Effect of Auditory vs. Textual Anchors on Vocal Severity Ratings

(Accepted for publication in the *Journal of Voice*)

Shaheen N. Awan, Ph.D.
Bloomsburg University of Pennsylvania

Laura Lawson, M.S. CCC-SLP
Berks County IU, Reading PA

Abstract

This study investigated anchor type (textual, auditory, or a combination) and effect on rater reliability. Forty judges were selected to rate voice samples of various voice quality types in terms of vocal severity using four different methods (No Anchor, Textual Anchor, Auditory Anchor, and Combined Textual/Auditory Anchors). Results indicated that auditory anchors produced significantly improved interjudge reliability across all voices as well as improved reliability for rating mild voice disorders. Textual anchors showed improvement over the no anchors condition, but were not as strong as the use of auditory anchors. However, combined textual/auditory anchors generally showed the greatest degree of reliability.

Summary

Attempts have been made to improve the validity and reliability of the perceptual evaluation of the voice by conducting ratings in conjunction with the use of external standards that can act as ‘anchors’ or references to which listeners can compare the voice sample being judged. Despite previous research suggesting the beneficial use of anchors, there have been no studies that address the anchor modality (i.e. auditory vs. visual) which would be most likely to improve reliability and stabilize one’s internal standard. It would be important to determine if a reliable and valid method for conducting perceptual analyses can be achieved to reduce the variability within and between judges. The goals of this study were to determine if anchors and training affect the perceptual rating of various voice types and seventies, compared to receiving training alone; if so, does the modality in which the anchor is presented affect rater reliability. The anchor type or modality refers to whether the anchor is presented auditorily, visually via a written definition (referred to as a ‘textual’ anchor) or a combination of both anchor types. If textual anchors and auditory anchors are equally reliable, then the requirement of creating and implementing a standard set of auditory anchors may be unnecessary.

Forty inexperienced judges were selected to rate 36 voice samples of various voice types (i.e. normal, breathy, hoarse, rough) in terms of perceived vocal severity using four different methods (No Anchor, Textual Anchor, Auditory Anchor, Combined Textual/Auditory Anchors). The selected voices samples had been previously rated in Awan and Roy (2005; 2006) for (a) voice quality type (normal; breathy; hoarse; rough) and (b) severity (rated on a hybrid categorical/visual analog scale) by 12 judges. All of the judges in the Awan and Roy studies had (a) completed a graduate course in voice disorders, (b) been exposed to the terminology to be used in the rating task, (c) participated in classroom exercises
in the perceptual evaluation of voice, and (d) had clinical experience with voice disordered patients. The use of multiple judges was employed to reduce the potential for any interjudge differences to create spurious experimental conclusions. The 36 voice samples selected included 9 normals, 9 breathy, 9 hoarse, and 9 rough samples (Total N = 36). Within each disordered voice type category, mild, moderate, and severe dysphonias were represented (3 of each severity category). These particular samples were chosen due to strong interjudge agreement in ratings of voice type and severity in the Awan and Roy studies.

For this particular study, forty (40) judges were selected from the undergraduate and graduate student population majoring in Speech Pathology to rate the aforementioned 36 voice samples in terms of voice type and severity. The judges included in this study may have had some limited exposure to speech and voice disorders in their previous academic coursework but should not have yet had (a) a formal course in voice disorders and voice evaluation, and (b) should not have been familiar with definitions for voice type and severity found in Awan (2001). Participants were 36 females and 4 males ranging from 19 to 44 years of age. A Kruskal-Wallis Nonparametric ANOVA indicated that there were no significant differences in the mean ages of each of the groups. All judges signed an Informed Consent Form and successfully passed a hearing screening of 25 dB at .5, 1, 2, and 4 kHz. The judges were randomly assigned to one of the aforementioned four groups.

A computer program was designed by for use by the judges when rating the voice samples. It was composed of four different screens, one for each of the four possible groups (i.e. No Anchor, Textual Anchor, Auditory Anchor, or Combined Textual/Auditory). Each screen included standard radio buttons to select quality (i.e. normal, breathy, hoarse, or rough) and a sliding scale (100 mm VAS) for severity that corresponds to the Overall Severity scale in the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) assessment form. The appearance of the program screen varied depending upon which group a judge was assigned to. The No Anchor group had only the standard buttons previously mentioned. For the anchor groups, there were additional buttons on the screen that represented each of the three quality types and the seventy types of mild and severe. In the Textual Anchor group, clicking on the anchor button would provide a written definition in a text box of the corresponding quality and severity. In the Auditory Anchor group, clicking on each anchor button would result in an auditory sample of the corresponding severity and quality. Clicking on the anchor buttons in the Combined Anchor group would result in both an auditory sample and a written definition of the anchor selected. The computer program also wrote all rating results to a .txt file for later analysis.

Results indicated that the combination of training and anchors significantly improves the reliability (in terms of both interjudge correlations and agreement, and 95% confidence intervals) of perceptual judgments. In particular, those subject groups that utilized auditory anchors within the rating task showed significantly improved interjudge reliability across all voices, as well improved reliability for rating mild voice disorders. Auditory anchors also resulted in reduced variability of ratings for both breathy and hoarse voice qualities. Textual anchors did show some improvement over training alone (i.e., no anchors), but were generally not as strong as the use of auditory anchors. However, the combination of textual and auditory anchors generally showed the greatest overall degree of reliability. Anchors were not only useful in improving measures of reliability amongst judges, but were able to do so without significantly increasing the amount of time spent on the rating tasks. Results and implications are discussed.

For more information, please contact Dr. S. N. Awan at sawan@bloomu.edu