strain	0.8	0.5	0	0	3

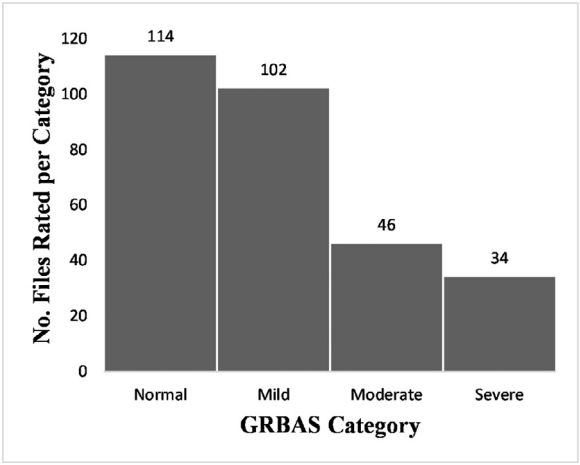


FIGURE 9. Average ratings for grade on the GRBAS scale. These are based on averages and the GRBAS categories depicted in this figure include the following thresholds: Normal = 0-0.5; Mild = 0.6-1.5; Moderate = 1.6-2.5; Severe = 2.6-3.0.

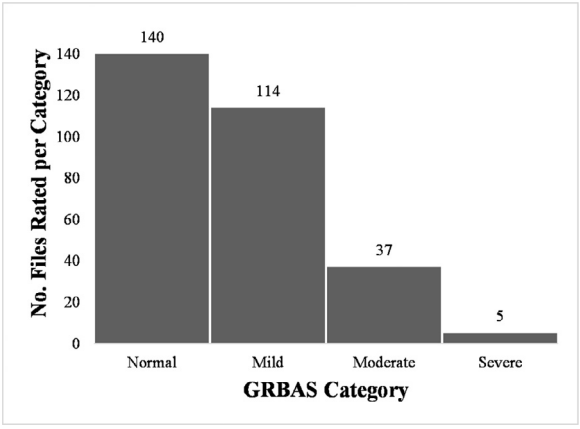


FIGURE 10. Average ratings for roughness on the GRBAS scale. These are based on averages and the GRBAS categories depicted in this figure include the following thresholds: Normal = 0-0.5; Mild = 0.6-1.5; Moderate = 1.6-2.5; Severe = 2.6-3.0.

Sample characteristics

Ratings for each audio file were averaged across all listeners. The characteristics reported below are based on these averages. Individual listener ratings for each audio sample are included in the database for both scales.

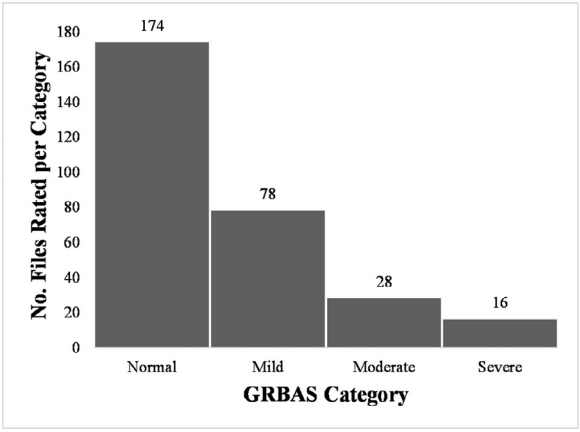


FIGURE 11. Average ratings for breathiness on the GRBAS scale. These are based on averages and the GRBAS categories depicted in this figure include the following thresholds: Normal = 0-0.5; Mild = 0.6-1.5; Moderate = 1.6-2.5; Severe = 2.6-3.0.

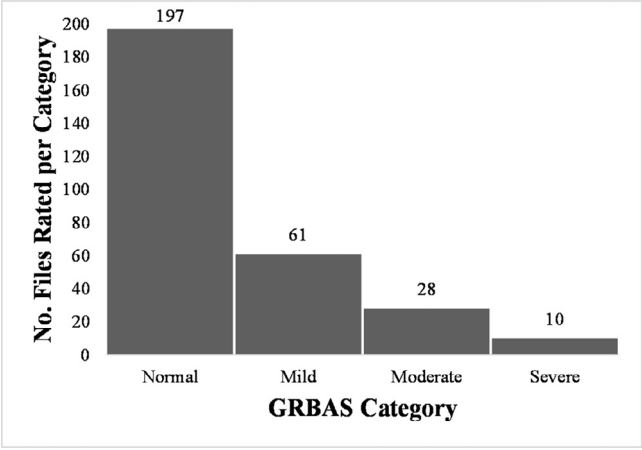


FIGURE 12. Average ratings for asthenia on the GRBAS scale. These are based on averages and the GRBAS categories depicted in this figure include the following thresholds: Normal = 0-0.5; Mild = 0.6-1.5; Moderate = 1.6-2.5; Severe = 2.6-3.0.

100-point visual analog scale characteristics

Table 1 depicts the mean, median, mode, minimum, and maximum values for the database as a whole on the 100-point visual analog scale. Averages across all raters were used to calculate these values (ie, average ratings for all audio files) and these values are presented to get a large-

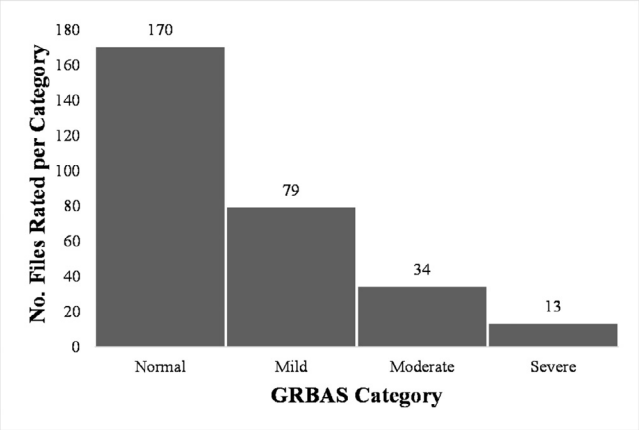


FIGURE 13. Average ratings for strain on the GRBAS scale. These are based on averages and the GRBAS categories depicted in this figure include the following thresholds: Normal = 0-0.5; Mild = 0.6-1.5; Moderate = 1.6-2.5; Severe = 2.6-3.0.

scale overview of the general severity distribution of each characteristic in the database. Figures 3-8 depict the frequency of average ratings by quality rated.

GRBAS characteristics

Table 2 depicts the mean, median, mode, minimum, and maximum values for the database as a whole on the GRBAS scale. Averages across all raters were used to calculate these values (ie, average ratings for all audio files). Figures 9-13 depict the frequency of average ratings by GRBAS quality.

Inter- and intra-rater reliability

For the 100-point visual analog scale, the overall intraclass correlation for inter-rater reliability was 0.86 (averages used as ratings), indicating a good overall inter-rater reliability

TABLE 3.
Intraclass Correlations by Quality Rated for Inter-rater Reliability

100-Point VAS	Intraclass Correlation (Averages Used as Ratings)	GRBAS	Intraclass Correlation (Averages Used as Ratings)
Severity	0.918	Grade	0.911
Roughness	0.789	Roughness	0.787
Breathiness	0.827	Breathiness	0.844
Strain	0.829	Asthenia	0.843
Pitch	0.856	Strain	0.845
Loudness	0.870		

TABLE 4.
Intraclass and Pearson Correlations Between Trials by Quality Rated for Intrarater Reliability

100-Point VAS	Intraclass Correlation (Assuming Averages Used)	GRBAS	Intraclass Correlation (Assuming Averages Used)
Severity	0.943	Grade	0.905
Roughness	0.896	Roughness	0.846
Breathiness	0.911	Breathiness	0.884
Strain	0.908	Asthenia	0.892
Pitch	0.878	Strain	0.862
Loudness	0.905		
	Pearson Correlations Between Trials by Quality Rated		Pearson Correlations Between Trials by Quality Rated
Severity	0.890	Grade	0.827
Roughness	0.814	Roughness	0.734
Breathiness	0.833	Breathiness	0.793
Strain	0.828	Asthenia	0.804
Pitch	0.772	Strain	0.757
Loudness	0.824		
	Overall Pearson Correlation Between Trials 1&2 = 0.839		Overall Pearson Correlation Between Trials 1&2 = 0.800

for this scale. For the GRBAS scale, overall intraclass correlation for inter-rater reliability was 0.859 (averages used as ratings), also indicating good inter-rater reliability for the GRBAS ratings. The overall intraclass correlation for intra-rater reliability on the 100-point visual analog scale was 0.913 (assuming averages used), indicating good intra-rater reliability for this scale. Overall intraclass correlation for intra-rater reliability for the GRBAS scale was 0.889, also indicating good intra-rater reliability overall. Table 3 depicts the intraclass correlations by feature (quality) rated for inter-rater reliability. Table 4 depicts both the intraclass correlations and Pearson correlations between trials by feature (quality) rated for intra-rater reliability. As can be seen in the tables, inter- and intra-rater reliability by feature was also good.

CONCLUSIONS

The *PVQD* is made up of 296 high quality audio recordings that represent a broad range of voice quality severities across age and sex. Visual depiction of the figures indicates severity levels skewed toward the "normal," "mild," and "moderate" severities. For educational purposes, this is advantageous, as perception of mild-moderate severities seems to be more difficult than normal and severe severities.[29] The inter- and intra-rater reliabilities indicate that the ratings included in the database may be used with relative confidence to create educational materials to teach auditory-perceptual evaluation of voice as well as to use the samples and ratings in research.

A limitation of the database is the lack of pediatric speakers. Although we attempted to include pediatric speakers, no samples including a speaker younger than 14 years old were collected. A further limitation is the use of a value to mark severity of pitch and loudness variation. Without further notations, it is impossible to discern whether a large pitch severity value is "too high" or "too low." Similarly, the score reported for loudness ratings could indicate "too loud" or "too soft." The database user will need to make that determination. Last, users of the database are encouraged to carefully listen to each file, as it was sometimes difficult to remove clinician instructions. Further, recordings were made in an authentic clinical environment rather in a sound-treated space. Therefore, some background noise may be present.

Acknowledgments

This project was funded by *The Voice Foundation* (Advancing Scientific Voice Research Grant). I would like to express gratitude for this support. Further, many individuals contributed to this database through collecting samples, rating samples, editing and organizing audio files, and providing guidance along the way. For their countless hours of help, I would like to thank Jackie Gartner-Schmidt, Amanda Gillespie, Leah Helou, Ryan Branski, Aaron Johnson, Stratos Achlatis, Shirley Gherson, Edie Hapner, Laurel Directo, Wendy LeBorgne, Erin Donahue, Jennifer Khayumov, Rayna Naraine, Karen Perta, Nilsa Perez, Maurice Goodwin, Christine Estes, Amy Harris, Maria Claudia Franca, Starr Cookman, Sweta Soni, Scott Sussman, Chandler Thompson, Ana Claudia Harten, Abigail Dueppen, Rachel Agron, Martha Pena, Kimberly Brownell, Gaida Hinnawi, Jenny Pierce, Wenli Chen, and Trudy Lynch. I would also like to thank Ms. Maria Russo, Executive Director of *The Voice Foundation*, for all her help and patience with me along the way.

REFERENCES

1. (ASHA) AS-L-HA. Preferred Practice Patterns for the Profession of Speech-Language Pathology. American Speech-Language-Hearing Association. doi:10.1044/policy.PP2004-00191.
2. Gerratt BR, Kreiman J, Antonanzas-Barroso N, et al. Comparing internal and external standards in voice quality judgments. *J Speech Lang Hear Res*. 1993;36:14–20. https://doi.org/10.1044/jshr.3601.14.
3. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Lang Hear Res*. 1993;36:21–40. https://doi.org/10.1044/jshr.3601.21.
4. Nagle KF. Emerging scientist: challenges to CAPE-V as a standard. *Perspect ASHA Spec Interest Groups*. 2016;1:47–53.
5. Eadie TL, Van Boven L, Stubbs K, et al. The effect of musical background on judgments of dysphonia. *J Voice*. 2010;24:93–101.
6. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19:555–573. https://doi.org/10.1016/j.jvoice.2004.08.008.
7. Helou LB, Solomon NP, Henry LR, et al. The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings of postthyroidectomy voice. *Am J Speech Lang Pathol*. 2010;19:248–258.
8. Sofranko JL, Prosek RA. The effect of levels and types of experience on judgment of synthesized voice quality. *J Voice*. 2014;28:24–35. https://doi.org/10.1016/j.jvoice.2013.06.001.
9. Sofranko JL. The effect of experience and the relationship among subjective and objective measures of voice quality. Published online 2012. Available at: https://etda.libraries.psu.edu/catalog/15247. Accessed May 7, 2017.
10. Ghio A, Dufour S, Wengler A, et al. Perceptual evaluation of dysphonic voices: can a training protocol lead to the development of perceptual categories? *J Voice*. 2015;29:304–311. https://doi.org/10.1016/j.jvoice.2014.07.006.
11. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20:527–544. https://doi.org/10.1016/j.jvoice.2005.08.007.
12. Eadie TL, Kapsner-Smith M. The effect of listener experience and anchors on judgments of dysphonia. *J Speech Lang Hear Res*. 2011;54:430–447. https://doi.org/10.1044/1092-4388(2010/09-0205).
13. Massachusetts Eye and Ear Infirmary. Voice Disorders Database (Version 1.03 Cd-Rom). *Voice Disorders Database (Version 1.03 Cd-Rom)*. Lincoln Park, NJ: Kay Elemetrics Corporation; 1994.
14. Awan SN, Roy N. Acoustic prediction of voice type in women with functional dysphonia. *J Voice*. 2005;19:268–282. https://doi.org/10.1016/j.jvoice.2004.03.005.
15. TalkBank. Voice Disorders Database. Available at: http://www.talkbank.org/. Accessed July 13, 2017.
16. University of Oxford. British National Corpus. Available at: http://www.natcorp.ox.ac.uk/. Accessed July 13, 2017.
17. Putzer M, Barry WJ. Saarbruecken voice database. Published May 23, 2007. Available at: http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4. Accessed July 13, 2017.
18. Connor N. Voice disorders: simulations & games. Available at: https://csd.wisc.edu/slpgames/index.html. Accessed June 14, 2017.

19. Walden P. Perceptual voice qualities database (PVQD). 2020;3. doi: https://doi.org/10.17632/9dz247gnyb.3.
20. Kempster G. CAPE-V: development and future direction. *SIG 3 Perspect Voice Voice Disord*. 2007;17:11–13. https://doi.org/10.1044/vvd17.2.11.
21. Mendeley Ltd. Mendeley Data. Available at: https://data.mendeley.com/. Accessed August 18, 2020.
22. Hirano M. *Clinical Examination of Voice*. New York: Springer-Verlag; 1981.
23. Kreiman J, Gerratt BR. Perceptual assessment of voice quality: past, present, and future. *Perspect Voice Voice Disord*. 2010;20:62. https://doi.org/10.1044/vvd20.2.62.
24. American Speech-Language-Hearing Association. Special Interest Group 3, Voice and Upper Airway Disorders. American Speech-Language-Hearing Association. Available at: https://www.asha.org/SIG/03/. Accessed August 14, 2020.
25. University of Iowa. Voiceserve. Published nd. Available at: https://list.healthcare.uiowa.edu/read/all_forums/?forum=. Accessed August 14, 2020.
26. Qualtrics. *Qualtrics Survey Software*. Available at: https://www.qualtrics.com/uk/. Accessed August 14, 2020.
27. Soundcloud Limited. *SoundCloud*. Available at: https://soundcloud.com/. Accessed August 14, 2020.
28. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428. https://doi.org/10.1037/0033-2909.86.2.420.
29. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice*. 2009;23:341–352. https://doi.org/10.1016/j.jvoice.2007.10.006.