Published in final edited form as:

J Speech Lang Hear Res. 2012 August ; 55(4): 1208–1219. doi:10.1044/1092-4388(2011/11-0048).

Perceptual Measures of Speech From Individuals With Parkinson's Disease and Multiple Sclerosis: Intelligibility and Beyond

Joan E. Sussman^a and Kris Tjaden^a

^aUniversity at Buffalo, State University of New York

Abstract

Purpose—The primary purpose of this study was to compare percent correct word and sentence intelligibility scores for individuals with multiple sclerosis (MS) and Parkinson's disease (PD) with scaled estimates of speech severity obtained for a reading passage.

Method—Speech samples for 78 talkers were judged, including 30 speakers with MS, 16 speakers with PD, and 32 healthy control speakers. Fifty-two naive listeners performed forced-choice word identification, sentence transcription, or visual analog scaling of speech severity for the Grandfather Passage (Duffy, 2005). Three expert listeners also scaled speech severity for the Grandfather Passage.

Results—Percent correct word and sentence intelligibility scores did not cleanly differentiate speakers with MS, PD, or control speakers. In contrast, both naive and expert listener groups judged reading passages produced by speakers with MS and PD to be more severely impaired than reading passages produced by control talkers.

Conclusion—Scaled estimates of speech severity appear to be sensitive to aspects of speech impairment in MS and PD not captured by word or sentence intelligibility scores. One implication is that scaled estimates of speech severity might prove useful for documenting speech changes related to disease progression or even treatment for individuals with MS and PD with minimal reduction in intelligibility.

Keywords

perception; measures; intelligibility

Perceptual evaluation of dysarthria continues to be the “gold standard” for clinical decisions according to Kent and colleagues (Bunton, Kent, Duffy, Rosenbek, & Kent, 2007; Kent, 1996). Perceptual studies of dysarthria have had listeners judge vowels (Sapir, Spielman, Ramig, Story, & Fox, 2007), single words (Kent, Weismer, Kent, & Rosenbek, 1989), sentences or phrases (Hustad, 2007; Liss, Spitzer, Caviness, & Adler, 2002), spoken reading passages (Tjaden & Wilding, 2004), and even extemporaneous speech materials (Bunton et al., 2007; Bunton & Keintz, 2008; Tjaden & Wilding, 2010). Listening paradigms have included forced-choice methods and orthographic transcription as well as scaling techniques,

such as direct magnitude estimation, visual analog scaling (VAS), and equal appearing interval scales (see reviews in Kent, 1996; Weismer, 2008). Thus, a range of speech materials and listening paradigms has been used for the perceptual evaluation of dysarthria.

Traditionally, when perceptually assessing speech produced by individuals with motor speech disorders, sentence or word intelligibility measures are used to represent the status of overall functional speech ability. This statement is supported by the widespread use of the Assessment of Intelligibility of Dysarthric Speech (Yorkston & Beukelman, 1981) or the computerized version of this test, the Sentence Intelligibility Test (SIT; Yorkston & Beukelman, 1996). As discussed by Weismer (2008), intelligibility measures of this type largely focus on segmental articulation abilities and yield percent correct scores based on the number of words correctly transcribed or identified as compared with those in a target stimulus. Thus, for example, if a listener correctly transcribes 90% of a talker's speech, this individual would be considered to have mild dysarthria characterized by understandable, reasonably adequate speech (Yorkston, Beukelman, Hakel, & Dorsey, 2007).

Despite their widespread use in both research and clinical applications for dysarthria, intelligibility scores have a variety of limitations (Kent et al., 1989; Weismer, 2008; Weismer & Martin, 1992). Hustad (2008), for example, noted problems with the binomial scoring used in sentence transcription intelligibility tasks. Specifically, because words are scored as either correct or incorrect relative to the target word, a percent correct score gives equal weight to all word classes such that scores are not transparent as to whether the listener was able to decipher the meaning or intent of the spoken message. Hustad (2008) proposed that comprehension measures might be used in addition to intelligibility scores to provide a more complete picture of communicative competence for individuals with dysarthria. That is, percent correct intelligibility scores derive from accurately perceiving the surface elements of speech and thus provide information about the integrity of the acoustic signal. In contrast, comprehension refers to the listener's ability to interpret the meaning of a message produced by a speaker with dysarthria, without regard for accuracy of phonetic or lexical parsing.

In Hustad's (2008) study, 144 listeners provided judgments of intelligibility and comprehension for narratives produced by 12 speakers with cerebral palsy (CP). Speakers were operationally assigned to mild, moderate, severe, and profound severity groups on the basis of sentence intelligibility test scores (Yorkston, Beukelman, & Tice, 1996). Speech stimuli consisted of thematic narratives composed of 10 related sentences. For the intelligibility task, narratives were segmented into individual sentences and presented to listeners for orthographic transcription. For the comprehension task, listeners heard a narrative in its entirety and then answered 10 questions that assessed understanding of both inferential and factual information. Different narratives were used to obtain judgments of intelligibility and comprehension. Results indicated that comprehension scores were not significantly correlated with intelligibility scores after controlling for variation in speaker severity. In addition, descriptive statistics indicated that comprehension scores were consistently better than intelligibility scores for all dysarthria severity groups. Hustad (2008) suggested that the comprehension task elicited a "more global" impression such that listeners were able to glean the meaning of a narrative even when phonetic segments were not always correctly orthographically transcribed.

Hustad's (2008) results seem to suggest that percent correct intelligibility scores combined with global comprehension judgments provide a more complete picture of a dysarthric speaker's spoken communicative competence than either measure alone. It is interesting to note, however, that other studies suggest that global or overall perceptual constructs such as speech severity, acceptability, and naturalness do not seem to provide additional or unique information regarding spoken communication beyond that provided by judgments of intelligibility (Dagenais, Watts, Turnage, & Kennedy, 1999; Southwood & Weismer, 1993; Weismer, Jeng, Laures, Kent, & Kent, 2001). For example, Whitehill, Ciocca, and Yiu (2004) reported a strong, significant correlation between percent correct sentence intelligibility scores and interval scaling estimates of acceptability for a connected speech sample produced by speakers with a variety of neurological diagnoses, including Parkinson's disease (PD), stroke, and CP. Multiple regression analysis further suggested that intelligibility was the most important of five variables in predicting judgments of acceptability, accounting for 73% of the variance in acceptability ratings. Strain-strangled voice was the second most important predictor variable, accounting for an additional 10% of the variance in acceptability above that predicted by intelligibility. Relatedly, Weismer et al. (2001) obtained magnitude estimates of severity and intelligibility for sentences produced by healthy controls as well as speakers with dysarthria secondary to amyotrophic lateral sclerosis (ALS) and PD. When judging intelligibility, listeners were simply instructed to scale intelligibility with a focus on articulatory precision, and when judging severity, listeners were instructed to base their judgments on articulation, speaking rate, respiration, and voice. Results indicated no significant difference in scaled estimates of intelligibility or severity. These types of results seem to suggest that perceptual constructs, such as intelligibility, acceptability, severity, and normalcy, are interpreted similarly by listeners and thus are redundant.

An alternative approach to perceptual assessment of dysarthria is to obtain numerous, component-specific perceptual judgments. This methodology is at the heart of the Darley, Aronson, and Brown (1969) approach of assessing dysarthria, wherein listeners judge multiple, specific perceptual parameters from the same speech sample (e.g., "pitch level, pitch breaks, monopitch, voice tremor, monoloudness, short rushes of speech," etc. [pp. 248–249]) by using interval scaling techniques. Although Bunton et al. (2007) concluded that this approach showed promise as a technique for identifying features of dysarthria, other dysarthria studies acknowledge reliability concerns with this approach (Zeplin & Kent, 1996; Zyski & Weisiger, 1987). Perceptual studies conducted using speech samples produced by individuals with impairments in voice or resonance also have reported poor reliability for listeners' judgments of component-specific speech attributes, such as hoarseness, harshness, breathiness, or degree of nasality (e.g., Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Laczi, Sussman, Stathopoulos, & Huber, 2005). Interestingly, Kreiman et al. (1993) recommended the use of more global ratings of overall speech competence, such as "good/poor voice" or "not impaired/severely impaired."

Although intelligibility measures or even component-specific judgments of speech, such as breathiness, nasality, or imprecise consonants, might continue to have some clinical value, the preceding review indicates that there presently is no consensus as to which perceptual constructs capture the impact of speech impairment secondary to dysarthria. In an effort to

address this complex issue, the current study explored the feasibility of obtaining global perceptual judgments of speech severity for individuals with multiple sclerosis (MS) and PD and comparing these judgments of global speech severity with more traditional metrics of intelligibility. More specifically, we were interested in determining whether a global measure of speech severity that incorporated impressions of voice, resonance, articulatory precision, and prosody would provide different information about the speech of individuals with MS and PD who had only mildly reduced intelligibility, compared with percent correct word and sentence intelligibility scores. A related approach for obtaining scaled, global estimates of speech severity has been reported by Weismer et al. (2001), although it should be noted that Weismer et al. used a different type of scaling paradigm and speech materials and studied speakers with PD and ALS who had greater reductions in intelligibility than did the speakers with MS and PD in the current study. We used standard, published measures of word intelligibility (Kent et al., 1989) and sentence intelligibility (Yorkston et al., 1996) in the present experiment and obtained scaled estimates of overall speech severity for a reading passage. Perceptual judgments of speech also were obtained for a group of age- and sex-matched healthy control talkers. Inclusion of speakers with MS as well as speakers with PD ensured that findings could be generalized beyond a single neurologic diagnosis and thus provided a replication of the effect of interest.

Purpose

Our main research question was whether a scaled estimate of speech severity would convey the impact of speech impairment secondary to MS or PD that is not captured by measures of word or sentence intelligibility for speakers with largely preserved intelligibility. We further hypothesized that scaled estimates of speech severity would separate speakers with PD and MS from control talkers owing to the influence of articulatory precision, resonance, voice, and rhythm on overall speech naturalness.

Method

Speakers and Speech Stimuli

A grand total of 78 talkers provided the speech stimuli. The 32 control speakers included 10 men (range = 25–70 years, $M = 56$ years) and 22 women (range = 27–77 years, $M = 57$ years). The 16 speakers with a medical diagnosis of PD included 8 men (range = 55–78 years, $M = 67$ years) and 8 women (range = 48–78 years, $M = 69$ years). Finally, the 30 speakers with a medical diagnosis of MS included 10 men (range = 29–60 years, $M = 51$ years) and 20 women (range = 27–66 years, $M = 50$ years). Three additional control speakers, two women (ages 22 and 53) and one man (age 49), produced foil words for use as distractors in the word intelligibility task but were not part of the larger study. With the exception of these latter three control talkers, all speakers are part of ongoing research investigating the acoustic-perceptual basis of dysarthria. We have previously reported the relationship between speech-language pathologists' scaled estimates of speech severity and moment coefficients of the long-term average spectrum for a subset of these speakers (Tjaden, Sussman, Liu, & Wilding, 2010).

Participants with medical diagnoses of PD or MS were recruited through patient support groups and newsletters for PD or MS in the western New York area, while control speakers were recruited through posted flyers and advertisements. All speakers were native speakers of American English, had achieved at least a high school diploma, and had visual acuity or corrected acuity adequate for reading printed materials. Hearing aid use was an exclusion criterion. Pure-tone thresholds were obtained by an audiologist at the University at Buffalo's Speech-Language and Hearing Clinic for the purpose of providing speakers with a global indication of their auditory status. However, no speaker was excluded on the basis of pure-tone thresholds. Participants with MS and PD were taking a variety of symptomatic medications, but no participant had undergone neurosurgical treatment for MS or PD. Speakers with PD ranged from 2 to 32 years postdiagnosis ($M = 9$ years, $SD = 7.8$ years), and speakers with MS ranged from 2 to 47 years postdiagnosis ($M = 14$ years, $SD = 11$ years). Five participants with MS had a primary progressive disease course, 18 participants had a relapsing remitting disease course, and 7 participants had a secondary progressive disease course. All speakers scored at least 26 of 30 on the Standardized Mini-Mental State Examination (Molloy, 1999), with the exception of one male speaker with MS who received a score of 25 of 30. Speakers were paid a modest fee for their participation.

Each speaker was audio-recorded while producing a variety of speech materials, including 70 words from the Kent et al. (1989) single-word intelligibility test, 11 sentences from the Sentence Intelligibility Test (Yorkston et al., 1996), and the Grandfather Passage (Duffy, 2005). Thus, all speakers produced the same 70 words as well as the Grandfather Passage, but SIT sentences differed among speakers. The order in which these speech materials were recorded varied among speakers. In addition, four orderings of words from the Kent et al. (1989) test were generated, and approximately 20 speakers were assigned to produce each ordering. Online audio-recording of speech samples took place in a sound-treated or quiet room. The acoustic signal was transduced with an AKG C410 head-mounted microphone positioned 10 cm and 45° to 50° from the left oral angle. The acoustic signal was preamplified, low pass-filtered at 9.8 kHz, and digitized directly to computer hard disk at a sampling rate of 22 kHz using TF32 (Milenkovic, 2005). Speakers with PD were recorded approximately 1 hr prior to taking anti-Parkinsonian medications. Cyclical medication effects have not been documented in MS. Thus, participants with MS were recorded at a variety of times of the day.

Listeners and Perceptual Tasks: Overview

The three perceptual tasks included (a) single-word intelligibility, (b) sentence intelligibility, and (c) VAS of speech severity for the Grandfather Passage. Listeners included 52 participants recruited from flyers posted at the University at Buffalo (naive listeners) as well as three speech-language pathologists (expert listeners), for a grand total of 55 listeners. The rationale for having both naive and expert listeners judge speech severity is discussed below.

The 52 naive listeners were divided into groups by perceptual task because the large number of speakers and stimuli precluded having all listeners judge all speakers in all tasks. Thus, four groups of naive listeners ($n = 42$) participated in both the word and sentence intelligibility tasks. These 42 listeners attended two experimental sessions occurring within a

1-week period. During the first experimental session, listeners completed the word intelligibility task, and during the second session, the same listeners completed the sentence intelligibility task. A fifth group of 10 listeners as well as the three expert listeners completed the VAS speech severity task in a single experimental session. Perceptual tasks were completed in a double-walled Industrial Acoustics Company audiometric booth, and stimuli were presented binaurally via Sony MDR V300 headphones. Presentation of the stimuli was controlled with custom software developed for use in the Speech Perception Laboratory at the University at Buffalo (Johnson & Sussman, 2008). All stimuli were presented at the same sound pressure levels at which they were naturally produced by the speakers and ranged from 68 to 83 dB SPL (Bruel and Kjaer SLM, Model 2215, C Scale) for peak amplitude of each stimulus.

Each of the 52 naive listeners passed a hearing screening at 20 dB SPL HL (American National Standards Institute, 2004) for octave pure-tone frequencies from 250 to 8000 Hz, bilaterally. These listeners also had achieved at least a high school diploma and reported minimal exposure or experience with speech produced by individuals with motor speech disorders. Naive listeners ranged in age from 19 to 44 years (word and sentence listeners: 29 women, mean age = 22 years, $SD = 3.5$ years; and 13 men, mean age = 24.5 years, $SD = 7.7$ years; VAS and naive listeners: 8 women, mean age = 24.5 years, $SD = 5.1$ years; 2 men, mean age = 22 years, $SD = 1.4$ years) and were paid for their participation. Each of the three speech pathologists had at least 10 years of clinical experience with motor speech disorders and was paid a modest consultation fee.

Word Intelligibility Task: Stimuli and Procedures

A computerized four-alternative forced-choice paradigm (Johnson & Sussman, 2008) presented four words per item (target word and three foils) on a computer screen in four equally spaced boxes. Listeners used a computer mouse to click the box corresponding to their response, and the response was saved to computer hard disk for further analysis. Ten percent of trials were repeated for the purpose of computing intrajudge reliability.

Forty-two naive listeners were divided into four groups so that each listener could comfortably judge both word and sentence stimuli in two 90-min sessions. To ease the listeners' task by reducing listener uncertainty (e.g., Carney, Widin, & Viemeister, 1977), we divided stimuli according to speaker group and distributed the number of stimuli approximately equally across listener groups, except for a final group of speakers who were recruited late during the data collection period. Written instructions for listeners indicated that some but not all of the speech samples they would be hearing were produced by speakers with neurological diagnoses like PD or MS. During the actual perceptual tasks, listeners were unaware of the identity or neurological status of the speech sample being judged.

One subset of 11 listeners judged stimuli produced by speakers with PD (PD listeners), while the other three groups of 10 to 11 listeners heard speech produced by a subset of the speakers with MS (MS1 listeners, MS2 listeners, MS3 listeners). PD listeners heard a total of 2,563 words, whereas the MS1 and MS2 listeners heard 2,486 words, and the MS3 listener group judged 1,031 words. In addition to stimuli produced by MS and PD speakers,

included in these stimuli totals are stimuli produced by a smaller group of control speakers who were matched by sex and age within 3 years of the targeted disordered speakers assigned to their listening group. Thus, stimuli for each group of listeners had approximately one third of the stimuli produced by typical speakers and two thirds of the stimuli produced by speakers with PD or MS. Foil words produced by the three additional control talkers were also presented for word identification to provide additional variation in stimuli to help prevent listener bias as discussed by Weismer (2008). Having foil words presented for identification allowed listeners to expect that those could be possible words to label as alternatives even though PD and MS speakers produced only the 70 target words selected from the Kent et al. (1989) task. All listeners heard all of the trials from the three additional foil speakers. We generated two different orders of words and sentences, and listeners were randomly assigned to a given order and group (i.e., PD listeners, MS1 listeners, and so forth).

For a given listener's data to be included, we required that intrajudge reliability be at least .60, as measured by a Pearson product-moment correlation. We excluded data from five additional listeners because their reliability levels were below this criterion level. Finally, for each listener, we calculated a percent correct score for the 70 words produced by each speaker for use in the statistical analysis.

Sentence Intelligibility Task: Stimuli and Procedures

The same 42 naive listeners who judged single-word intelligibility also judged sentence intelligibility on a customized computerized sentence transcription task developed for the purposes of the study. Sentence stimuli for each speaker included 11 SIT sentences plus one extra for computation of intrajudge reliability. The same groups of listeners (PD1, MS1, MS2, MS3) transcribed sentences produced by the same speakers who had produced words that were judged in the prior session. Each listener orthographically transcribed 314 (PD), 303 (MS1, MS2), or 97 (MS3) sentences. Listeners heard each sentence via headphones and then typed their response on a computer keyboard. The listeners were then asked to confirm their typing and enter the response before moving to the next sentence, which was administered 500 ms afterward. We generated two random orderings of sentences for each group of listeners (i.e., PD listeners, MS1 listeners, and so forth) and randomly assigned listeners to a given order.

Sentence transcriptions were first scored by the computer program. If the transcription was identical to the target stimulus, then a match for the total number of words in the sentence was recorded. When exact matches did not occur, two trained research assistants manually counted the number of words correctly transcribed. The total number of matching words was compared with the total number of words in the target sentences. Spelling errors and homonyms were disregarded when determining a word match. Intrajudge and interjudge reliability for word-match judgments was calculated for 10% of each of the research assistants' data at 98% and 99%, respectively. In addition, we assessed listeners' intrajudge reliability with a Pearson product-moment correlation with a required performance of at least .60. We did not use data from the same five listeners as in the word task because of

poor intrajudge reliability. Finally, we tabulated for each listener an overall percent correct score for each speaker's sentences for use in the statistical analysis.

VAS of Speech Severity: The Grandfather Passage

An additional group of 10 naive listeners as well as the three expert listeners scaled speech severity for the Grandfather Passage. Two groups of listeners were included as there is some discrepancy in the literature as to whether expert and naive listeners judge dysarthric speech differently on perceptual scaling or rating tasks (see discussion of this issue by Bunton et al., 2007; Kent, 1996; Zyski & Weisiger, 1987). It was further speculated that the scaling task would be more open to interpretation compared with forced-choice word identification or orthographic transcription of sentences, and we were interested in empirically determining whether listener experience would impact findings. In the voice literature (e.g., Kreiman et al., 1993), expert listeners have been found to judge voice disorders more harshly than naive listeners, whereas in judging nasality, expert listeners have been shown to judge less severely compared with naive listeners (Laczi et al., 2005). Monsen (1978, 1983) also reported that naive and expert listeners had different ratings of intelligibility for speech produced by children with hearing impairment. Monsen (1978) further showed that after practice with the listening tapes, differences due to experience were reduced. However, such potential differences due to experience motivated the current interest in determining whether expert and naive listeners might differ in scaling speech severity for individuals with MS and PD.

Speech severity was judged for passages produced by all 78 speakers on a continuous 150-mm vertical scale (Cannito, Burch, Watts, & Rappold, 1997). Scale end points ranged from 0 at the bottom (*no impairment*) to 1.0 at the top (*severe impairment*). Approximately 10% of passages ($n = 8$) were presented to listeners twice for use in calculating intrajudge reliability. Each listener was required to have an intrajudge reliability correlation coefficient for original and reliability trials of at least .60 to be included for study. Written instructions were displayed on a computer screen and were read to each listener. These instructions—broadly similar to those used by Weismer et al. (2001) in their study reporting magnitude estimates of scaled severity in dysarthria—are shown below.

You will be hearing samples of paragraph readings that for the most part, are highly understandable. We want you to rate the overall severity of the speech sample. Some speakers have neurological diagnoses (e.g., have diseases like Parkinson's Disease or Multiple Sclerosis) and some do not. Please pay attention to the following things when you listen to the passages:

1. Voice (quality—breathy, noisy, gurgly, high pitch, too low pitch, or OK)
2. Resonance (too nasal, not nasal in the right places, sounds like they have a cold, or OK)
3. Articulatory precision (some sounds are crisp or slurred or somewhere in between or OK), and
4. Speech rhythm (the timing of speech doesn't sound right or is OK).

In other words, pay attention to overall speech naturalness and prosody (melody and timing of speech). Do not focus on the speaker's

intelligibility or how understandable each passage is. Rather, scale your overall impression of the speech/voice output from *no impairment* (at the bottom of the scale) to *severely impaired* (at the top of the scale).

Listeners were presented with seven practice trials to become familiar with the task and to ensure that they understood how to use the VAS before beginning experimental trials. Training trials were composed of passage readings produced by speakers with MS, speakers with PD, and control talkers who have participated in some of our previous studies but were not included in the current study (e.g., Tjaden & Sussman, 2006).

We used custom computer software (Johnson & Sussman, 2008) to convert the position of the mouse pointer on the VAS line to a scale value ranging from 0 to 1.0. This scale value was subsequently multiplied by 100 for ease of comparison across tasks. Scale values for the 10 listeners were then averaged to provide a single scale value for each Grandfather Passage for use in the statistical analysis.

Data Analysis

Prior to statistical analysis, all percentage values were arcsine transformed as recommended by Winer (1981) to normalize the distributions of data. For both word and sentence intelligibility scores, we used one-way analysis of variance (ANOVA) to evaluate potential differences in intelligibility for speakers with MS, speakers with PD, and healthy controls, divided into groups by sex because of the different numbers of speakers in each group. Scaled estimates of speech severity for the naive and expert listeners were strongly correlated (Pearson product-moment correlation coefficient $r = .909$, $p < .01$) and were not significantly different from one another as observed in a one-way ANOVA, $F(1, 155) = 3.502$, $p = .063$. Thus, only data for the larger naive listener group were included in the one-way ANOVA examining potential differences in scaled estimates of speech severity as a function of speaker group and sex (i.e., control speaker female, control speaker male, MS female, MS male, PD female, and PD male). For each of the three ANOVAs, we used a Bonferroni-corrected alpha level of .017 to ascertain statistical significance ($.05/3 = .017$). We used Tukey's honestly significant difference (HSD) post hoc tests to further analyze significant main effects, with a standard alpha level of .05 to ascertain significance of post hoc tests. Finally, we used Pearson correlations to assess the relationship among word intelligibility, sentence intelligibility, and scaled estimates of speech severity. Again, owing to the strong correlation between naive and expert listeners' scaled estimates of speech severity, only data for naive listeners were used in this correlation analysis.

Results

For all perceptual tasks, results are reported by speaker group and sex so that readers can see all trends in the data, particularly since there were different numbers of speakers in each of the groups (control, PD, and MS). Mean percent correct scores also are reported to facilitate data interpretation, although percent correct scores were transformed for the statistical analyses. Figure 1 reports average percent correct scores and SDs for the word intelligibility task. As shown in Figure 1, on average, word intelligibility scores were greater than 94%. Listeners were least accurate identifying words produced by female speakers with PD, but

there was no significant main effect of speaker group in the ANOVA, $F(5, 78) = 2.67, p = .029$.

Figure 2 shows average percent correct scores and *SDs* for the sentence intelligibility task. The ANOVA indicated a significant effect of speaker group, $F(5, 78) = 6.72, p < .0001$. Post hoc Tukey's HSD comparisons further showed that the female speakers with PD were significantly different from both the male and female control speakers as well as from the male and female speakers with MS. Thus, the female PD group was responsible for the significant main effect of speaker group in the ANOVA. Figure 2 further indicates that, on average, sentence intelligibility exceeded 90%, with the exception of sentences produced by female speakers with PD. Variability, as shown by *SD* bars, also tended to be greater for most disordered speaker groups compared with control talkers.

Figures 3 and 4 report scaled estimates of speech severity for the Grandfather Passage as a function of speaker group and sex. As reported below, all data analyses focused on naive listeners, but findings for expert listeners are provided in Figure 3 for completeness. For ease of comparison with the word and sentence intelligibility tasks, we converted the *y*-axis in Figures 3 and 4 from the 0–1.0 VAS by multiplying by 100, as the exact value is not critical but the relationship is preserved. As indicated in Figures 3 and 4, scaled estimates of speech severity averaged approximately 50%–65% for expert listeners and 40%–50% for naive listeners for both the MS and PD groups. Those scaled scores were significantly different from those obtained for the control talkers, whose scores averaged at approximately 20% for both expert and naive listener groups, close to the “not impaired” end of the VAS.

As previously noted, scaled estimates of speech severity for expert and naive listener ratings were highly correlated and were not significantly different from one another. Thus, we calculated an ANOVA using only the naive listener ratings. Results indicated a significant main effect of speaker group, $F(5, 78) = 7.22, p < .001$. Post hoc pairwise comparisons using a Tukey's HSD indicated that control speakers were judged as a separate homogeneous group from the speakers with MS and PD ($p < .021$). The female speakers with PD were rated as the most severe group. In summary, findings for scaled estimates of speech severity differed from those for word intelligibility and sentence intelligibility. The scaled estimates of speech severity—but not word or sentence intelligibility scores—differentiated control speakers from speakers with MS and PD. Stated differently, MS and PD speakers who were just as intelligible as controls at the word and sentence level were judged to have significantly poorer speech severity for the Grandfather Passage compared with control talkers.

Finally, the correlation analysis indicated that word and sentence intelligibility scores were significantly correlated (Pearson $r = .551; p < .01$, two-tailed). Word intelligibility and scaled estimates of speech severity ($r = -.526; p < .01$, two-tailed) as well as sentence intelligibility and scaled estimates of speech severity ($r = -.598; p < .01$, two-tailed) were significantly correlated. That is, when word or sentence intelligibility scores were higher or better, scaled estimates of speech severity also were better (i.e., toward *no impairment*, or 0). Thus, all types of perceptual judgments tended to be moderately correlated despite different results in the ANOVAs for intelligibility and scaled speech severity.

Discussion

Results from the current study suggest that percent correct word or sentence intelligibility scores do not fully capture listeners' impressions of speech adequacy for speakers with MS and PD, at least when intelligibility is reasonably preserved. Word intelligibility was nearly perfect when groups of naive listeners judged words in a four-alternative forced-choice paradigm, even when foil words were included in the task to reduce listener bias, as has been recommended (Weismer, 2008). Similarly, on average, sentence intelligibility was over 90% for all speaker groups, except for female speakers with PD and, even then, over 82% correct on average. Word and sentence intelligibility scores for speakers in the present study were found to be significantly correlated. Other studies also have reported a strong relationship between measures of word and sentence intelligibility in dysarthria, although it is notable that metrics of intelligibility for speakers with neurological diagnoses in some of these previous studies encompassed a different range (e.g., Bunton & Keintz, 2008; Weismer, Yunusova, & Bunton, 2012; Yorkston & Beukelman, 1978; Yunusova, Weismer, Kent, & Rusche, 2005). For example, word intelligibility for speakers with ALS or PD in the Yunusova et al. (2005) study ranged from 68% to 98% ($M = 84\%$), whereas word intelligibility for speakers with MS or PD in the current study ranged from only 88% to 99% ($M = 96\%$). The fact that sentence intelligibility scores in the current study were not obviously higher or better than word intelligibility scores, as reported previously for speakers with mild dysarthria (Hustad, 2007; Yorkston & Beukelman, 1978), likely is due to the fact that intelligibility metrics in the present study were close to ceiling.

In contrast to the finding that word and sentence intelligibility scores did not differ among speakers with MS, speakers with PD, and healthy controls, when listeners heard the Grandfather Passage and then used a continuous VAS to register their global impression of speech severity, perceptual judgments were found to be significantly poorer for speakers with MS or PD. These results are broadly consistent with those of Weismer et al. (2001), who found that a perceptual scaling technique that encompassed factors such as voice quality and prosody was useful in separating ALS and PD speaker groups with similar word intelligibility scores. The fact that even naive listeners in the current study judged speakers with MS and PD to have moderately impaired overall speech severity, which in turn was significantly poorer than that of healthy control talkers, may have real-world implications for individuals ranging in age from 50 to approximately 65 years. Those individuals might still need or want to be employed or involved in social activities where typical speech proficiency is assumed. Control speakers in the present study, who were of similar age to speakers with MS and PD, had average scale values for speech severity of approximately 20%. The finding that scaled estimates of speech severity for control speakers were not at zero—indicating no impairment—likely reflects listeners' sensitivity to normal age-related changes in speech, perhaps especially changes related to voice or prosody. The written instructions indicating that “some speakers have neurological diagnoses like Parkinson's Disease or Multiple Sclerosis and some do not” also may have biased listeners to expect some level of speech impairment.

We speculate that the reason for the different findings for scaled estimates of speech severity as compared with word or sentence intelligibility had to do with the instructions given to

listeners who were directed to scale the overall severity of speech not according to how understandable a given passage was but according to how impaired the speaker sounded on the basis of voice, resonance, articulatory precision, and rhythm. In addition to differences in the listening paradigm (i.e., forced-choice identification, orthographic transcription, and VAS), however, speech materials also varied (i.e., words, sentences, and paragraph reading). It therefore cannot be stated definitively that results are attributable to a true difference between intelligibility and scaled speech severity (for related discussion concerning scaling techniques and multiple-choice single-word identification in dysarthria, see Weismer et al., 2001). Future investigations can now build on the current findings by comparing scaled estimates of intelligibility and scaled estimates of overall speech severity for the same speech materials. Studies including speakers with greater reductions in intelligibility also are needed to determine whether a metric of scaled severity might be useful only for certain portions of the overall severity continuum.

The present study was not intended to identify the speech component or components (i.e., articulatory precision, resonance, voice, rhythm, and vocal intensity) responsible for the group differences in scaled speech severity. Nor was the current study intended to evaluate the extent to which listeners might have differentially weighted the various speech components (i.e., articulatory precision, resonance, voice, rhythm, and vocal intensity) when making their judgments of overall severity. As noted by Weismer et al. (2012), however, an important topic for future research is to identify the speech production variables that explain the types of severity judgments for speakers with MS and PD reported in the present study. Anecdotally, it was noted that many of the speakers with PD had reductions in segmental precision as well as a breathy, monotonous voice quality. Speakers with MS also tended to exhibit reduced segmental precision, although prosodic and voice deficits tended to be more variable for the MS group, with some talkers exhibiting a slow speech rate coupled with excess and equal stress, while other talkers presented with voice quality changes in the form of increased vocal harshness or hoarseness. Thus, in the same way that a given intelligibility score (i.e., 75% sentence intelligibility) can be explained by a variety of underlying speech production variables, a given scaled severity score also could be explained by a variety of underlying speech production variables.

Our suggestion that perceptual measures in addition to percent correct single word and sentence intelligibility scores are needed to characterize speech impairment in dysarthria is not unique. A variety of investigators have pointed out the need for metrics in addition to single-word intelligibility, as single-word tasks do not adequately capture intelligibility of everyday connected speech samples (e.g., Bunton & Keintz, 2008; Duffy & Giolas, 1974; Yorkston & Beukelman, 1977; for a comprehensive treatment of this issue, see Weismer, 2008). However, single-word tasks still continue to be included in published intelligibility tests for dysarthria (e.g., Enderby & Palmer, 2008; Yorkston et al., 1996), no doubt because these tests provide at least some information regarding segmental accuracy and are relatively easy to administer and score.

Some investigators have advocated that sentence stimuli be used when obtaining perceptual judgments of dysarthria, including perceptual judgments of intelligibility (e.g., Frearson, 1985; Giolas, 1966; Weismer et al., 2001). Studies by Monsen (1978, 1983) that examined

transcription sentence intelligibility for children with hearing impairment showed that results could vary as a result of many factors, including the phonotactic and/or syntactic complexity of the stimuli, whether listeners were experienced or naive, whether sentences were presented in their contextual setting or separated from that, and whether listeners were able to view speakers as the sentences were produced or they could hear only the productions. Attempts to control for such variables have no doubt improved the validity of traditional intelligibility measures. Indeed, Frearson (1985) pointed out the usefulness of the sentence portion of Yorkston and Beukelman's (1981) Assessment of Intelligibility of Dysarthric Speech, because reliability across judges was significantly better for transcription of sentences than for transcription of spontaneous speech materials. Sentence intelligibility scores overestimated spontaneous speech intelligibility, however. Weismer et al.'s (2001) finding that scaled estimates of sentence intelligibility separated ALS and PD speaker groups differently than single-word intelligibility also would seem to support using sentence-level stimuli when obtaining perceptual judgments of dysarthria, although it is possible that listening paradigm differences (i.e., multiple-choice scoring format versus scaling technique) contributed to this result.

It also has been suggested that spontaneous speech should be used when obtaining perceptual judgments of dysarthria (e.g., Bunton & Keintz, 2008; Frearson, 1985; Kempler & Van Lancker, 2002). For example, Kempler and Van Lancker (2002) investigated how four groups of listeners judged the speech from one individual with PD as he produced spontaneous speech, repeated the spontaneous speech sample, read a passage, and sang. For all speech tasks, listeners orthographically transcribed stimuli. Findings showed that spontaneous speech was significantly less intelligible than other speech materials. More recently, Tjaden and Wilding (2010) concluded that a reading task showed promise for representing intelligibility of extemporaneous speech from a monologue. Using direct magnitude estimation scaling techniques, they found no significant difference in scaled intelligibility for reading and monologue tasks produced by speakers with PD.

In conclusion, results from the current study suggest that a measure of scaled severity for a paragraph reading task shows promise in providing different or supplementary information regarding speech impairment as compared with traditional metrics of percent correct single word or sentence intelligibility, at least for speakers with neurological diagnoses who have relatively preserved intelligibility. In this manner, a metric of scaled severity might prove useful for documenting speech changes related to disease progression or even treatment-related changes for individuals with MS and PD with minimal reduction in intelligibility. Our results, as well as those of others (e.g., Hustad, 2008; Weismer, 2008; Weismer et al., 2001), would seem to argue for the importance of conducting a more comprehensive set of perceptual metrics for dysarthria beyond measures of percent correct word or sentence intelligibility. As Kent et al. (1989, p. 495) stated, "Different tests are suited to different purposes." Their original word intelligibility test was designed to provide a "phonetic interpretation of impaired intelligibility" (Kent et al., 1989, p. 495). Overall, a growing body of evidence suggests that investigators and clinicians incorporate perceptual measures that can tap into a variety of "levels," including surface acoustics–phonetics that could be revealed from intelligibility tests (e.g., Kent et al., 1989; Yorkston et al., 2007), semantic-related comprehension measures using targeted questions about the meaning of narratives

(e.g., Hustad, 2008), and overall speech severity measures that appear to incorporate listeners' sensitivity to voice, prosody, and other speech parameters.

Acknowledgments

This research was supported by National Institute on Deafness and Other Communication Disorders Grant R01 DC004689. Results were presented at the 2009 American Speech-Language-Hearing Association Annual Convention, in New Orleans, LA. We thank Gary Weismer for his helpful comments. We thank Stacy Blackburn and Ken Johnson for their assistance with data collection and reduction.

References

- American National Standards Institute. Specifications for audiometers. New York, NY: Author; 2004. ANSI S3.6-2004
- Bunton K, Keintz CK. Effects of a concurrent motor task on speech intelligibility in speakers with Parkinson disease. *Journal of Medical Speech-Language Pathology*. 2008; 16:141–155. [PubMed: 21637738]
- Bunton K, Kent RD, Duffy J, Rosenbek J, Kent J. Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language, and Hearing Research*. 2007; 50:1481–1495.
- Cannito M, Burch A, Watts C, Rappold P. Disfluency in spasmodic dysphonia: A multivariate analysis. *Journal of Speech, Language, and Hearing Research*. 1997; 40:627–641.
- Carney AE, Widin GP, Viemeister NF. Noncategorical perception of stop consonants differing in VOT. *The Journal of the Acoustical Society of America*. 1977; 4:961–970.
- Dagenais P, Watts C, Turnage L, Kennedy S. Intelligibility and acceptability of moderately dysarthric speech by three types of listeners. *Journal of Medical Speech-Language Pathology*. 1999; 7:91–96.
- Darley F, Aronson A, Brown J. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*. 1969; 12:246–269. [PubMed: 5808852]
- Duffy J. *Motor speech disorders: Substrates, differential diagnosis, and management*. 2. St. Louis, MO: Elsevier Mosby; 2005.
- Duffy J, Giolas T. Sentence intelligibility as a function of key word selection. *Journal of Speech and Hearing Research*. 1974; 17:631–637. [PubMed: 4444284]
- Enderby, P., Palmer, R. *Frenchay Dysarthria Assessment*. 2. Austin, TX: Pro-Ed; 2008.
- Frearson B. A comparison of the AIDS Sentence List and Spontaneous Speech Intelligibility scores for dysarthric speech. *Australian Journal of Human Communication Disorders*. 1985; 13:5–21.
- Giolas T. Comparative intelligibility scores of sentence tests and continuous discourse. *Journal of Auditory Research*. 1966; 6:31–38.
- Hustad KC. Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with cerebral palsy. *Folia Phoniatrica et Logopaedica*. 2007; 59:306–317. [PubMed: 17965573]
- Hustad KC. The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research*. 2008; 58:562–573.
- Johnson, K., Sussman, J. *Experimental software for perceptual experiments*. Buffalo, NY: University at Buffalo, State University of New York; 2008.
- Kempler D, Van Lancker D. Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. *Brain and Language*. 2002; 80:449–464. [PubMed: 11896652]
- Kent RD. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*. 1996; 5:7–23.
- Kent RD, Weismer G, Kent J, Rosenbek J. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*. 1989; 54:482–499. [PubMed: 2811329]
- Kreiman J, Gerratt B, Kempster G, Erman A, Berke G. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*. 1993; 36:21–40. [PubMed: 8450660]

- Laczi E, Sussman J, Stathopoulos E, Huber J. Perceptual evaluation of hypernasality compared to HONC measures: The role of experience. *Cleft Palate & Craniofacial Journal*. 2005; 42:202–211.
- Liss J, Spitzer S, Caviness J, Adler C. The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America*. 2002; 112:3022–3030. [PubMed: 12509024]
- Milenkovic, P. TF32 [Computer program]. Madison, WI: University of Wisconsin—Madison; 2005.
- Molloy, D. Standardized Mini-Mental State Examination. Troy, NY: New Grange Press; 1999.
- Monsen RB. Toward measuring how well hearing-impaired children speak. *Journal of Speech and Hearing Research*. 1978; 21:197–219. [PubMed: 703272]
- Monsen RB. The oral speech intelligibility of hearing-impaired talkers. *Journal of Speech and Hearing Disorders*. 1983; 48:286–296. [PubMed: 6621019]
- Sapir S, Spielman J, Ramig L, Story B, Fox C. Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*. 2007; 50:899–912.
- Southwood M, Weismer G. Listener judgments of the bizarreness, acceptability, naturalness and normalcy of dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology*. 1993; 1:151–161.
- Tjaden K, Sussman J. Perception of coarticulatory information in normal speech and dysarthria. *Journal of Speech, Language, and Hearing Research*. 2006; 49:888–902.
- Tjaden K, Sussman J, Liu G, Wilding G. Long-term average spectral (LTAS) measures of dysarthria and their relationship to perceived severity. *Journal of Medical Speech-Language Pathology*. 2010; 18:125–132.
- Tjaden K, Wilding G. Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*. 2004; 47:766–783.
- Tjaden K, Wilding G. Effects of speaking task on intelligibility in Parkinson's disease. *Clinical Linguistics & Phonetics*. 2010; 1:1–14.
- Weismer, G. Speech intelligibility. In: Ball, MJ.Perkins, MR.Muller, N., Howard, S., editors. *The handbook of clinical linguistics*. Oxford, England: Blackwell; 2008. p. 568-582.
- Weismer G, Jeng J, Laures J, Kent R, Kent J. Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica*. 2001; 53:1–18. [PubMed: 11125256]
- Weismer, G., Martin, R. Acoustic and perceptual approaches to the study of intelligibility. In: Kent, RD., editor. *Intelligibility in speech disorders*. Philadelphia, PA: John Benjamins; 1992. p. 67-118.
- Weismer G, Yunusova Y, Bunton K. Measures to evaluate the effects of DBS in speech production. *Journal of Neurolinguistics*. 2012; 25:74–94. doi:10.1016.2011.08.006. [PubMed: 24932066]
- Whitehill T, Ciocca V, Yiu E. Perceptual and acoustic predictors of intelligibility and acceptability in Cantonese speakers with dysarthria. *Journal of Medical Speech-Language Pathology*. 2004; 12:229–233.
- Winer, B. *Statistical principles of experimental design*. New York, NY: McGraw-Hill; 1981.
- Yorkston K, Beukelman D. A communication system for the severely dysarthric speaker with an intact language system. *Journal of Speech and Hearing Disorders*. 1977; 42:265–270. [PubMed: 67278]
- Yorkston K, Beukelman D. A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders*. 1978; 11:499–512. [PubMed: 739065]
- Yorkston K, Beukelman D. Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders*. 1981; 46:296–301. [PubMed: 7278175]
- Yorkston, K., Beukelman, D. *Sentence Intelligibility Test for Windows*. Lincoln, NE: Tice Technologies; 1996.
- Yorkston, K., Beukelman, D., Hakel, M., Dorsey, M. *Sentence Intelligibility Test for Windows*. Lincoln, NE: Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital; 2007.

- Yorkston, K., Beukelman, D., Tice, R. Sentence Intelligibility Test. Lincoln, NE: Tice Technologies; 1996.
- Yunusova Y, Weismer G, Kent RD, Rusche NM. Breath-group intelligibility in dysarthria: Characteristics and underlying correlates. *Journal of Speech, Language, and Hearing Research*. 2005; 48:1294–1310.
- Zeplin, J., Kent, RD. Reliability of auditory perceptual scaling of dysarthria. In: Robin, DA, Yorkston, KM., Beukelman, DR., editors. *Disorders of motor speech: Assessment, treatment, and clinical characterization*. Baltimore, MD: Brookes; 1996. p. 145-154.
- Zyski B, Weisiger B. Identification of dysarthria types based on perceptual analyses. *Journal of Communication Disorders*. 1987; 20:367–378. [PubMed: 3655012]

Figure 1.

Mean percent correct intelligibility scores and *SDs* for the word intelligibility task are reported for female and male speakers in each of the three speaker groups (control, multiple sclerosis, and Parkinson's disease). CSF = control speakers, female; CSM = control speakers, male; MSF = multiple sclerosis, female; MSM = multiple sclerosis, male; PDF = Parkinson's disease, female; PDM = Parkinson's disease, male.

Figure 2. Mean percent correct intelligibility scores and *SDs* for the sentence intelligibility task are reported for female and male speakers in each of the three speaker groups (control, MS, and PD).

Figure 3. Mean visual analog scaling (VAS) speech severity ratings and *SDs* are reported for expert listeners' judgments of the Grandfather Passage. Severity ratings are reported separately for female and male speakers in each of the three speaker groups (control, MS, and PD).

Figure 4. Mean VAS speech severity ratings and *SDs* are reported for naive listeners' judgments of the Grandfather Passage. Severity ratings are reported separately for female and male speakers (control, MS, and PD).