Auditory-Perceptual Evaluation of Voice Quality of Cochlear-implanted and Normal-hearing Individuals: A Reliability Study

***Ana Cristina Coelho, †Alcione Ghedini Brasolotto, ‡Ana Carolina Nascimento Fernandes, §Daniela Malta de Souza Medved, ‖Eduardo Magalhães da Silva, and *Fayez Bahmad Júnior,**

**‡§‖Brasília and †Bauru, Brazil*

Summary: Objective. This study aimed to present an experience in rating voices of adults with normal hearing and adults with cochlear implants and critically examine the outcomes, discussing pros and cons of the methodology used.

Study design. This is a cross-sectional, prospective study.

Methods. One hundred and fifty voice samples, consisting of 50 sustained vowels, 50 samples of connected speech, and 50 samples of conversational speech, belonging to 25 adults with hearing impairment with cochlear implants and 25 adults with normal hearing, were perceptually analyzed for inter-rater agreement and intra-rater reliability. Three experienced judges rated the voice samples using visual analog scales of parameters considered relevant for cochlear-implanted population such as articulation, intonation, and resonance. The raters participated in three training sessions for calibration and had 1 month to complete the ratings individually. Twenty percent of the samples were repeated randomly to verify intra-rater reliability. The levels of agreement and reliability were verified using the interclass correlation coefficient.

Results. The inter-rater agreement varied widely across the parameters and speech tasks, from poor to excellent agreement. The only parameter for which the raters maintained consistently good or excellent agreement for all groups and emissions was the pitch. For intra-rater reliability, two of the raters presented excellent reliability for most parameters across all of the speech tasks, whereas one rater presented more inconsistencies.

Conclusions. In this reliability study, factors such as extensive deadline for the auditory perceptual evaluation, lack of periodic recalibration, speech tasks, and familiarity with the population studied were identified as factors that contributed to inconsistent reliability results.

Key Words: Voice–Voice quality–Auditory-perceptual evaluation–Voice assessment–Voice disorders.

INTRODUCTION

Voice assessment is part of a construct that involves acoustic and aerodynamic measures, laryngeal imaging, self-assessment, and auditory-perceptual evaluation.[1] Auditory-perceptual evaluation is commonly used[2–6] for research and clinical settings,[7,8] even with its subjective and multidimensional features.[9,10] It is subjective because the voice is by nature a perceptual phenomenon, and its characteristics are recognized based on what is perceived by a listener to be normal or altered.[11]

To achieve criteria such as reliability, validity, and responsiveness to change,[11] auditory-perceptual evaluation must go through a carefully designed process, which includes capturing of the vocal signal, selection of the vocal tasks, and selection of the instrument of evaluation and the rater panel. The most commonly used auditory-perceptual rating scales are the grade, roughness, breathiness, asthenia and strain (GRBAS) scale and the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), used by clinicians to rate the severity of the voice using one rating of overall severity and some descriptive perceptual parameters.[12] To date, however, no single method of auditory perceptual voice analysis has achieved these criteria satisfactorily, nor has any instrument been used consistently and with similar results across studies, which still confounds communication of clinical and research findings.[11] Listeners very often disagree with each other in their ratings of voice quality,[13] contributing to this phenomenon, and both intra-rater agreement and inter-rater reliability fluctuate greatly among studies,[14] even in some studies using the GRBAS scale or the CAPE-V scale.

Listeners rate the vocal quality according to their stored mental representations, which serve as internal standards and can vary between one listener and another over time and under different circumstances.[8] Some important influences on judgments include the listeners' cultural and clinical references, their academic education, the duration of the samples, the number of sessions, the environmental setting to perform the ratings,[15] the internal standards for different types of voices,[7–9,13] difficulty in isolating one particular feature of the emission, the type of scale, the magnitude of the attribute being measured,[13] the time of experience,[7] the particular features of the population that is being assessed, and the type and amount of training involved.

There may be difficulty in agreement not only on the kind of voice quality, but also on the amount of that dimension present

Accepted for publication February 22, 2017.

From the *Postgraduate Program in Health Sciences, University of Brasília, Campus Universitário Darcy Ribeiro, Brasília, Distrito Federal 70910-900, Brazil; †Speech-Language Pathology Department, Bauru School of Dentistry, University of São Paulo, Al Octávio Pinheiro Brisola, 9-75, Bauru, São Paulo 17012-901, Brazil; ‡Speech-Language Pathology Course, Centro Universitário Planalto do Distrito Federal, Avenida Pau Brasil, lote 02 S/N, Águas Claras, Brasília, Distrito Federal, Brazil; §Brasília Teaching Hospital, University of Brasília, SGAN 605, Av. L2 Norte, Brasília, Distrito Federal 70840-901, Brazil; and the ‖Speech-Language Pathology Course, University of Brasília, Centro Metropolitano, Conjunto A, lote 01, Brasília, Distrito Federal 72220-275, Brazil.

Address correspondence and reprint requests to Ana Cristina Coelho, SQNW 108, Bloco J, ap 105., Noroeste, Brasília, DF 70686-200, Brazil. E-mail: anacrisccoelho@yahoo.com.br

Journal of Voice, Vol. ■■, No. ■■, pp. ■■-■■
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2017.02.012

on a given voice.[16] Perceptual context may also cause a drift in ratings. Authors exemplify that if a listener hears a moderately rough voice after hearing several mildly rough voices, the rating for the moderate voice may become more severe, because hearing several mild voices may have shifted the listener's internal standards.[17,18] The level of agreement and reliability changes according to the severity of the attribute measured. There is a tendency of obtaining better agreement in the ratings of normal and severely altered voices. The ratings of intermediate voices produce more controversies among raters.[9] Listeners may also differ in using the range of the rating scale and how they interpret different points on the same rating scale.[17]

The matter of rater experience in perceptual ratings is controversial. Whereas some studies state that experienced raters are less reliable because their repertoires of internal standards vary more and use different rating strategies, others found that experienced raters are more consistent in their assessment.[7,9] Some authors state that experienced listeners may introduce more variability into judgments of voice quality because they use a flexible strategy to determine salient perceptual features, making continual adjustments as they fine-tune their decisions.[11,19] Also, there are reports that inexperienced raters agree on evident voice characteristics, for both pathologic and normal voices, whereas experienced raters do not easily agree.[15] Fatigue, lack of attention, and transcription errors may also interfere on reliability.[9,15,20,21]

Perceptual training or calibration is a common resource for obtaining valid and reliable results, because it increases agreement and consistency.[7] Training and exposure to a variety of voice samples help model the factors related to the listener to obtain better agreement and reliability[15,21] by calibrating and stabilizing internal standards.[7,8] Although it is not clear which type and amount of training is required to acquire and maintain consistency in the auditory-perceptual evaluation, there is sufficient evidence to support the statement that training with voice samples is valid. Also, to maintain consistency, periodic recalibrations may be necessary. In summary, training aims to capacitate the rater in using the selected rating instrument, understanding about the attributes being measured, and calibrating the presence or absence and the severity of each parameter.

Another form of addressing variability of raters' internal standards is the use of external anchors[7] (models of normal and altered voices in the same speech tasks), aiming to decrease variability.[15] However, external anchors have traditionally been used in studies designed to assess sustained vowels rather than connected speech[7] or conversational speech, which are important tasks that allow the assessment of relevant aspects of vocal function. These anchor samples serve as consistent examples that override a listener's variable internal standards and create a stable listening context for rating voice samples.[17]

Less studied altered voices, such as voices of cochlear-implanted individuals, may be more difficult for raters to assess. This can be due to lack of references, less practice or listening ability, and the specific alterations of the population, in this case, resonance and suprasegmental disorders, in addition to the glottal source alterations.

Based on these statements, the purpose of this study is to present an experience in rating voices of adults with normal hearing and adults with cochlear implant using a variety of perceptual parameters and critically examine the outcomes, discussing pros and cons of the methodology used.

METHODS

Voice samples

For this study, 50 voice samples of three different tasks were used from a research database, totaling 150 voice samples collected from 50 individuals of both genders. The voice tasks consisted of the sustained /a/ vowel, connected speech (counting from 1 to 10), and conversational speech lasting from 20 to 30 seconds at comfortable frequency and intensity. These samples belonged to 25 adults with hearing impairment with cochlear implants (CI group) and 25 adults with normal hearing (NH group), ranging in age from 18 to 45 years.

None of the individuals had history of smoking, drinking, using the voice professionally, menopause, previous laryngeal surgery, or previous voice therapy for diagnosed laryngeal alterations. All voice samples were recorded in a quiet environment, using a laptop computer, with the audio interface M-audio Fast Track Pro, AKG C512 headset positioned at 3 cm from the mouth, and *Sony Sound Forge* 10.0 software with a sampling rate of 44,100 Hz, 16 bit, mono channel, AKG and sony.

Perceptual ratings

The ratings of this study represent a first attempt to obtain satisfactory inter-rater agreement and intra-rater reliability among three speech-language pathologists using a method that consisted of consensus training followed by individual analysis of each voice sample. All raters are specialized in voice disorders and had experience in auditory-perceptual evaluation of normal and altered voices, in clinical and research contexts. The raters participated in three training sessions lasting 4 hours each, using 10 additional voice samples. In each session, the raters listened to the voice samples of both adults with CI and adults with NH using a free-field speaker in a quiet room, discussing all the parameters to be assessed and reaching a consensus on the definition, the presence or absence, and the severity of each parameter on the rating scale. To reach a consensus, the raters were allowed to listen to the voice samples as many times as necessary. In the first training session, the raters trained with the sustained /a/ vowel; in the second session, with connected speech; and in the third, with conversational speech. After the training session, each rater had 1 month to complete the ratings individually. The raters were instructed to perform the ratings in a quiet environment, use headphones always in the same volume, and not to exceed the evaluation of 20 voice samples per day. This number was determined to avoid potential perceptual errors caused by fatigue. The samples were separated by voice tasks in a randomized order. The rater had the information of age and gender of each voice sample. In addition, 20% of the samples were repeated at random to verify intra-rater reliability. The raters were free to listen to the stimuli as many times as necessary.

For the auditory perceptual evaluation, undifferentiated visual analog scales (VAS) were used. The parameters were selected from an ongoing study about the voice of individuals with hearing loss and include intelligibility, articulation, intonation, speech rate, speech-breathing coordination, laryngeal resonance, pharyngeal resonance, hyponasality, hypernasality, anterior resonance, posterior resonance, strain, breathiness, roughness, instability, pitch, loudness, and overall severity (Appendix). Hyponasality was not considered for the /a/ vowel as it is an oral vowel. The suprasegmental features (intelligibility, articulation, intonation, speech rate, and speech-breathing coordination) were considered only for the conversational speech. Although some of these parameters, such as the resonance parameters, are not commonly used on research that include auditory-perceptual evaluation of the voice, they were considered relevant for the population being studied. During the training, the raters were provided with written definition of the vocal parameters as a textual anchor, and they were allowed to consult the definitions whenever necessary while rating the voice samples individually.

Statistical analysis

The levels of inter-rater agreement and intra-rater reliability were verified using the interclass correlation coefficient (ICC). The intra-rater reliability was tested with 20% of stimuli repetition. The following ICC values were considered: less than 0.40, poor; between 0.40 and 0.59, fair; between 0.60 and 0.74, good; and between 0.75 and 1.00, excellent.[22]

RESULTS

Inter-rater agreement

Inter-rater agreement was obtained by calculating the ICC across the three raters. The ICC was calculated for the overall population and separately, for the CI group and for the NH group. The data presented consist of the average measures (reliability of different raters averaged together) and the confidence interval range (Tables 1–3). The inter-rater agreement varied widely across the parameters and speech tasks, from poor to excellent. The only parameter for which the raters maintained consistently good or excellent agreement for all groups and emissions was the pitch. For the sustained vowel, the ICC of the overall population ranged from 0.218 for laryngeal resonance to 0.823 for loudness (Table 1). For the connected speech, ICC varied from 0.210 for breathiness to 0.868 for loudness (Table 2). Finally, for the conversational speech, ICC varied from 0.185 for breathiness to 0.879 for intonation (Table 3). For the conversational speech, the raters presented with good or excellent agreement for a greater number of parameters than for the other speech tasks. There were also no consistent results regarding better or worse agreement for assessing the CI group and the NH group. For some parameters, however, by an empirical observation of the results, the raters achieved much better agreement for the CI group than for the NH group. The level of agreement increased expressively for posterior resonance in the sustained vowel (Table 1); for hyponasality, anterior resonance, posterior resonance, strain, roughness, and overall severity in the connected speech (Table 2); and for intelligibility, articulation, intonation, pharyngeal resonance, hyponasality, strain, instability, and overall severity in the conversational speech (Table 3) for the CI group in comparison with the NH group.

Intra-rater reliability

Intra-rater reliability was also obtained by calculating the ICC, by randomly repeating the ratings of 20% of the voice samples. Tables 4–6 show the ICC average measures and the confidence interval range for the three raters in each speech task. Raters 1 and 2, especially rater 2, presented with excellent reliability for most parameters across all of the speech tasks. Rater 3 presented more inconsistencies, ranging from poor to excellent reliability for a greater number of parameters across the speech tasks, being more reliable for the conversational speech.

TABLE 1.
Interclass Correlation Coefficient Across Three Raters for the Sustained Vowel According to the Overall Population, the CI Group, and the NH Group

Sustained Vowel—Inter-rater Agreement						
	Overall		CI Group		NH Group	
	ICC	Confidence Interval	ICC	Confidence Interval	ICC	Confidence Interval
Laryngeal	0.218	–0.089 to 0.469	0.385	–0.040 to 0.681	0.126	–0.584 to 0.569
Pharyngeal	0.278	–0.050 to 0.526	0.377	–0.120 to 0.691	0.138	–0.436 to 0.550
Hypernasality	0.533	0.294 to 0.703	0.522	0.087 to 0.772	0.537	0.150 to 0.773
Anterior	0.235	–0.057 to 0.476	0.182	–0.166 to 0.513	0.305	–0.122 to 0.628
Posterior	0.652	0.456 to 0.781	0.708	0.424 to 0.862	0.485	0.017 to 0.754
Strain	0.265	–0.039 to 0.508	0.234	–0.108 to 0.550	0.039	–0.333 to 0.425
Breathiness	0.588	0.295 to 0.757	0.626	0.270 to 0.823	0.616	0.222 to 0.821
Roughness	0.240	–0.082 to 0.493	0.212	–0.312 to 0.589	0.316	–0.171 to 0.650
Instability	0.505	0.253 to 0.684	0.489	0.093 to 0.751	0.313	–0.332 to 0.674
Pitch	0.792	0.633 to 0.880	0.707	0.367 to 0.868	0.758	0.532 to 0.885
Loudness	0.823	0.687 to 0.898	0.747	0.464 to 0.885	0.866	0.727 to 0.938
Overall severity	0.490	0.222 to 0.677	0.316	–0.226 to 0.660	0.468	–0.490 to 0.751

Abbreviation: ICC, interclass correlation coefficient.

TABLE 2.
Interclass Correlation Coefficient Across Three Raters for the Connected Speech According to the Overall Population, the CI Group, and the NH Group

Connected Speech—Inter-rater Agreement						
	Overall		CI Group		NH Group	
	ICC	Confidence Interval	ICC	Confidence Interval	ICC	Confidence Interval
Laryngeal	0.303	−0.028 to 0.555	0.218	−1.650 to 0.553	0.193	−0.440 to 0.598
Pharyngeal	0.607	0.378 to 0.762	0.470	−0.002 to 0.746	0.451	0.023 to 0.726
Hyponasality	0.845	0.753 to 0.907	0.848	0.705 to 0.928	0.019	−0.931 to 0.539
Hypernasality	0.610	0.383 to 0.764	0.488	0.076 to 0.746	0.411	−0.410 to 0.705
Anterior	0.463	0.162 to 0.671	0.608	0.265 to 0.811	0.291	−0.137 to 0.619
Posterior	0.705	0.530 to 0.823	0.606	0.262 to 0.809	0.406	−0.067 to 0.706
Strain	0.686	0.410 to 0.829	0.701	0.377 to 0.863	0.075	−0.263 to 0.427
Breathiness	0.210	−0.131 to 0.485	0.126	−0.397 to 0.530	0.214	−0.281 to 0.583
Roughness	0.371	0.046 to 0.607	0.453	0.000 to 0.732	0.121	−0.332 to 0.506
Instability	0.437	0.125 to 0.654	0.157	−0.398 to 0.559	0.209	−0.142 to 0.534
Pitch	0.806	0.668 to 0.888	0.798	0.603 to 0.905	0.822	0.636 to 0.917
Loudness	0.868	0.786 to 0.921	0.876	0.754 to 0.941	0.826	0.665 to 0.917
Overall severity	0.852	0.764 to 0.911	0.767	0.476 to 0.897	0.191	−1.050 to 0.496

Abbreviation: ICC, interclass correlation coefficient.

DISCUSSION

Modeling sources of listener variability in voice quality assessment is the first step in developing reliable, valid protocols for measuring quality, and provides insight into the reasons that listeners disagree in their quality assessments.[13] This study presents an experience in rating normal and altered voices using a variety of perceptual parameters directed to a cochlear-implanted population.

It is important to emphasize that this study does not intend to present the voice characteristics of individuals with cochlear implant, but to analyze the rating process of these voices. Usually, the literature discusses the challenges, advantages, and

TABLE 3.
Interclass Correlation Coefficient Across Three Raters for the Conversational Speech According to the Overall Population, the CI Group, and the NH Group

Conversational Speech—Inter-rater Agreement						
	Overall		CI Group		NH Group	
	ICC	Confidence Interval	ICC	Confidence Interval	ICC	Confidence Interval
Intelligibility	0.752	0.587 to 0.855	0.729	0.420 to 0.878	0.471	−0.023 to 0.749
Articulation	0.851	0.752 to 0.913	0.753	0.495 to 0.886	0.437	0.089 to 0.733
Intonation	0.879	0.803 to 0.928	0.892	0.765 to 0.952	0.059	−0.795 to 0.550
Speech rate	0.785	0.658 to 0.871	0.782	0.581 to 0.896	0.791	0.593 to 0.901
Coordination	0.392	0.075 to 0.622	0.124	−0.511 to 0.554	0.077	−0.345 to 0.464
Laryngeal	0.300	−0.024 to 0.551	0.089	−0.242 to 0.434	0.283	−0.252 to 0.598
Pharyngeal	0.464	0.169 to 0.671	0.550	0.156 to 0.783	−0.178	−0.775 to 0.333
Hyponasality	0.790	0.662 to 0.874	0.747	0.511 to 0.880	0.178	−0.555 to 0.605
Hypernasality	0.621	0.396 to 0.771	0.401	−0.460 to 0.697	0.399	−0.046 to 0.696
Anterior	0.286	−0.048 to 0.522	0.403	−0.032 to 0.696	0.120	−0.142 to 0.422
Posterior	0.747	0.597 to 0.848	0.599	0.250 to 0.806	0.473	0.039 to 0.741
Strain	0.781	0.593 to 0.880	0.759	0.539 to 0.885	0.233	−0.109 to 0.549
Breathiness	0.185	−0.088 to 0.434	0.054	−0.327 to 0.430	0.025	−0.154 to 0.281
Roughness	0.246	−0.065 to 0.506	0.205	−0.144 to 0.531	0.141	−0.084 to 0.417
Instability	0.626	0.203 to 0.813	0.544	0.137 to 0.787	0.066	−0.083 to 0.288
Pitch	0.758	0.614 to 0.854	0.764	0.522 to 0.890	0.742	0.506 to 0.877
Loudness	0.402	0.074 to 0633	0.151	−0.619 to 0.594	0.423	−0.029 to 0.712
Overall severity	0.847	0.745 to 0.910	0.700	0.399 to 0.860	0.106	−1.000 to 0.372

Abbreviation: ICC, interclass correlation coefficient.

TABLE 4.
Interclass Correlation Coefficient for the Sustained Vowel for Intra-rater Reliability

Sustained Vowel—Intra-rater Reliability

	Rater 1		Rater 2		Rater 3	
	ICC	Confidence Interval	ICC	Confidence Interval	ICC	Confidence Interval
Laryngeal	0.834	0.348 to 0.958	0.926	0.707 to 0.981	0.323	–1.659 to 0.831
Pharyngeal	0.140	–1.689 to 0.770	0.991	0.967 to 0.998	0.338	–0.520 to 0.798
Hypernasality	0.818	0.262 to 0.995	0.998	0.989 to 0.999	0.803	0.207 to 0.951
Anterior	0.342	–1.825 to 0.839	0.969	0.882 to 0.992	0.371	–1.552 to 0.844
Posterior	0.900	0.617 to 0.975	0.948	0.791 to 0.987	0.287	–2.747 to 0.833
Strain	0.381	–0.086 to 0.832	0.901	0.597 to 0.975	0.388	–2.081 to 0.855
Breathiness	0.820	0.242 to 0.956	0.913	0.618 to 0.979	0.779	0.178 to 0.944
Roughness	0.821	0.342 to 0.954	0.772	0.017 to 0.944	0.651	–0.353 to 0.913
Instability	0.844	0.389 to 0.962	0.927	0.710 to 0.982	0.655	–0.518 to 0.916
Pitch	0.952	0.814 to 0.988	0.985	0.939 to 0.996	0.754	–0.001 to 0.939
Loudness	0.900	0.591 to 0.975	0.977	0.911 to 0.994	0.892	0.553 to 0.973
Overall severity	0.828	0.336 to 0.957	0.982	0.927 to 0.996	0.387	–0.706 to 0.829

Abbreviation: ICC, interclass correlation coefficient.

disadvantages of perceptual ratings in populations with voice deviations of different natures rather than hearing loss, and one of the contributions of this study is to present the rating process that occurred with this specific population, and reflect in aspects that occur with voice analysis of different populations other than this particular one.

Both inter-rater agreement and intra-rater reliability results reflect considerable variability across parameters. Other studies that assessed reliability coefficients also found variable results using the CAPE-V scale, which is also composed of VAS of voice parameters, one performed with adults[11] and another with children.[23] In this study, pitch and loudness were the most consistently rated parameters.

Considering the inter-rater agreement results, it seems that, despite the period of training and calibration, the raters' internal standards prevailed in the auditory-perceptual evaluation. An issue that came up is that the period of 1 month to complete the ratings may have been too long. This ample amount of time was given to prevent fatigue and attention lapses; however, the results suggest that the raters may have fallen out of calibration and their internal standards shifted over that time. Also, providing a written definition of the vocal parameters as a textual anchor alone was not effective in improving agreement.

The ICC for calculating inter-rater agreement was done for the overall population and for the CI group and the NH group. The level of agreement increased significantly for posterior resonance in the sustained vowel (Table 1), and several parameters of the connected speech (Table 2) and conversational speech (Table 3) for the CI group in comparison with the NH group. This finding is probably because these vocal characteristics were

TABLE 5.
Interclass Correlation Coefficient for the Connected Speech for Intra-rater Reliability

Connected Speech—Intra-rater Reliability

	Rater 1		Rater 2		Rater 3	
	ICC	Confidence Interval	ICC	Confidence Interval	ICC	Confidence Interval
Laryngeal	0.546	–1.11 to 0.891	0.744	0.074 to 0.935	0.618	–0.745 to 0.908
Pharyngeal	0.776	0.062 to 0.945	0.362	–0.960 to 0.829	0.605	–0.827 to 0.905
Hyponasality	0.951	0.804 to 0.988	0.831	0.280 to 0.959	0.465	–1.630 to 0.870
Hypernasality	0.812	0.299 to 0.952	0.889	0.568 to 0.972	0.921	0.688 to 0.980
Anterior	0.947	0.800 to 0.987	0.424	–0.851 to 0.848	0.550	–0.889 to 0.890
Posterior	0.945	0.790 to 0.986	0.797	0.225 to 0.949	0.791	0.123 to 0.948
Strain	0.774	0.076 to 0.944	0.342	–2.394 to 0.845	0.123	–2.494 to 0.782
Breathiness	0.643	–0.544 to 0.913	—	—	0.241	–1.717 to 0.806
Roughness	0.319	–2.001 to 0.834	0.990	0.954 to 0.998	0.169	–3.291 to 0.804
Instability	0.933	0.749 to 0.983	0.925	0.714 to 0.981	0.225	–1.110 to 0.782
Pitch	0.950	0.809 to 0.987	0.958	0.833 to 0.989	0.925	0.697 to 0.981
Loudness	0.820	0.251 to 0.956	0.646	–0.181 to 0.907	0.720	–0.220 to 0.932
Overall severity	0.982	0.926 to 0.995	0.959	0.832 to 0.990	0.623	–0.544 to 0.907

Abbreviation: ICC, interclass correlation coefficient.

TABLE 6.
Interclass Correlation Coefficient for the Conversational Speech for Intra-rater Reliability

Conversational Speech—Intra-rater Reliability						
	Rater 1		Rater 2		Rater 3	
	ICC	Confidence Interval	ICC	Confidence Interval	ICC	Confidence Interval
Intelligibility	0.918	0.686 to 0.979	0.992	0.970 to 0.998	0.992	0.970 to 0.998
Articulation	0.750	−0.018 to 0.938	0.980	0.925 to 0.995	0.978	0.915 to 0.994
Intonation	0.860	0.452 to 0.965	0.937	0.741 to 0.984	0.869	0.451 to 0.968
Speech rate	0.872	0.482 to 0.968	0.939	0.758 to 0.985	0.919	0.670 to 0.980
Coordination	0.392	−1.995 to 0.856	0.994	0.975 to 0.998	0.740	−0.064 to 0.936
Laryngeal	0.799	0.230 to 0.949	0.750	−0.240 to 0938	0.343	−1.054 to 0.824
Pharyngeal	0.852	0.449 to 0.963	1,000	0.999 to 1.000	0.736	−0.011 to 0.934
Hyponasality	0.834	0.331 to 0.959	0.689	−0.094 to 0.920	0.278	−1.754 to 0.818
Hypernasality	0.869	0.476 to 0.967	0.999	0.995 to 1.000	0.970	0.865 to 0.993
Anterior	0.521	−1.298 to 0.885	1,000	0.998 to 1.000	0.966	0.805 to 0.992
Posterior	0.894	0.588 to 0.973	0.997	0.998 to 0.999	0.867	0.496 to 0.966
Strain	0.950	0.811 to 0.987	0.991	0.953 to 0.998	0.855	0.419 to 0.964
Breathiness	0.890	0.588 to 0.972	0.972	0.895 to 0.993	0.241	−2.860 to 0.821
Roughness	0.784	0.162 to 0.946	0.954	0.824 to 0.988	0.496	−0.805 to 0.871
Instability	0.846	0.378 to 0.962	0.998	0.991 to 0.999	0.787	0.153 to 0.947
Pitch	0.867	0.503 to 0.966	0.995	0.981 to 0.999	0.965	0.864 to 0.991
Loudness	0.849	0.441 to 0.962	1,000	—	0.867	0.433 to 0.968
Overall severity	0.980	0.925 to 0.995	0.986	0.945 to 0.996	0.870	0.483 to 0.968

Abbreviation: ICC, interclass correlation coefficient.

much more salient for the CI group, contributing to better agreement. Although there are studies that state that normal voices are rated more reliably, Law et al[24] also found better agreement for the disordered population, relating this fact to lack of well-defined differentiation between normal and abnormal voices and different internal standards for pathologic voice qualities, and therefore, raters might not agree on what constitutes a normal voice. They also state that discrete pathologic-type vocal characteristics may be found in the normal population, contributing to poorer reliability because raters disagree on the level of severity of these mild alterations.

Another relevant observation for the inter-rater agreement is that the voice task was also relevant. The number of parameters for which the raters had good or excellent agreement increased expressively from the sustained vowel to the other tasks, which involved speech. With the exception of CAPE-V,[25] which has pre-established phonatory tasks, perceptual-auditory assessment methods do not specify detailed methods of voice sampling and recording, even though these are important considerations because of their potential influence on perceptual ratings. Some studies agree that speech tasks promote better agreement,[14,20,24] whereas another found better agreement results for the sustained /a/ vowel and counting numbers than for conversational speech[26] and another found no difference in agreement among different tasks.[27]

The most commonly used emission in research is the sustained vowel,[21] because it offers information of voice quality at a glottal level, without the interference of aspects of speech. The inclusion of speech tasks, however, is equally important to understand the habitual voice standards in a situation that is similar to the use of the voice in daily situations. Moreover, the overall severity of the voice is not only influenced by the type of voice at a glottal level such as roughness and breathiness, but is highly influenced by suprasegmental information such as intonation and speech rate.[16]

As for intra-rater reliability, rater 2 was the most consistent of all three. Although all three raters had extensive experience in rating voice quality, rater 2 in particular was the only one who had experience in assessing and treating speech and voice disorders in individuals with cochlear implant. Difficulties isolating individual attributes in complex acoustic voice patterns are pointed out to be a relevant issue for rating voice quality,[13] and the voices of this population have peculiarities regarding aspects such as resonance and suprasegmental features.

Studies have shown that listeners often disagree with one another, not as a result of flawed or inconsistent perceptual abilities, but of the methods used to gather ratings.[13] In this study, extensive deadline for completing ratings, lack of periodic recalibration, speech tasks, and familiarity with the population studied were identified as factors that contribute to inconsistent reliability results. Perceptual training should be even more intensified and periodic so that all the parameters to be assessed could be better defined and engraved for its physiological or acoustic aspects, discussed and agreed on by the raters. Some voice characteristics may contain more than one dimension, which the listeners may focus on in different manners,[20] making discussion and modeling of the different sources of reliability important for the perceptual ratings.[13] Moreover, the ability of isolating one parameter from another seems to rely on more explicit definitions, discussion, and training, because a rater may be more sensible to a parameter than another parameter owing to their individual experiences.

Inter-rater agreement depends on whether raters can create similar internal standards for a particular voice signal and thus assign consistent ratings. However, internal standards differ from one another. Therefore, the use of a consistent voice sample alone does not adjust for differences in internal standards between raters, suggesting that the use of training and external anchors is imperative for the creation of similar internal standards.[24]

The use of external anchors seems to contribute in solving many of the problems found regarding reliability in auditory-perceptual evaluation. A study that tested reliability in several different settings reports that listeners agreed 96% of the time in the task using custom comparison stimuli and a conversational scale.[13] Another study that compared the effect of the use of anchors with training alone found better agreement and reliability results using training, auditory anchors, and written anchors combined.[28]

Limitations on the methodology of this study should be noted: After training, the raters were provided with too much time to perform rating with only written external anchor; no recalibration sessions were provided; two of the raters did not have experience with the CI population, and therefore had little familiarity with some of the parameters chosen; and because each rater assessed the voice samples individually, some listening settings such as equipment and listening environment were not controlled. However, as done in this study, it remains important to select different speech tasks, a reliable scale such as the VAS, and parameters that encircle all relevant voice characteristics of the population being studied, and not only glottal source, in this case, adults with cochlear implant.

CONCLUSION

This study presents the reliability results of the rating process of voices of adults with CI and adults with NH. ICC showed inconsistent values, suggesting some lapses in the rating method, such as extensive deadline for the auditory-perceptual evaluation, lack of periodic recalibration, speech tasks, and familiarity with the population studied.

Consistency in methodology and improvements in the auditory-perceptual evaluation of voice quality remain challenging. Describing the severity of auditory-perceptual characteristics of the voice quality, among different populations, in a way that can be communicated among clinicians and research centers, is essential for supporting the perceptual ratings' reliability and validity in the voice assessment.

Future direction includes the combination of factors such as structured training and rating sessions, and use of consistent external anchors, rating scale, and speech tasks, all to achieve the main goal, which is standardized and reliable perceptual assessment of voice quality.

APPENDIX

Definition of vocal features assessed in this study:

- Intelligibility: how understandable the speech is.
- Articulation: correct production of speech sounds.
- Intonation: melody and frequency variation standards in speech.
- Speech rate: fastness or slowness of speech in a sentence.
- Speech-breathing coordination: coordination between breathing and speech.
- Laryngeal resonance: low resonance; voice sounds muffled.
- Pharyngeal resonance: resonance centered in the oropharynx; voice sounds metallic.
- Hyponasality: insufficient use of the nasal cavity during speech.
- Hypernasality: excessive use of the nasal cavity during speech.
- Anterior resonance: excessive oral resonance.
- Posterior resonance: decreased oral resonance.
- Strain: excessive phonatory effort.
- Breathiness: audible air escape in the voice.
- Roughness: irregularity of the voicing source.
- Instability: unstable quality of the emission in terms of frequency and intensity.
- Pitch: perceptual correlation of the fundamental frequency.
- Loudness: perceptual correlation of intensity.
- Overall severity: global, integrated impression of the voice.

REFERENCES

1. Barsties B, De Bodt M. Assessment of voice quality: current state-of-the-art. *Auris Nasus Larynx*. 2015;42:183–188.
2. Master S, Biase N, Pedrosa V, et al. The long-term average spectrum in research and in the clinical practice of speech therapists. *Pro Fono*. 2006;18:111–120.
3. Gama ACC, Alves CFT, da Silva Berto Cerceau J, et al. Correlation between acoustic-perceptual data and voice-related quality of life in elderly women. *Pro Fono*. 2009;21:125–139.
4. Karnell MP, Melton SD, Childes JM, et al. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21:576–590.
5. Yamasaki R, Madazio G, Leão SH, et al. Auditory-perceptual evaluation of normal and dysphonic voices using the voice deviation scale. *J Voice*. 2017;31:67–71.
6. Behlau M, Madazio G, Oliveira G. Functional dysphonia: strategies to improve patient outcomes. *Patient Relat Outcome Meas*. 2015;6:243–253.
7. Iwarsson J, Reinholt Petersen N. Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *J Voice*. 2012;26:304–312.
8. Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatr Logop*. 2009;61:49–56.
9. Eadie TL, Van Boven L, Stubbs K, et al. The effect of musical background on judgments of dysphonia. *J Voice*. 2010;24:93–101.
10. Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *J Voice*. 2004;18:299–304.
11. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20:14–22.
12. Gould J, Waugh J, Carding P, et al. A new voice rating tool for clinical practice. *J Voice*. 2012;26:e163–e170.
13. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am*. 2007;122:2354–2364.
14. Brinca L, Batista AP, Tavares AI, et al. The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *J Voice*. 2015;29:776, e7-e14.
15. dos Santos Freitas SAV. Avaliação Acústica e Áudio Percetiva na Caracterização da Voz Humana [thesis]. Porto, Portugal: Universidade do Porto; 2012. 251 pp [in Portuguese].
16. Eadie TL, Doyle PC. Classification of dysphonic voice: acoustic and auditory-perceptual measures. *J Voice*. 2005;19:1–14.

17. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20:527–544.
18. Gerratt BR, Kreiman J, Antonanzas-Barroso N, et al. Comparing internal and external standards in voice quality judgments. *J Speech Hear Res*. 1993;36:14–20.
19. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *J Acoust Soc Am*. 2000;108:1867–1876.
20. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19:555–573.
21. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.
22. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychol Assess*. 1994;6:284–290.
23. Kelchner LN, Brehm SB, Weinrich B, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. *J Voice*. 2010;24:441–449.
24. Law T, Kim JH, Lee KY, et al. Comparison of rater's reliability on perceptual evaluation of different types of voice sample. *J Voice*. 2012;26:666, e13-e21.
25. Kempster GB, Gerratt BR, Verdolini Abbott K, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
26. Lu FL, Matteson S. Speech tasks and interrater reliability in perceptual voice evaluation. *J Voice*. 2014;28:725–732.
27. Maryn Y, Roy N. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *J Soc Bras Fonoaudiol*. 2012;24:107–112.
28. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice*. 2009;23:341–352.