Speech Tasks and Interrater Reliability in Perceptual Voice Evaluation

*Fang-Ling Lu and †Samuel Matteson, *†*Denton, Texas*

Summary: Objective/Hypothesis. The optimal selection of speech task is essential for more reliable perceptual ratings and a better understanding of the perceptual qualities of pathologic voices. Nevertheless, researchers have rarely explored this issue using the GRBAS scale. This study investigates the effect of speech task selection on interrater reliability during perceptual voice assessment.

Study Design. Experimental study.

Methods. Sixty subjects, 39 dysphonic subjects and 21 normal controls, performed 13 speech tasks including three 5-second sustained vowel sounds (/ɑ/, /i/, and /u/) each at three pitch levels (high, habitual, and low), maximum phonation of the vowel /ɑ/, pitch glide, counting from 1 to 10, and oral reading of the Rainbow Passage. A group of 18 graduate students in speech-language pathology served as perceptual judges and rated the dysphonic severity for the speech samples based on three parameters in the GRBAS scale—Grade, Roughness, and Breathiness. The formalism of the AC_1 statistic proposed by Gwet was applied to determine relative reliability between the speech tasks and the raters.

Results. The counting task and sustained vowel /ɑ/ in high, habitual, and low registers exhibited the most reproducibility and consequently the highest reliability statistic.

Conclusions. The counting task and sustained /ɑ/ phonation are the optimal tasks for perceptual voice judgment in regard to interrater reliability. Future perceptional studies may benefit from this finding to determine the relationship between speech task selection and the validity of any given perceptual rating system in terms of sensitivity and specificity.

Key Words: Sustained vowel phonation–Contextual speech–Interrater reliability–GRBAS scale.

INTRODUCTION

Auditory perceptual assessment plays a vital role in voice evaluation despite its inherent subjectivity and the lingering debate regarding its reliability and validity.[1–7] The most evident advantages of using the perceptual voice assessment are accessibility of the test materials and simplicity in implementation procedures. Several popular perceptual evaluation scales such as the GRBAS scale[8] or the CAPE-V system[9,10] have been well studied and are readily accessible to clinicians. Although auditory perceptual measures are often used as a reference for other objective voice assessment tools such as acoustic analysis, the link between objective measures and perceptual assessment of dysphonic voices remains disappointingly weak and inconclusive due to intrinsic shortcomings of each measurement system.

Auditory perceptual assessment itself is an intricate process involving numerous complex interrelated elements, many of which are subjective by nature and not well understood. Research studies suggest that judgments of vocal qualities are inherently unstable and prone to measurement error caused by many known and unknown variables. A number of factors such as the listener's professional training in voice disorders, the listener's bias derived from a prior knowledge of the speaker's medical or voice history, voice features in a perceptual rating scale, or the type of speech stimuli to be judged are known to have significant impact on the listener's ability to differentiate between pathologic and nonpathologic voices.[1–7,11–18]

Researchers have recommended three evidence-based approaches to mitigate rater-related variability, including provision of listening training, limited choice of voice features for evaluation, and selection of suitable speech stimuli.[5,19,20] Although the former two approaches have been extensively studied, the issue concerning appropriate speech task selection has not received significant attention.

Studies have shown that inter- and intrarater reliability may improve when the listeners receive listening training before testing.[11,13,21–23] The level of agreement among listeners may also improve if only three voice features—overall severity, roughness, and breathiness are judged.[7,9,11,12,14,15,20,23,24] When selecting speech stimuli for perceptual evaluation, the general recommendation is to include both sustained vowels and contextual speech in testing, given the inherent features unique in both types of speech samples.[25–28] Although sustained vowel stimuli naturally produce a higher level of interrater reliability due to their innate stability and consistency, they are poor representatives of daily voice usage and are prone to underestimation of the severity of voice deviance.[1,16,21,29–31] On the other hand, although inherent physiological complexity and naturalness in the contextual speech stimuli provide a more accurate estimation of deviant voice quality,[16,21,29,30,32] this type of speech sample is inclined to produce a lower rater reliability because of their intrinsic variability in speaking pattern (eg, dialect, speaking rate, prosody) and voice quality.[30] To date, the selection of speech tasks for perceptual voice assessment is generally

Accepted for publication January 31, 2014.

Portions of this article were presented at the 2012 National Convention of American Speech-Language-Hearing Association; Atlanta, Georgia, November 16, 2012.

From the *Department of Speech and Hearing Sciences, University of North Texas, Denton, Texas; and the †Department of Physics, University of North Texas, Denton, Texas.

Address correspondence and reprint requests to Fang-Ling Lu, Department of Speech and Hearing Sciences, University of North Texas, 1155 Union Circle # 305010, Denton, TX 76203-5017. E-mail: flu@unt.edu

Journal of Voice, Vol. 28, No. 6, pp. 725-732
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.01.018

limited to three speech tasks—sustained phonation of the vowel sounds /ɑ/ and /i/, as well as a short reading of phonetically balanced texts such as the Rainbow Passage. In the field of voice research, there seemed to be a lack of attention heretofore, in general, that prompts an exploration of the utility of other speech stimuli for perceptual voice evaluation.

The research evidence from objective measurement studies using acoustical analyses, electroglottographic measures, or strobolaryngoscopic examinations has shown significant relationships between speech tasks and the structural configuration and positioning of the larynx and vibratory pattern of the vocal folds.[33–37] For examples, the vowel type,[33,34,38–46] pitch level,[41,47,48] vocal effort,[39,47,49,50] vowel- versus text-based context,[27,33–36,51–53] utterance length,[40,54–56] and speaking rate[57,58] are shown to influence vocal tract configuration, voice onsets, or pauses between syllables or words. Based on these findings and the evidence from foregoing perceptual voice research, it may be safe to postulate that speech features (eg, sustained phonation, pitch, loudness, articulation, speaking rate, and so forth) play a crucial role in influencing a speaker's voice quality and may also have a significant effect on the listener's ability to reliably and accurately judge the speaker's voice quality.[33–37]

The present study was designed to determine the effect of speech task selection on interrater reliability in perceptual voice judgment. The relationship between interrater reliability and the selection among 13 speech tasks was investigated. The authors hypothesized that the level of agreement between the listeners in judging dysphonic and nondysphonic voices could be influenced by the selection of speech samples carrying certain salient features (eg, vowels, contextual speech, various pitch levels, utterance lengths, and so forth), and such findings could provide a starting point to find optimal speech stimuli that could provide a respectable level of rater reliability and adequate identification of dysphonia.

MATERIALS AND METHODS

Subjects and voice samples

Voice samples were obtained from 39 dysphonic subjects and 21 normal controls. The dysphonic group consisted of 15 males and 24 females with an age range of 18–81 years; the average age was 32.9 ± 2.7 years. Subjects of the dysphonic group were diagnosed with a wide range of laryngeal pathologies, which were verified through the strobovideolaryngoscopic examinations performed by otolaryngologists or speech-language pathologists. Diagnoses of vocal fold pathology among dysphonic subjects included 13 cases of laryngitis, 13 cases of vocal nodules, five cases of vocal polyp or polyps, three cases of presbylaryngis, two cases of muscle tension dysphonia, one case of spasmodic dysphonia, and two cases of unilateral vocal fold paralysis. The control group consisted of one male and 20 females ranging in age from 21 to 35 years (mean = 25.5 years; standard deviation = 0.9). All control subjects exhibited normal laryngoscopic findings and reported no current history of dysphonia. Informed consent was obtained from all the subjects in the study, and the protocol was approved by the University of North Texas Institutional Review Board.

Speech samples were recorded in a quiet room. Each subject was fitted with a headset microphone (TalkPro Xpress Headset; VXI Corp., Rollinford, NH) coupled with a high-quality digital audio recorder (Olympus LS-10 Linear PCM recorder; Olympus Imaging America Inc., Cypress, CA). The microphone was maintained at a distance of 7.5 cm from the subject's mouth and slightly off center to avoid breath or plosive noise. The recording volume was monitored continually to maintain an optimal dynamic range in the recording system to avoid distortion. Recorded speech samples were saved as .wav files at 48 000 Hz sampling rate with 16 bits of amplitude resolution.

During each recording session, the subject performed 13 speech tasks as follows: (Tasks 1–9) 5-second sustained phonation of three vowel sounds (/ɑ/, /i/, and /u/), each at three pitch levels (high, habitual, and low); (Task 10) maximum prolongation of the vowel /ɑ/ in one breath; (Task 11) a pitch glide saying the "ah" sound; (Task 12) counting from 1 to 10; and (Task 13) oral reading of the first paragraph of the Rainbow Passage. Before recording, each subject received instructions to perform habitual-pitched phonation of three vowels, maximum phonation of /ɑ/, counting, and passage reading at a comfortable pitch and loudness level. Subjects were also instructed to maintain a natural pace in counting (10 words) and reading of the Rainbow Passage (100 words). During high- and low-pitched vowel productions, subjects were shown by the investigator (the first author) to sustain the vowel sounds without reaching the range of falsetto (ie, loft register) or glottal fry (ie, pulse register). For the pitch glide task, subjects were instructed to begin the /ɑ/ vowel at habitual pitch level, followed by an ascending glide to reach the highest pitch level without causing any pitch breaks and then a descending glide to reach the lowest pitch level without producing any glottal fry. The investigator closely monitored each subject's voice volume to avoid unintended loudness shifts during high- or low-pitched phonation. Subjects were asked to produce each task twice consecutively. In all, each subject produced 26 speech samples (13 tasks × 2 trials).

Each sustained vowel sample was approximately 5 seconds in duration, whereas other speech samples had a wide range of lengths from 60 subjects. The average length for the maximum phonation task was between 5.2 and 29.3 seconds; for the pitch glide task, between 1.2 and 15.3 seconds; for the counting task, between 4.2 and 10.4 seconds, resulting in an average speaking rate of 59–143 words per minute (WPM); and for the Rainbow Passage reading, between 22.4 and 57.1 seconds, resulting in an average speaking rate of 105–268 WPM. Altogether, 120 voice samples were generated by 60 subjects in each speech task, resulting in a total of 1560 voice samples from all 13 tasks (13 tasks × 2 trials × 60 subjects). The recorded samples were saved in a computer, and speech samples in each task set were organized in a random order to mitigate potential learning effects on listeners during testing.

Judges and listening training

Eighteen 2nd-year graduate students in speech-language pathology were recruited to be judges without any monetary or