Use of Natural Anchors for Dysarthria Assessment: An Exploratory Study on Improving Rater Reliability

Thushani Umesha Munasinghe[1], Deepthi Crasta[1], Kaila Stipancic[2], Mili Kuruvilla-Dugdale[1]
[1]*University of Iowa, Iowa, USA*
[2]*University at Buffalo, USA*
thushani-munasinghe@uiowa.edu, deepthi-crasta@uiowa.edu, klstip@buffalo.edu, mili-kuruvilla-dugdale@uiowa.edu,

Abstract

The aim of this project was to determine whether the use of anchors improves interrater and intrarater reliability when nonexpert listeners rated five features salient to hypokinetic dysarthria: overall severity, reduced loudness, articulatory imprecision, short rushes of speech, and monotony. Fourteen nonexperts rated 82 sentences recorded from individuals with Parkinson's disease and healthy controls using five separate equal appearing interval (EAI) scales to indicate their perception of the five features mentioned above. The listeners rated the samples twice, once without and once with external anchors. Interrater reliability and intrarater reliability were calculated using intraclass correlation coefficients (ICCs). Findings revealed an overall increase in both interrater and intrarater reliability for most features in the anchor condition, except for monotony, where a decrease in single-measures ICC was noted for the anchor compared to non-anchor condition. These preliminary findings highlight how external anchors can benefit interrater and interrater reliability when rating perceptual dimensions of dysarthria.

Keywords: scaling, dysarthria, reliability, anchors

1. Introduction

When listeners rate different speech samples, they develop internal standards for what counts as different intervals on a rating scale, and they rely on these standards to guide their judgements (Kreiman et al., 1993). Internal standards can be influenced by factors such as experience and training of the listener, or context, as well as other listener characteristics such as memory and attention (Gerratt et al., 1993; Kreiman et al., 1992). Internal standards are believed to be developed gradually over years, for example, when expert clinicians are exposed to disordered speech samples over time. Internal standards are known to drift and may be unstable while getting established, resulting in variable ratings. In addition to these listener-related factors, external factors such as acoustic context (e.g., a listener may perceive a speech sample with moderate severity as more severe if it is presented after a series of voices with mild dysarthria severity) and task (e.g., reading task vs. spontaneous conversation) can also affect internal standards.

The reliance on unstable internal standards for perceptual judgments could be one reason for the high variability in rater reliabilities reported in the literature. In the context of dysarthria, reduced rater reliability poses a significant challenge when employing auditory-perceptual assessments to diagnose and measure specific subsystem features (Stipancic et al., 2023). To improve rater reliability, voice researchers have used stable external standards with the intent of replacing the idiosyncratic internal standards (Awan & Lawson, 2009; Chan & Yiu, 2002). In these studies, the external anchor served as a reference against which raters could compare the experimental stimuli. Both natural (i.e., speech samples from speakers) and synthetic (i.e., computer-generated speech) anchors have been studied. Natural anchors seem to be the best references to use if they resemble the target stimuli to be rated by listeners.

Several studies have been conducted to examine the improvements to interrater and intrarater reliability when employing anchors in voice assessment, as well as the type of anchor (natural versus synthetic), and the mode of presentation (auditory, visual/textual, or a combination of both) that affects reliability most (Awan & Lawson, 2009; Santos et al., 2021). A previous study also explored how listener experience influenced the use of anchors (Eadie & Kapsner-Smith, 2011). The findings suggest that external auditory anchors enhance both interrater and intrarater reliability compared to other modalities. Experience level did not influence reliability, and both experienced and novice listeners demonstrated greater reliability when using anchors compared to the condition without anchors. Novice listeners often play a role in speech assessment, to determine the real-world impact of the communication disorder, and enables the recruitment of a large participant pool for research studies. However, no previous studies have been conducted to examine if and how rater reliability changes with external anchor use when nonexperts judge dysarthric speech features.

In the context of dysarthria assessment, interval scales are often used because they are less time consuming and easy to use in a clinical setting (Kreiman et al., 1993). The equal appearing interval scale (EAI) is one such scale, which has predefined intervals that are equidistant from each other. In an anchored condition, the experimenter provides a reference stimulus for each interval and raters can use the references to guide their ratings of the experimental stimuli. Although previous voice studies have explored the reliability of employing anchors with the EAI scale (Gerratt et al., 1993), similar investigations have yet to be conducted for dysarthria.

The aim of the present study was to compare the reliability of ratings completed with an EAI scale by nonexpert listeners, without and with the presence of anchors. Both interrater and intrarater reliability were examined for five salient hypokinetic dysarthria features, including overall speech impairment severity, reduced loudness, articulatory imprecision, short rushes of speech, and monotony.

2. Methods

This study was approved by the Institutional Review Board of the University of Iowa. All participants gave written informed consent before completing study procedures.

2.1 Participants

Two groups of participants were included: 1) speakers; and 2) listeners.

10.21437/issp.2024-7

2.1.1 Speakers

The speakers consisted of 43 individuals with Parkinson's disease (PD; 18 females, 25 males) and 25 neurologically healthy speakers (11 females, 14 males). The inclusion criteria were: (i) no history of speech, language, or hearing disorders; (ii) no co-occurring neurological diagnosis for the participants with PD, and absence of any neurological diagnosis for the controls; (iii) no history of head and neck surgery; (iv) not wearing a hearing aid or having a prescription for hearing aids; and (v) be a monolingual, native speaker of American English.

2.1.2 Listeners

Fourteen neurologically healthy participants (M_{age} = 26.5 years, SD = 3.55) were recruited as listeners. The inclusion criteria were (i) be between the ages 19-90 years; (ii) pass a bilateral hearing screening at 25 dB HL at 500 Hz, 1 kHz, 2 kHz, and 4 kHz; (iii) no history of speech, language, or hearing disorders; (iv) use English as the primary language of communication; (v) have minimal exposure to communication disorders.

2.2 Experimental Tasks

2.2.1 Speech task

The speakers were recorded reading 11 unique sentences from The Speech Intelligibility Test (SIT; Yorkston et al., 2007); one sentence that presented with the greatest number of dysarthria features of interest was selected from each speaker. To determine which features were present, each sentence was rated on the Dysarthria Rating Scale (Darley et al., 1969 a, 1969 b), independently by two trained research assistants. Consensus was sought if they disagreed about features, and the consensus ratings were used to select the final stimulus set.

2.2.2 Auditory-perceptual scaling task

A total of 68 samples (i.e., 43 PD and 25 control) were used to determine interrater reliability; 20% of the samples (n = 14) were randomly selected and repeated for intrarater reliability.

The perceptual experiment was conducted in a quiet laboratory setting. Each listener attended two sessions which were one week apart and lasted approximately one hour each week. Listeners used calibrated headphones to listen to the speech samples. Ratings were completed using a custom MATLAB GUI that displayed five separate EAI scales at once, one for each feature (i.e., overall speech impairment severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony).

Definitions of each feature were provided by the experimenter to the listeners. They were instructed to rate overall severity based on their general impression of severity rather than understandability of the sentences. Reduced loudness was assessed based on the softness or quietness of the voice in the sample, while articulatory imprecision was evaluated based on how crisply and clearly the speech sounds were produced. Short rushes of speech was rated by identifying instances of rapid speech characterized as rushed segments preceded and followed by pauses. For monotony, the listeners were instructed to consider the flatness of the speech sample in terms of pitch, loudness, or duration.

Listeners used a 5-point EAI scale either without anchors or with anchors. The 5-point EAI scale had the following intervals: 1=typical, 2=mild, 3=moderate, 4=severe, and 5=profound. For the session without anchors, listeners were asked to rate the first three features after listening to a sample once and then rate the next two features after listening to the sample again. For the anchor condition, reference samples were provided for each scale interval for each feature. The listeners played the anchors of the first three features before listening to the sample and rating the three features. They followed the same steps for the next two features. The anchors reappeared after every eight samples. The listeners were encouraged to use the entire scale for the ratings during both the sessions.

The anchor for each scale interval was selected by two experts. First, an experienced speech-language pathologist rated dysarthria speech samples from the *Audio Seminar Series* (Darley et al., 1975) and chose samples for mild, moderate, severe, and profound levels for each feature. Then the last author rated the chosen samples independently for features and severity levels. Discrepancies between the experts were resolved through consensus before the anchors for each interval of each feature were selected.

2.3 Data Analysis

2.3.1 Statistical analysis

SPSS statistical software version 28 (SPSS Inc, Chicago, IL) was used for statistical analyses. Both interrater reliability and intrarater reliability were estimated using intraclass correlation coefficients (ICCs). For interrater reliability, single- and average-measures consistency from 2-way random-effects models (Koo & Li, 2016) with 14 raters across 68 samples was used to obtain the ICCs and their respective 95% confidence intervals. For intrarater reliability, single- and average-measures consistency from 2-way mixed-effects models (Koo & Li, 2016) for the 14 raters were calculated along with their relevant 95% confidence intervals.

Inter- and intra-rater reliability ICCs were calculated for each of the five speech features for both anchor conditions (i.e., with and without anchors). ICC values were descriptively compared between the anchor conditions for each feature. A difference in ICC values between anchor conditions was considered meaningful if there was a switch to a higher or lower reliability category with the use of anchors.

3. Results

3.1 Interrater reliability

Compared to the non-anchor condition, there was an overall increase in both single- and average-measures ICC for all features, including overall speech impairment severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony for the anchor condition (Table 1). The average-measures ICCs indicated good or excellent reliability regardless of the anchor condition. However, for short rushes of speech, a change in the reliability category was observed, where reliability increased from moderate to excellent with anchors. Single-measure ICC values for all features ranged from poor to moderate levels regardless of the anchor condition. However,

reliability for overall severity and reduced loudness improved from poor to moderate when anchors were used.

Table 1: *Interrater reliability of all features rated with an equal appearing interval (EAI) scale with and without anchors.*

Speech Feature	Interrater Reliability Intraclass Correlation Coefficient (CIs)		
	Measure	No Anchors	Anchors
Overall severity	Single	**0.492** (.403-.593)	**0.587** (.501-.680)
	Average	0.931 (.904-.953)	0.952 (.934-.967)
Articulatory imprecision	Single	0.513 (.425-.613)	0.597 (.512-.689)
	Average	0.937 (.912-.957)	0.954 (.936-.969)
Reduced loudness	Single	**0.423** (.337-.526)	**0.520** (.432-.619)
	Average	0.911 (.877-.940)	0.938 (.914-.958)
Short rushes of speech	Single	0.295 (.219-.393)	0.410 (.324-.513)
	Average	**0.854** (.797-.901)	**0.907** (.870-.936)
Monotony	Single	0.433 (.346-.536)	0.495 (.407-.596)
	Average	0.914 (.881-.942)	0.932 (.906-.954)

Note. CI=Confidence interval; ICC values less than 0.5 are indicative of poor reliability; values between 0.5 and 0.75 indicate moderate reliability; values between 0.75 and 0.9 indicate good reliability; and values greater than 0.90 indicate excellent reliability. The bold values indicate the category shift in the ICC values.

3.2 Intrarater reliability

The single- and average-measures ICC values for intrarater reliability ranged from moderate to excellent in both anchor conditions (Table 2). Except for overall severity and monotony, reliability measures increased for both single and average measures when anchors were used. The single measure ICC for articulatory imprecision moved from moderate to good reliability, but there was no shift in reliability categories for any of the other features.

Although several tests are available to determine statistical differences between ICC measures (e.g., Fisher's Z test, Konishi-Gupta modified Z-test, the likelihood ratio test, and Alsawalmeh-Feldt F-test), for the present exploratory study, we compared the ICC measures in a more qualitative manner, as the study was underpowered (Donner et al., 2002).

4. Discussion and Conclusion

In this study, we investigated interrater reliability and intrarater reliability among raters who assessed speech samples from talkers with PD and healthy controls using an EAI scale. The preliminary findings presented here suggest that there are benefits to combining natural anchors with an interval scale for evaluating dysarthric speech.

A meaningful change in interrater reliability (i.e., switch to a higher reliability category) was observed for three out of the five speech features, namely overall severity (single-measures ICC), reduced loudness (single-measures ICC), and short rushes of speech (average-measures ICC). Despite this improvement, the moderate reliability observed for overall severity and reduced loudness when using anchors is insufficient for clinical purposes, which contrasts with the average measures for both anchor conditions, which are highly acceptable. Voice studies have reported similar magnitudes of improvements in reliability when using external anchors combined with training. However, these studies show increased interrater variability when anchors were used without training, suggesting limited use for anchors alone (Chan & Yiu, 2006). Most prior studies only include average-measures ICC, presumably because the individual rating is unreliable, and ICCs based on average measures are always higher than those based on single measures. Hayen and colleagues (2007) emphasize that average measures should not be used when determining ICCs unless there are specific situations where averaged ratings apply. In dysarthria assessment, the measurement from a single rater is typically the basis of the actual measurement, suggesting the importance of considering single measures ICCs.

Table 2: *Intrarater reliability of all features rated with an equal appearing interval (EAI) scale with and without anchors.*

Speech Feature	Intrarater Reliability Intraclass Correlation Coefficient (CIs)		
	Measure	No Anchors	Anchors
Overall severity	Single	0.824 (.773-.864)	0.824 (.773-.864)
	Average	0.903 (.872-.927)	0.903 (.872-.927)
Articulatory imprecision	Single	**0.734** (.663-.793)	**0.808** (.753-.852)
	Average	0.847 (.797-.884)	0.894 (.859-.920)
Reduced loudness	Single	0.761 (.695-.814)	0.813 (.759-.855)
	Average	0.864 (.820-.898)	0.897 (.863-.922)
Short rushes of speech	Single	0.650 (.561-.724)	0.719 (.643-.780)
	Average	0.788 (.719-.840)	0.836 (.783-.877)
Monotony	Single	0.734 (.662-.792)	0.668 (.583-.739)
	Average	0.847 (.797-.884)	0.801 (.736-.850)

The switch to a higher intrarater reliability category was observed only for single-measures ICC of articulatory imprecision. However, for most features, the magnitudes of both single and average measures increased with the use of anchors, except for monotony and overall severity. Similar improvements in intrarater reliability in the presence of auditory anchors have been observed in voice studies (Chan & Yiu, 2002; Eadie & Kapsner-Smith, 2011). Regarding monotony, previous dysarthria studies indicated that this feature behaves differently from other hypokinetic dysarthria features. Stipancic et al. (2023) showed that ratings of monotony had the poorest criterion validity and reliability compared to ratings of overall speech impairment, articulatory imprecision, and slow rate.

Another study by Stipancic (in press) identified monotony as a metathetic feature compared to the four other features in this study, which were identified as prothetic continua. Therefore, future work is necessary to identify the perceptual properties of monotony to delineate why it behaves differently, and to incorporate the findings for future research (e.g., have listeners rate the subordinate dimensions of monotone speech rather than overall monotony). When comparing single-measures ICC for interrater reliability and intrarater reliability, it was evident that the single-measures ICC for intrarater reliability were higher than for interrater reliability across anchor conditions and features. This indicates the raters are more consistent within themselves. In contrast, the average-measures ICC of intrarater reliability were lower than the interrater ICC values for both conditions. This contrasts with the findings of previous voice studies, which showed that intrarater reliability values are generally better than interrater reliability measures with the use of anchors (Awan & Lawson, 2009). The observed differences may stem from methodological variations across studies, including differences in subject populations, task complexities, and stimuli. Follow-up studies are warranted to investigate deeper into potential reasons underlying these disparities.

There may be potential reasons for the variability in interrater and intrarater reliability across different speech features. For the present study, we used the EAI scale to rate hypokinetic dysarthria features since it is one of the most used scales in research and clinical settings. However, an EAI scale might not be suited for assessing speech features that are prothetic because they are best scaled by direct magnitude estimation (DME), while metathetic features can be scaled using EAI or DME. Results of a recent study indicated that except for monotony, all the other features in the current study were prothetic, suggesting that a DME scale, rather than an EAI scale is the best fit to assess overall speech impairment severity, articulatory imprecision, reduced loudness, and short rushes of speech (Stipancic, in press). When selecting a scale for dysarthria assessment, it is essential to consider both reliability and validity. Even though the EAI scale shows increased reliability in rating dysarthria features with the use of anchors, it is also essential to consider if the EAI scale is the best fit for each feature. Critical next steps will be to examine both EAI and DME scales without and with anchors to see if reliability changes similarly across the different scales.

It is also important to consider when anchors should be used. A study by Stipancic et al. (2023) investigated the effect of auditory training on perceptual ratings of dysarthric speech, in which external anchors were only used during training and not during the post-training ratings. Results showed that there was little improvement in rater reliability as a result of training. One of the reasons might be that in this previous study, external anchors were only used during training, and the internal standards of the nonexpert listeners may have been unstable and insufficient on their own to improve reliability. Therefore, it is recommended to use anchors during the actual ratings to avoid overreliance on shifting or developing internal standards, particularly with nonexpert listeners.

It is important to systematically investigate the use of anchors for rating salient speech features of other dysarthria types. In the current study, an overall improvement in interrater and intrarater reliability was observed with the addition of external anchors to an EAI scale. The feasibility of incorporating anchors for other types of scales, such as DME and visual analog scales will be investigated in future work.

5. Acknowledgements

This research was supported by the National Institutes of Health (R15DC016383 and R21DC019952; PI: Kuruvilla-Dugdale). We are grateful to the research assistants and subjects who participated in the study. Special thanks to Dahlia Cukierkorn, Lexi Jacobsmeyer, Morgan Linneweh, Ella Meier, and Anna Mae Williams for helping with data collection and analysis. Chaewon Park and Minguang Song helped modify the original MATLAB GUI for the anchor condition.

6. References

Awan, S. N., & Lawson, L. L. (2009). The Effect of Anchor Modality on the Reliability of Vocal Severity Ratings. *Journal of Voice*, *23*(3), 341–352.

Chan, K. M. K., & Yiu, E. M.-L. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research: JSLHR*, *45*(1), 111–126.

Eadie, T. L., & Kapsner-Smith, M. (2011). The Effect of Listener Experience and Anchors on Judgments of Dysphonia. *Journal of Speech, Language & Hearing Research*, *54*(2), 430–447.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing Internal and External Standards in Voice Quality Judgments. *Journal of Speech, Language, and Hearing Research*, *36*(1), 14–20.

Hayen, A., Dennis, R. J., & Finch, C. F. (2007). Determining the intra- and inter-observer reliability of screening tools used in sports injury research. *Journal of Science and Medicine in Sport*, *10*(4), 201–210.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, *36*(1), 21–40.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research*, *35*(3), 512–520.

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, *17*(1), 45–56.

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40*(3), 699–704.

Santos, P. C. M. D., Vieira, M. N., Sansão, J. P. H., & Gama, A. C. C. (2021). Effect of synthesized voice anchors on auditory-perceptual voice evaluation. *CoDAS*, *33*(1), e20190197.

Stipancic, K. L., Golzy, M., Zhao, Y., Pinkerton, L., Rohl, A., & Kuruvilla-Dugdale, M. (2023). Improving Perceptual Speech Ratings: The Effects of Auditory Training on Judgments of Dysarthric Speech. *Journal of Speech, Language, and Hearing Research*, 1–23.

Stipancic, K. L., Whelan, B., Laur, L., Zhao, Y., Rohl, A., & Kuruvilla-Dugdale, M. (in press). Tipping the Scales: Indiscriminate Use of Interval Scales to Rate Diverse Dysarthric Features. *Journal of Speech, Language, and Hearing Research.*