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