ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Validity of Off-the-Shelf Automatic Speech Recognition for Assessing Speech Intelligibility and Speech Severity in Speakers With Amyotrophic Lateral Sclerosis

Sarah E. Gutz,[a] Kaila L. Stipancic,[b] Yana Yunusova,[c,d,e] James D. Berry,[f] and Jordan R. Green[a,g]

[a] Program in Speech and Hearing Bioscience and Technology, Harvard Medical School, Boston, MA [b] Department of Communicative Disorders and Sciences, University at Buffalo, NY [c] Department of Speech-Language Pathology, University of Toronto, Ontario, Canada [d] Hurvitz Brain Sciences Program, Sunnybrook Research Institute, Toronto, Ontario, Canada [e] Toronto Rehabilitation Institute, University Health Network, Ontario, Canada [f] Sean M. Healey and AMG Center for ALS, Massachusetts General Hospital, Boston [g] Department of Communication Sciences and Disorders, MGH Institute of Health Professions, Boston, MA

ARTICLE INFO

Article History:
Received November 2, 2021
Revision received January 21, 2022
Accepted March 15, 2022

Editor-in-Chief: Bharath Chandrasekaran
Editor: Kate Bunton

https://doi.org/10.1044/2022_JSLHR-21-00589

ABSTRACT

Purpose: There is increasing interest in using automatic speech recognition (ASR) systems to evaluate impairment severity or speech intelligibility in speakers with dysarthria. We assessed the clinical validity of one currently available off-the-shelf (OTS) ASR system (i.e., a Google Cloud ASR API) for indexing sentence-level speech intelligibility and impairment severity in individuals with amyotrophic lateral sclerosis (ALS), and we provided guidance for potential users of such systems in research and clinic.
Method: Using speech samples collected from 52 individuals with ALS and 20 healthy control speakers, we compared word recognition rate (WRR) from the commercially available Google Cloud ASR API (Machine WRR) to clinician-provided judgments of impairment severity, as well as sentence intelligibility (Human WRR). We assessed the internal reliability of Machine and Human WRR by comparing the standard deviation of WRR across sentences to the minimally detectable change (MDC), a clinical benchmark that indicates whether results are within measurement error. We also evaluated Machine and Human WRR diagnostic accuracy for classifying speakers into clinically established categories.
Results: Human WRR achieved better accuracy than Machine WRR when indexing speech severity, and, although related, Human and Machine WRR were not strongly correlated. When the speech signal was mixed with noise (noise-augmented ASR) to reduce a ceiling effect, Machine WRR performance improved. Internal reliability metrics were worse for Machine than Human WRR, particularly for typical and mildly impaired severity groups, although sentence length significantly impacted both Machine and Human WRRs.
Conclusions: Results indicated that the OTS ASR system was inadequate for early detection of speech impairment and grading overall speech severity. While Machine and Human WRR were correlated, ASR should not be used as a one-to-one proxy for transcription speech intelligibility or clinician severity ratings. Overall, findings suggested that the tested OTS ASR system, Google Cloud ASR, has limited utility for grading clinical speech impairment in speakers with ALS.

Correspondence to Jordan R. Green: jgreen2@mghihp.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

The demand for automated speech analysis systems is increasing due to their potential value as biomarkers for a variety of mental and physical health conditions (Low et al., 2020; Toth et al., 2018). The current standard for

assessing speech impairment severity requires licensed speech-language pathologists (SLPs), who use ordinal descriptors (e.g., mild, moderate, and severe; King et al., 2012; Tjaden & Liss, 1995). While found to be reliable (Stipancic et al., 2021), this assessment requires experienced listeners, can be time intensive and costly, and can be biased by the assessor's familiarity with the speaker, speech disorder, and subject matter (King et al., 2012; Tjaden & Liss, 1995). The need for more objective and automatic methods for assessing speech severity in motor speech disorders is widely recognized for a variety of research and clinical applications, including improved diagnosis, symptom monitoring, and intervention design (King et al., 2012; Tjaden & Liss, 1995).

Off-the-shelf automatic speech recognition (OTS ASR) systems are an attractive candidate for this application, because they are low cost, simple to implement, and widely available. The proportion of words incorrectly recognized by these systems, or word error rate (WER), could presumably serve as a quantitative index of overall speech impairment. Because OTS ASR platforms are trained on typical speech, recognition accuracy degrades as speech becomes more atypical or less intelligible (De Russis & Corno, 2019; Mustafa et al., 2015). Prior research in people both with and without speech impairment has linked ASR accuracy to speech intelligibility (Ferrier et al., 1995; Jacks et al., 2019; McHenry & Laconte, 2010; Riedhammer et al., 2007), with correlations reported between .80 and .98 (Ferrier et al., 1995; Jacks et al., 2019; Riedhammer et al., 2007). Similarly, Tu et al. (2016) found a strong correlation between ASR WER and perceptually rated severity WER (Pearson $r = .80$).

Despite these advantages, the efficacy of ASR has been understudied for speech severity grading. Moreover, previous work has identified several threats to ASR validity, including biases from language models. For example, language models can boost accuracy by aiding word prediction, or they can decrease accuracy when errors lower the probability of correctly selecting nearby words (Keshet, 2018). The word-level transcription of most OTS ASR systems could further obfuscate subtle differences at the phone level, as a mild distortion and a major articulatory deviation could lead to the selection of the same incorrect word (Keshet, 2018). There is also evidence that some speech deviations affect ASR accuracy more than others (Benzeghiba et al., 2007; Goldwater et al., 2010; Keshet, 2018; Tu et al., 2016), such that speakers with certain dysarthria etiologies or subsystem impairments could be erroneously classified as more or less severe regardless of actual dysarthria severity. Moreover, reports that humans and ASR systems produce different errors when transcribing speech limit the reliability of ASR for tracking functional speech changes (Mulholland et al., 2016). These limitations suggest that ASR may be unpredictable when indexing severity and problematic for judging clinically relevant speech differences.

In the current project, we tested the clinical validity of a widely available OTS ASR system (Google Cloud ASR) for grading speech severity in persons with ALS (Goldsack et al., 2020; Google LLC, 2020). Clinical validation is defined as whether a measure "acceptably identifies, measures, or predicts a meaningful clinical, physical, functional state, or experience, in the stated context of use" (Goldsack et al., 2020). Other groups have made inroads in clinically validating ASR for dysarthria by investigating the relationship between perceptual severity measures and ASR transcription (Tu et al., 2016). For example, comparing human transcription and ASR transcription, Jacks et al. (2019) found very high correlations (Spearman $\rho = .96$–$.98$) using IBM Watson for speakers with aphasia and/or apraxia of speech (AOS) following a stroke; Maier et al. (2010) reported Spearman *rho* between −.88 and −.90 for a hidden Markov model (HMM)—based ASR system used on head and neck cancer patients with dysglossia and dysphonia; and Ballard et al. (2019) found agreement of 75.7% between human and ASR (using CMU Pocket-Sphinx) judgments of word-level productions by people with aphasia and AOS following stroke. Looking at the relationship between Google ASR accuracy and clinician-rated severity, Tu et al. (2016) found a moderate correlation (Pearson $r = .69$) for speakers with dysarthria when using the Google ASR engine. Still, the evidence for clinical validity of ASR—as applied to a specific clinical population—remains scant and has primarily been evaluated for aphasia and AOS (e.g., Ballard et al., 2019; Jacks et al., 2019).

We focused on the Google Cloud Speech ASR system as it is a top-performing freely available ASR system, with a low WER for healthy speakers and a documented gradient for speakers with dysarthria: WER of 3.95% for healthy speakers, 16.11% for mildly or unimpaired speakers with ALS or cerebral palsy (CP), and 78.21% for severely impaired speakers with ALS or CP (De Russis & Corno, 2019). Furthermore, Google Cloud Speech Recognition is accessible through several coding languages and platforms, such as MATLAB and python (e.g., MathWorks Audio Toolbox Team, 2022; Zhang, 2017) as well as consumer-accessible programs like the Google search function. Thus, although Google Cloud ASR was not developed for clinical purposes, it is a functional system that researchers and clinicians alike can easily use (e.g., De Russis & Corno, 2019; Tu et al., 2016). Throughout this article, we refer to the results from Google Cloud Speech Recognition simply as "ASR" or "Machine WRR" (see below).

We addressed the limitations of previous work by using a relatively large ($N = 72$ speakers) data set encompassing neurologically healthy control (HC) speakers and the full speech severity range in patients with just one dysarthria etiology, amyotrophic lateral sclerosis (ALS). We focused on clinically validating ASR in people with dysarthria secondary to

ALS—a disease characterized by the progressive degeneration of upper and lower motor neurons (resulting in spastic, flaccid, or mixed spastic–flaccid dysarthria; Tomik & Guiloff, 2010)—because speech severity is a common metric of bulbar disease progression in ALS (Green et al., 2013). We closely considered ASR's relationship with speech intelligibility due to intelligibility's strong relationship with severity (Rong, Yunusova, Wang, et al., 2015) and because it is commonly used as a severity proxy or stratification tool in research (Gutz et al., 2019; Rong, Yunusova, & Green, 2015; Stipancic et al., 2018, 2021). Speech intelligibility is measured as the percentage of spoken words that a listener can correctly transcribe (Miller, 2013). While intelligibility is a component of impairment severity, many factors influence severity, including resonance, articulation, respiration, phonation, prosody, and intelligibility (Darley et al., 1969). Thus, the two measures allow distinct insights into the clinical validity of ASR.

Clinicians and researchers use other metrics to assess dysarthria as well. For example, standardized assessments such as the Frenchay Dysarthria Assessment (Enderby, 1980) and patient report measures such as the Revised Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) (Cedarbaum et al., 1999) are often used to determine the presence or severity of dysarthria. Clinicians and researchers may also use acoustic and kinematic measures such as rate, consistency, and lip movement on diadochokinesis tasks (DDK) to assess dysarthria and AOS in Parkinson's disease, multiple sclerosis, stroke, traumatic brain injury, and ALS (Rowe et al., 2020; Tjaden & Watling, 2003; Ziegler, 2002). Speaking rate, which degrades before intelligibility in ALS, has also been used to rate speech severity in people with dysarthria, including those with ALS (Rong, Yunusova, Wang, et al., 2015). Additionally, while some work indicates that many features that influence severity are redundant to human listeners and therefore add little beyond intelligibility (Weismer et al., 2001), a growing body of work suggests that features independent of intelligibility, such as comprehensibility and naturalness, impact clinical measures of speech severity (Hustad, 2008; Sussman & Tjaden, 2012). However, because our aim was to evaluate ASR as a measure of speech impairment severity, and not to characterize said impairment, we focused on clinician-rated severity and speech intelligibility, measures that have already been associated with ASR performance, as mentioned above.

To contextualize ASR's performance, we compared ASR with transcription intelligibility for all measures under consideration. Additionally, the data set was well curated in terms of clinical speech labels and ground-truth human orthographic transcriptions, completed by 10 experienced SLPs and previously reported in Stipancic et al. (2021).

Our study was designed to address the following three research questions: (RQ1) Is the OTS ASR system output a valid representation of transcription speech intelligibility (convergent validity)? (RQ2) Is the OTS ASR system output a valid proxy for clinician-rated severity groupings (known-groups validity)? and (RQ3) Is the OTS ASR system accuracy reliable across sentences of varying length and content (internal reliability)?

Method

Procedure

This project used data collected as part of the same data set used by Stipancic et al. (2021), who previously reported data for the speakers, listeners, and the reliability of human transcription intelligibility used in this study.

Participants—speakers. We collected audio samples from 72 speakers (see Table 1), who participated in a larger study on bulbar motor involvement in ALS (Green et al., 2013). Our sample included both HC speakers and speakers with ALS. Of the participants with ALS, 25 had spinal onset, 21 had bulbar onset, three had mixed onset, and three had unknown onset. The mean ALSFRS-R score was 31 (SD = 9.62).

Participants—listeners. Ten SLPs served as clinician listeners. All listeners had a master's degree in speech-language pathology and had been practicing for a mean of 6.6 years (range: 2–14 years, SD = 5.1 years) and had experience working with patients with dysarthria (M = 6.25 years, range: .5–.14 years, SD = 5.2 years). SLP listeners rated the speech samples remotely through the online survey platform REDCap (Harris et al., 2009). Listeners could listen to each sample recording twice, using headphones. Listeners completed an online hearing screening tool before completing the REDCap survey to control for individual listening volume (Miracle Ear, 2018). Listener participants and data collection protocol (described below) were previously reported in Stipancic et al. (2021).

Speech sample. Speech samples were collected from the Speech Intelligibility Test (SIT; Yorkston et al., 2007). Each speaker read a set of 11 randomly selected sentences increasing incrementally in length from five to 15 words. Audio (32-bit, mono, 44.1 kHz) was recorded with a head-fixed microphone 1 in. from the speaker's mouth.

ASR Measures

Machine word recognition rate (WRR) was calculated as 1 minus the WER and then multiplied by 100. WER, a common metric for evaluating ASR accuracy, is presented as the rate of transcription errors (insertions, deletions, and substitutions) relative to the sentence length and was calculated as the Levenshtein distance (Levenshtein, 1966) between the target SIT text and the text transcribed by the Google Cloud Speech API (Google LLC, 2020). We accessed Google Cloud Speech through the python library Speech

Table 1. Participant stratification and demographic information, as well as onset site, ALSFRS-R score, Human word recognition rate (WRR), and Machine WRR.

Group	Clinician-rated severity	*n*			Age (years) *M* (SD)	Sp.	Onset site (*n*)			ALSFRS-R *M* (*SD*)	*M* (*SD*; %)	
		All	F	M			Bul.	Mix.	Un.		Human WRR	Machine WRR
ALS & HC	Normal	23	12	11	62.87 (9.31)	n/a	n/a	n/a	n/a	n/a	97.82 (2.29)	83.12 (14.04)
HC	Normal	20	11	9	63.22 (9.37)	n/a	n/a	n/a	n/a	n/a	98.10 (1.99)	85.45 (13.46)
ALS	Normal	3	1	2	60.59 (10.55)	2	0	1	0	38.00 (2.83)	95.96 (3.71)	67.59 (5.60)
ALS	Mild	11	4	7	58.59 (8.36)	6	3	1	1	25.78 (10.60)	92.83 (3.41)	76.40 (10.74)
ALS	Moderate	11	3	8	58.64 (12.12)	8	2	0	1	30.75 (8.75)	77.13 (17.00)	48.02 (26.09)
ALS	Severe	14	4	10	56.28 (11.28)	4	9	1	0	33.22 (9.15)	58.41 (19.62)	17.70 (12.60)
ALS	Profound	13	10	3	61.61 (8.14)	5	7	0	1	32.25 (9.94)	18.09 (18.87)	2.86 (4.94)
	Overall	72	33	39	60.01 (9.93)	25	21	3	3	31.00 (9.62)	71.84 (32.1)	49.52 (35.81)

Note. F = female; M = male; Sp. = spinal; Bul. = bulbar; Mix. = mixed; Un. = unknown; ALS = participants with ALS diagnosis; HC = healthy controls; n/a = data not collected for HC participants.

Recognition (Zhang, 2017) and used it to transcribe each recorded sentence individually. We evaluated the efficacy of the Google Cloud API, because it is freely available (Zhang, 2017) and has been used in prior work to evaluate speech intelligibility (Dimauro et al., 2017; Vásquez-Correa et al., 2018). Although we focused on the Google Cloud Speech API, validation of other systems (e.g., IBM Watson, Microsoft, Amazon) is warranted for similar reasons.

Clinical Measures

Clinician-rated severity was our main standard of comparison. Clinician-provided perceptual severity ratings reflected expert opinion on speech based on the listener's total percept, which would account for factors such as naturalness and subsystem involvement in addition to intelligibility. Clinician raters were instructed to "Please indicate the severity of the speech for this individual" as normal, mild, moderate, severe, or profound. These data and methods were previously reported in Stipancic et al. (2021), who found high intrarater reliability (% agreement = .94) and high interrater reliability (weighted κ = .91) for the current sample. The clinician ratings were averaged across the two clinician listeners and rounded up to the nearest whole number (i.e., the more severe rating). Two listeners provided ratings to allow for interrater reliability calculation. Although we had data for HCs and people with ALS whom clinicians rated as "normal," we combined speakers from both groups into the clinician-rated "normal" category for analyses. This approach was in keeping with our aim of evaluating ASR relative to clinician-rated severity, rather than distinguishing the speech of HCs from people with ALS with "normal" speech.

Human WRR (sentence-level speech intelligibility) was derived at the sentence level using the SIT. Two SLP listeners orthographically transcribed each sample, writing word for word what they heard the speaker say. Each listener transcribed 24 unique SIT samples and two reliability samples. Accuracy for each sentence was calculated as the number of correctly identified words divided by the number of target words and then multiplied by 100 to obtain percent intelligibility. Average sentence intelligibility for each speaker was calculated as the mean sentence intelligibility of the 11 SIT sentences, averaged across the two listeners. For the sentence-level transcriptions of these data, Stipancic et al. (2021) found high intrarater reliability (average ICC = .91, $p < .001$) and high interrater reliability (ICC = .94, $p < .001$).

Statistical Methods

(RQ1) Convergent validity. To assess the convergent validity of Machine WRR with Human WRR, we computed Pearson correlation between Machine WRR and Human WRR. We tested the fit of linear and quadratic equations to assess if Machine WRR could be considered a one-to-one equivalent of Human WRR, or if their relationship was nonlinear.

To test how well Machine WRR tracks Human WRR for high and low intelligibility speakers, we compared Machine WRR between groups stratified by Human WRR. We stratified speakers by Human WRR using SIT upper limits reported in prior work; cutoffs for Human WRR Intelligibility Groups 1–5 were, respectively, 100%, 96%, 90%, 80%, and 50% (Stipancic et al., 2018, Table 2). In prior work, these groups have been labeled as having normal, mild, moderate, severe, and profound intelligibility impairment (Stipancic et al., 2018); however, we report them here as Groups 1–5 to distinguish them from the clinician-rated severity groups. Participants are often stratified by intelligibility as a proxy for severity (e.g., Blaney & Hewlett, 2007; Connaghan & Patel, 2017); thus, there is value in assessing the ability of Machine WRR to match intelligibility stratification independent of clinician-rated severity. We conducted pairwise *t*-test comparisons of Machine WRR between all Human WRR groups using Bonferroni correction for 10 comparisons, α = .005 (.05/10).

Table 2. Human word recognition rate (WRR) stratification (RQ1).

Intelligibility group	Human WRR range (%)	Human WRR (%)		Machine WRR (%)	
		M	*SD*	*M*	*SD*
1	96 ≤ WRR ≤ 100	98.76	1.04	82.09	13.06
2	90 ≤ WRR < 96	93.66	1.58	76.79	18.91
3	80 ≤ WRR < 90	83.43	1.81	48.84	21.72
4	50 ≤ WRR < 80	63.72	7.88	23.54	12.83
5	WRR < 50	21.14	16.81	3.99	4.84

Note. Mean and *SD* of Human WRR and of Machine WRR for each intelligibility group, as well as the range of Human WRR used to define the intelligibility group.

(RQ2) Known-groups validity. To assess known-groups validity, we compared Human WRR and Machine WRR between clinician-rated severity groups. We first used clinician-rated severity to group participants into normal, mild, moderate, severe, and profound severity groups (see Table 1). Then, we compared Machine WRR scores between each pair of clinician-rated severity groups using *t*-test comparisons with Bonferroni correction, α = .005 (.05/10). Similarly, we compared Human WRR scores between clinician-rated severity groups with Bonferroni correction, α = .005 (.05/10). Additionally, we used receiver operating characteristic (ROC) analyses through the pROC package (Robin et al., 2011) to determine optimal diagnosis thresholds between adjacent clinician-rated severity groups for both measures; we computed sensitivity, specificity, and accuracy of Machine WRR and Human WRR for classification. We further used classification and regression trees (CARTs) to compare the utility of using of Machine WRR or Human WRR for classifying speakers with different impairment severity levels and to test the diagnostic utility of both measures. For the CART decision trees, we used the package rpart (Therneau & Atkinson, 2019) and pruned the trees using the least cross-validated error. To address a possible ceiling effect in Machine WRR, we created CART for noise-augmented ASR (nASR; described below).

nASR was employed to counter a ceiling effect observed for the normal and mild severity groups for Machine WRR CART classification (speakers with Machine WRR ≥ 52%). We used nASR to explore how ASR could be modified to better stratify speakers with dysarthria. For nASR, we mixed speech samples with multitalker babble at integer signal-to-noise ratios (SNRs) from −5 to +10 dB, using a custom MATLAB script. We then input these noise-mixed signals through the ASR system, as reported in previous work from our laboratory (Gutz et al., 2021). We tested the ASR transcriptions at each SNR using the CART method described above. Here, we only report results for nASR with an SNR of 0 dB, because these CARTs had the highest accuracy of the tested SNRs. Even though ASR is trained on noise-perturbed samples (Park et al., 2020), a noisy speech signal can degrade ASR performance (Krishna et al., 2019), which could separate Machine WRRs for normal and mild groups.

(RQ3) Internal reliability. Because each SIT sentence is a different length, we assessed whether sentence length impacted WRR for Human and Machine intelligibility. We used mixed-effects models with participant as a random intercept, and we controlled for clinician-rated severity by including it as a fixed effect in the equation lmer(WRR ~ sentence + severity + [1 | participant]; Bates et al., 2015).

We further assessed intraspeaker reliability by comparing Machine WRR variation (standard deviation [*SD*]) across SIT sentences within a given speaker to the minimally detectable change (MDC), a clinical benchmark specific to each severity group. A change greater than or equal to the MDC indicates a change outside the measurement error (Stratford & Riddle, 2012). In other words, an *SD* larger than the MDC would indicate variation within the SIT sentence set that cannot be explained by measurement error or the speaker alone.

Stipancic et al. (2018) found that the MDC varied with a speaker's speech intelligibility and calculated the MDC as $1.96 \times \sqrt{2} \times$ the standard error of measurement for intelligibility, for a set of 196 speakers (147 with ALS and 49 HCs). Stipancic et al. (2018) calculated MDCs for groups that were stratified by intelligibility. However, in keeping with our aim of clinically validating ASR according to clinician-rated severity groups, we stratified participants by their clinician-rated severity groups. We then assigned each severity group an MDC from the corresponding intelligibility group from Stipancic et al. (2018); for example, we used the MDC calculated for the moderately impaired intelligibility group for the moderate clinician-rated severity group. We compared the *SD*s of Human and Machine WRRs with the MDC for each severity group. We also compared the *SD*s of Human WRR with those of Machine WRR. To test whether speakers' intelligibility varied across sentences more or less than the MDC benchmark, we conducted *t* tests using Bonferroni correction, α = .01 (05/5) for each family of tests and α = .05 for tests conducted for the full set of participants.

Results

(RQ1) Convergent Validity: Relationship Between Machine WRR and Human WRR

Human WRR and Machine WRR were strongly correlated, Pearson $r(70) = .87$, $p < .001$. Linear and quadratic models fit to the Machine WRR and Human WRR relationship demonstrated that the quadratic fit, $F(2, 69) = 207.4$, $p < .001$, $r^2 = .86$, was stronger than the linear fit, $F(1, 70) = 226.4$, $p < .001$, $r^2 = .76$. AIC for the quadratic model (AIC = 586.38) was less than the AIC for the linear model (AIC = 620.70), indicating that the relationship between Machine

Table 3. Model fits for Machine and Human word recognition rate (WRR; RQ1).

Model for Machine WRR and Human WRR	*df*	*F*	r^2
Linear	70	226.37	.76
Quadratic	69	207.39	.85

Note. Degrees of freedom (*df*), *F*-statistic (*F*), and adjusted r^2 for linear and quadratic models fit to Machine WRR (response variable) and Human WRR (predictor). $p < .0001$ for both models, with heteroskedasticity correction for the quadratic model. An analysis of variance comparing the two models showed a statistically significant difference between models, $F(1) = 45.26$, $p < .001$.

WRR and Human WRR was nonlinear (see Table 3 and Figure 1). Results from an analysis of variance indicated that the quadratic model was a significantly better fit than the linear model, $F(1) = 45.26$, $p < .001$. The quadratic model indicated a slope close to zero for low-intelligibility speakers and an increasing slope starting around Human WRR of 40%–60% (see Figure 1). Furthermore, while *t* tests showed that Machine WRR differed among most Human WRR stratification groups at the $\alpha = .005$ (.05/10) level, it did not differ between the two groups (Groups 1 and 2) with the highest Human WRRs ($p = .256$; see Figure 2 and Table 4).

(RQ2) Known-Groups Validity: Evaluating the Relationship of Machine WRR and Human WRR With Clinician-Rated Severity Groups

Summary statistics for Human WRR and Machine WRR showed generally higher Human WRR and Machine WRR for groups rated as less impaired by clinicians (see Table 1). Human WRR was different among most clinician-rated severity groups at the $\alpha = .005$ (.05/10) level using *t* tests, except between normal and mild ($p = .321$) and between mild and moderate ($p = .009$; see Figure 3 and Table 5). Machine WRR differed between

Figure 1. Machine and Human word recognition rate (WRR) linear and quadratic models (RQ1). Machine WRR plotted against Human WRR with linear (dashed line) and quadratic (solid line) models.

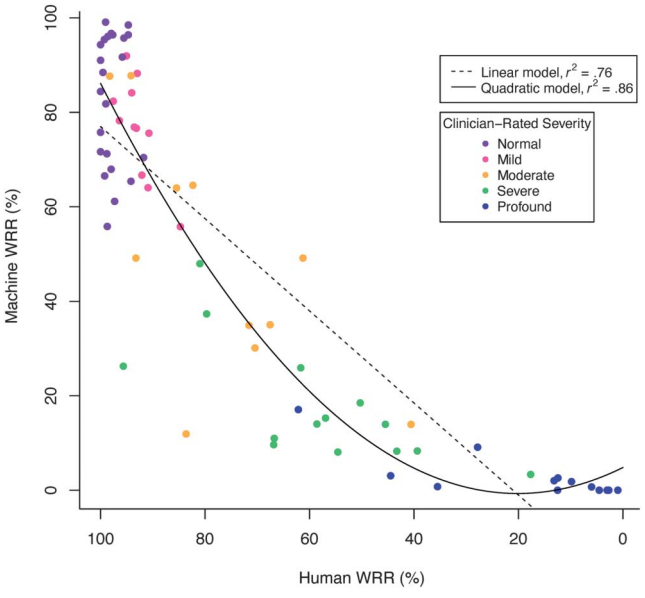


Figure 2. Machine word recognition rate (WRR) compared with Human WRR stratification groups (RQ1). Machine WRR (%) as a function of Human WRR groups. A lower group number indicates less impaired intelligibility. The Speech Intelligibility Test (SIT) score upper limits for Human WRR Groups 1–5 were, respectively, 100%, 96%, 90%, 80%, and 50%. Human WRR thresholds are listed below each stratification group. Except for Group 1—for which both numbers are included in the range—the lower number is inclusive, and the upper limit is exclusive, for example, Group 3 is defined as 90% > WRR ≥ 80%. Brackets indicate nonsignificant pairwise comparisons; all other pairwise comparisons were significant, $\alpha = .005$ (.05/10).

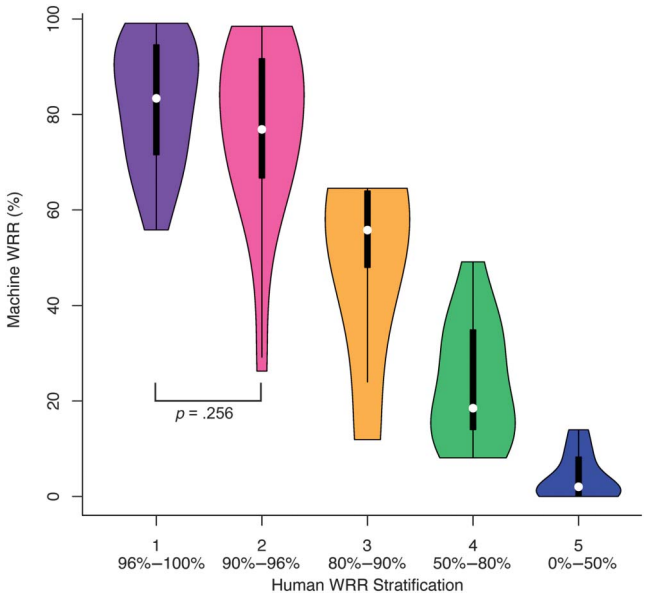
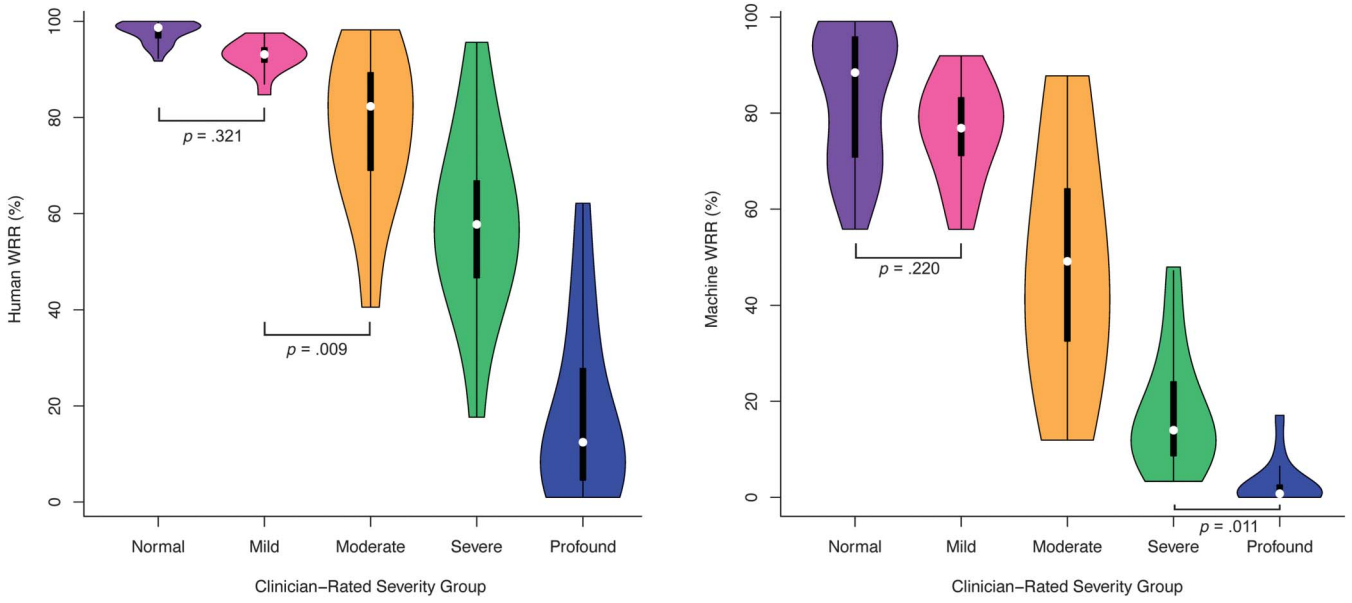


Table 4. Human word recognition rate (WRR) stratification pairwise comparisons (RQ1).

Human WRR groups compared	*t*	*p*
1 & 2	1.15	.256
1 & 3	4.74	< .001
1 & 4	11.71	< .001
1 & 5	16.87	< .001
2 & 3	3.91	< .001
2 & 4	10.30	< .001
2 & 5	15.12	< .001
3 & 4	3.42	.001
3 & 5	6.28	< .001
4 & 5	3.78	< .001

Note. *t* statistic (*t*) and *p* value (*p*) for each *t* test between pairs of intelligibility groups. All comparisons were evaluated at the $\alpha = .005$ (.05/10) level, $df = 67$.

Figure 3. Human word recognition rate (WRR) and Machine WRR by clinician-rate severity group (RQ2). Human WRR (left) and Machine WRR (right) by clinician-rated severity group. Brackets indicate nonsignificant pairwise comparisons; all other pairwise comparisons were significant, α = .005 (.05/10).



most clinician-rated severity groups except between normal and mild (p = .220) and between severe and profound (p = .011; see Figure 3 and Table 5).

(RQ2) Known-Groups Validity: ROC Curves Used to Determine the Utility of Machine WRR and Human WRR for Classifying Clinician-Rated Severity Groups

Human WRR ROC curves showed high accuracy (accuracy ≥ .90; see Table 6) for all severity group cutoffs; Machine WRR ROC curves had high accuracy (accuracy ≥ .90; see Table 6) for nearly all group cutoffs, except the cutoff between the normal and mild groups, which was slightly lower (accuracy = .81; see Table 6). While the Machine WRR ROC curve classifying the normal and mild groups had high sensitivity (1.00), it had lower specificity (.71). The cutoff for the severe and profound groups was 44.97% for Human WRR and 9.37% for Machine WRR; the cutoff point between the normal and mild groups was 94.17% for Human WRR and 55.80% for Machine WRR. Thus, the dynamic range for optimal thresholds for the Human WRR ROC curves was about 37% points higher than the dynamic range for the Machine WRR ROC curve thresholds (see Table 6).

Table 5. Clinician-rated severity group pairwise comparisons (RQ2).

Severity groups compared	Human WRR		Machine WRR	
	t	*p*	*t*	*p*
Normal & mild	1.00	.321	1.24	.220
Normal & moderate	4.15	< .001	6.48	< .001
Normal & severe	8.54	< .001	13.05	< .001
Normal & profound	16.89	< .001	15.64	< .001
Mild & moderate	2.71	.009	4.50	< .001
Mild & severe	6.28	< .001	9.85	< .001
Mild & profound	13.41	< .001	12.14	< .001
Moderate & severe	3.42	.001	5.09	< .001
Moderate & profound	10.59	< .001	7.45	< .001
Severe & profound	7.70	< .001	2.61	.011

Note. *t* statistic (*t*) and *p* value (*p*) for each *t* test between pairs of clinician-rated severity groups. One set of *t* tests compared Human word recognition rate (WRR; left), and another compared Machine WRR (right) between each pair of clinician-rated severity groups. All comparisons were evaluated at the α = .005 (.05/10) level, *df* = 67.

(RQ2) Known-Groups Validity: Decision Tree Classification Using CART

The unpruned Machine WRR tree had six branches, which were pruned to three; the unpruned Human WRR tree had five branches, pruned to four. The pruned Machine WRR tree (see Figure 4, middle) acted as a three-class classifier and sorted speakers into three groups: normal, severe, and profound, which resulted in an accuracy of 0 for both the mild and moderate groups. The pruned Human WRR tree (see Figure 4, left) classified speakers into four groups: normal, mild, severe, and profound. The Machine WRR decision tree had an overall

Table 6. Results from receiver operating characteristic (ROC) curves (RQ2).

Measure	Cutoff point between	Threshold	Specificity	Sensitivity	Accuracy	AUC
Human	Normal & mild	94.17	.90	.96	.92	.97
Human	Mild & moderate	84.18	.87	1.00	.93	.96
Human	Moderate & severe	81.66	.96	.89	.92	.96
Human	Severe & profound	44.97	.92	.93	.93	.98
Machine	Normal & mild	55.80	.71	1.00	.81	.90
Machine	Mild & moderate	52.47	.89	1.00	.94	.96
Machine	Moderate & severe	48.57	1.00	.89	.93	.98
Machine	Severe & profound	9.37	.92	.93	.93	.98

Note. Optimal threshold, specificity, sensitivity, and accuracy for each threshold, as well as area under the curve (AUC), for ROC curves created to classify each of the pairs of two groups listed, for Human word recognition rate (WRR; Human) and Machine WRR (Machine). For specificity and sensitivity, classification as the more severely impaired group is considered a positive classification.

accuracy of .67, and the Human WRR decision tree had an overall accuracy of .74 (see Table 7).

The highest performing nASR tree had speech mixed with noise at an SNR of 0 dB (see Figure 4, right). The pruned and unpruned nASR tree both had two branches. This nASR tree had an accuracy of .91 for the normal group and .73 for the mild group. Overall accuracy was .75 when predicted values for normal and mild from the nASR tree were combined with predicted values for moderate, severe, and profound from the unaltered Machine WRR tree (see Table 7).

(RQ3) Internal Reliability: Effect of Sentence Length on WRR

When controlling for clinician-rated severity, sentence length had a significant effect on Human WRR, $F(1, 719) = 29.46$, $p < .001$, and Machine WRR, $F(1, 719) =$

Figure 4. Decision Trees for Human word recognition rate (WRR), Machine WRR, and noise-augmented automatic speech recognition (nASR). Pruned decision trees using classification and regression tree (CART) for Human WRR (left), Machine WRR (middle), and nASR with a 0 dB signal-to-noise ratio (SNR) for normal/mild classification. Divisions show decision thresholds for group classification. The Human WRR tree had 74% accuracy; the Machine WRR tree had 67% accuracy; the nASR tree had 76% accuracy for the normal and mild groups. For the nASR tree, samples were chosen for classification using the noise-mixed signal if the Machine WRR CART classified them as normal (Machine WRR $\geq$ 52). Each box displays the predicted severity group with the number of actual speakers in each severity group predicted for that group (normal, mild, moderate, severe, and profound) for the Human WRR and Machine WRR trees; (normal, mild, and moderate) for the nASR tree.

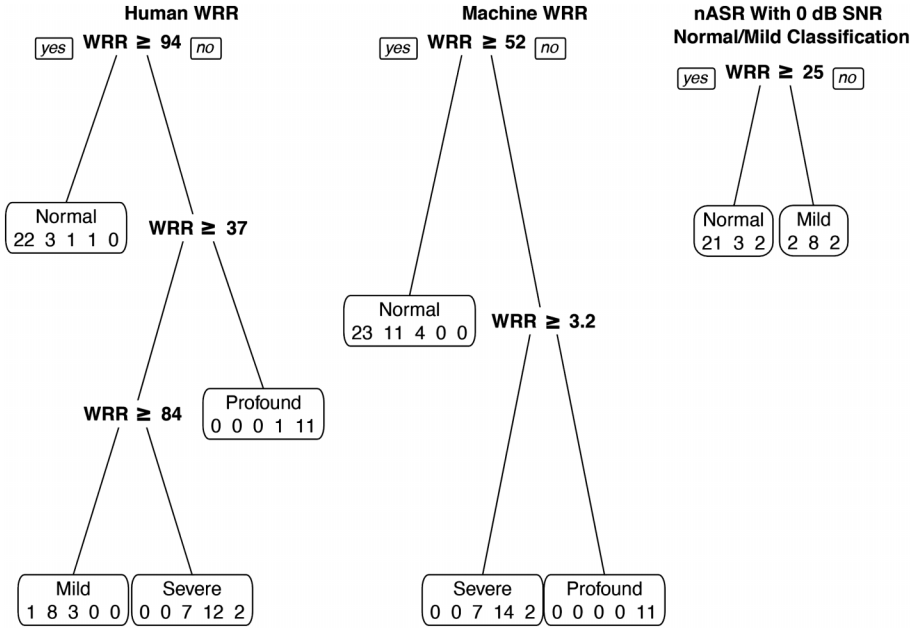


Table 7. Decision tree output and accuracy (RQ2).

		Pruned Human WRR decision tree					
		Predicted					
		Normal	**Mild**	**Moderate**	**Severe**	**Profound**	**Accuracy**
Actual	Normal	22	1	0	0	0	.96
	Mild	3	8	0	0	0	.73
	Moderate	1	3	0	7	0	.00
	Severe	1	0	0	12	1	.86
	Profound	0	0	0	2	11	.85
						Overall	*.74*
		Pruned Machine WRR decision tree					
		Predicted					
		Normal	**Mild**	**Moderate**	**Severe**	**Profound**	**Accuracy**
Actual	Normal	23	0	0	0	0	1.00
	Mild	11	0	0	0	0	.00
	Moderate	4	0	0	7	0	.00
	Severe	0	0	0	14	0	1.00
	Profound	0	0	0	2	11	.85
						Overall	*.67*
		Combined ASR (Machine WRR) & nASR decision trees					
		Predicted					
		Normal	**Mild**	**Moderate**	**Severe**	**Profound**	**Accuracy**
Actual	Normal	21	2	0	0	0	.91
	Mild	3	8	0	0	0	.73
	Moderate	2	2	0	7	0	.00
	Severe	0	0	0	14	0	1.00
	Profound	0	0	0	2	11	.85
						Overall	*.75*

Note. Classification and regression trees (CARTs). Confusion matrices for pruned decision trees for Human word recognition rate (WRR; top), Machine WRR (middle), and combined noise-augmented automatic speech recognition (nASR) with ASR (bottom). Actual group displayed in rows, predicted in columns. For combined trees (bottom), predicted values from the 0 dB signal-to-noise ratio (SNR) nASR decision tree are between dashed vertical lines. All other values are from the baseline ASR (Machine WRR) decision tree. Accuracy for the nASR tree alone classifying normal and mild groups was .76.

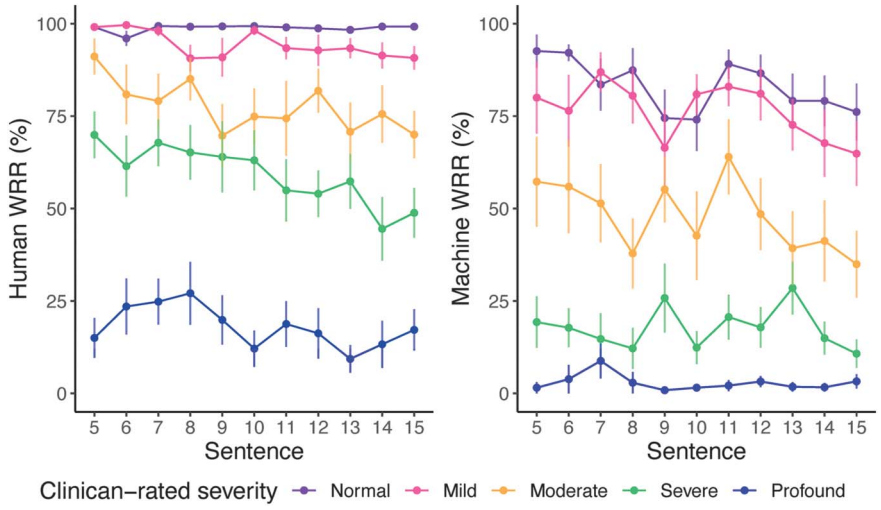
11.11, $p < .001$ (see Figure 5). For Machine WRR, there was no interaction between sentence and clinician-rated severity, $F(4, 715) = 1.32$, $p = .260$. However, for Human WRR, there was an interaction between sentence and clinician-rated severity, $F(4, 715) = 6.36$, $p < .0001$, such that the more severe groups (moderate, severe, and profound) showed overall greater decline in WRR as sentence length increased. Even controlling for this interaction, there was still a significant effect of sentence length on Human WRR, $F(4, 715) = 30.38$, $p < .001$. We observed a ceiling effect across all sentences in the normal group for Human WRR and a floor effect in the profound group for Machine WRR across all sentences.

(RQ3) Internal Reliability: Standard Deviation Compared With the MDC

Overall, intraspeaker variability (*SD*) was greater for Machine WRR (*SD* = 19.58%) than Human WRR (*SD* = 11.41%; see Table 8). When compared by individual severity group, Machine WRR *SD* was greater than that of Human WRR for the normal, mild, and moderate severity groups (see Figure 6 and Table 8). This difference was significant at the $\alpha = .05$ level when all severity groups were considered and at the $\alpha = .01$ (.05/5) level for the normal and mild groups (see Figure 6, right, and Table 8). Machine WRR *SD* was significantly less than Human WRR *SD* for the profound group (see Figure 6 and Table 8).

Comparing measures to the MDC, we found that the *SD* of Human WRR was significantly *less than* the MDC for the mild, severe, and profound severity groups, $\alpha = .01$ (.05/5; see Figure 6, left, and Table 8). At the $\alpha = .01$ (.05/5) level, the *SD* of Machine WRR for normal, mild, and moderate severity groups was significantly *greater than* the MDC (see Figure 6, right, and Table 8); Machine WRR *SD* for severe and profound was significantly *less than* the MDC (see Figure 6, right, and Table 8). However, due to a floor effect for Machine WRR scores in the profound

Figure 5. Human and Machine word recognition rate (WRR) scores across the Speech Intelligibility Test (SIT) sentence set (RQ3). Line plots of Human WRR (left) and Machine WRR (right) by sentence, grouped by clinician-rated severity. Error bars indicate standard error. WRR in both models differed significantly by sentence and clinician-rated severity ($p < .001$).



group (Machine WRR at or near 0% for most speakers), *SD* is not a meaningful measure for this group.

Discussion

Summary

In this study, we investigated the adequacy of an OTS ASR system trained on typical speech, for the clinical grading of impaired speech due to ALS. ASR estimates of speech intelligibility and severity were moderately correlated with those made by human raters, specifically experienced SLPs. The dynamic range of ASR was reduced relative to that of the clinician transcription intelligibility, defined by poor responsiveness to severity on the bottom end (floor effect) and dispersion on the top end, such that poor performance was observed even for participants with no or mild speech impairments. Accuracy and stability in Machine WRR performance were particularly poor for typical and mildly impaired severity groups, decreasing its utility as an early indicator of neurologic involvement. These findings suggest that the tested OTS ASR system, Google Cloud Speech, has limited utility for grading speech impairments and may indicate the need for ASR systems more generally to be trained on dysarthric speakers specifically for clinical purposes (see Table 9).

(RQ1) Convergent Validity: There Is a Nonlinear Association Between Machine WRR and Human WRR

We found a strong correlation between Machine and Human WRR, which was weaker than that found by some previous work, for example, Jacks et al. (2019) who

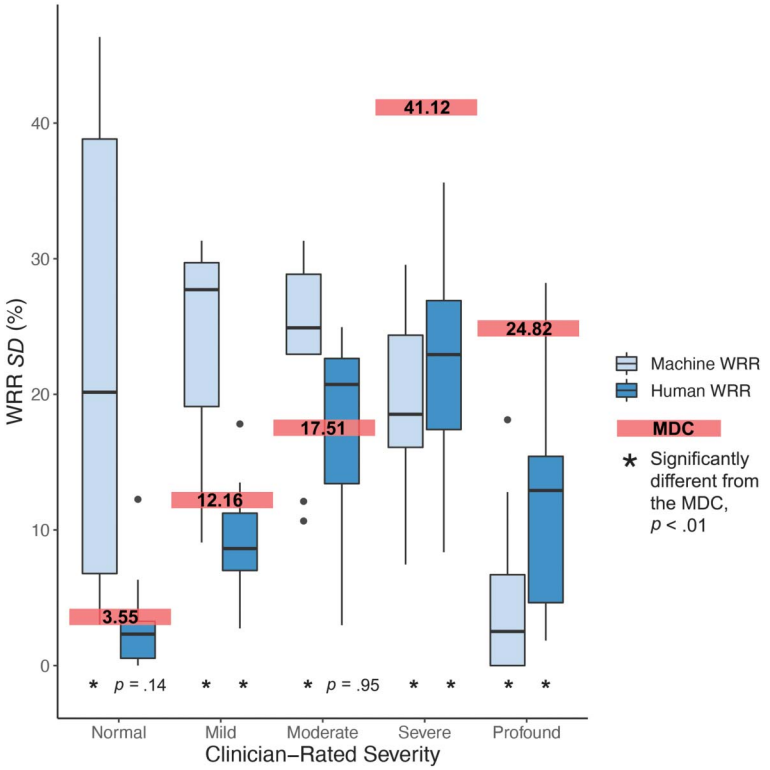
Table 8. Within-subject standard deviation of Human word recognition rate (WRR) and Machine WRR across Speech Intelligibility Test (SIT) sentences (RQ3).

Clinician-rated severity group	MDC	Human WRR *SD* vs. MDC			Machine WRR *SD* vs. MDC			Human vs. Machine *SD*	
		M (*SD*)	*t*	*p*	*M* (*SD*)	*t*	*p*	*t*	*p*
ALL	12.07	11.41 (9.33)	−0.60	.55	19.58 (12.73)	4.54	< .01[b]	4.05	< .01[c]
Normal	3.55	2.67 (2.77)	−1.53	.14	22.66 (16.34)	5.49	< .01[b]	5.66	< .01[c]
Mild	12.16	9.11 (4.27)	−2.37	.04[a]	24.22 (7.64)	5.24	< .01[b]	5.73	< .01[c]
Moderate	17.51	17.64 (7.43)	0.06	.95	23.64 (7.09)	2.74	.01[b]	1.89	.07
Severe	41.12	22.21 (6.81)	−10.39	< .01[a]	19.47 (12.73)	−12.43	< .01[a]	−1.09	.29
Profound	24.82	11.92 (9.33)	−0.60	< .01[a]	4.82 (16.34)	−12.68	< .01[a]	−2.58	.02[d]

Note. Superscripts indicate a statistically significant difference between the groups considered. Mean and *SD* (%) of the within-subject standard deviation for Human and Machine WRR across SIT sentences. MDC = minimally detectable change.

[a]*SD* < MDC. [b]*SD* > MDC. [c]Human WRR *SD* < Machine WRR *SD*. [d]Human WRR *SD* > Machine WRR *SD*. For "ALL" $\alpha = .05$. For all other family-wise comparisons (normal, mild, moderate, severe, and profound) $\alpha = .01$ (.05/5).

Figure 6. Variability of Human word recognition rate (WRR) and Machine WRR across the Speech Intelligibility Test (SIT) sentence set (RQ3). Variability (*SD*) across the SIT sentence set for Machine WRR (light blue, left) and Human WRR (dark blue, right) for each participant, grouped by clinician-rated severity. Minimally detectable change (MDC) for each severity group is labeled with a red bar.



found $r = .9$ for IBM Watson used for speakers with aphasia and/or AOS. However, our results aligned with other works that found a similar relationship, for example, Maier et al. (2010) reported $r \geq .8$ for an HMM-based model for speakers with dysglossia and/or dysphonia. Similarly, prior work found that for speakers with either ALS or CP, recognition by Google Speech, IBM Watson, and Microsoft Azure was worse for speakers with "severely distorted" speech (Google WER: 78.21%; IBM WER: 89.08%; Microsoft WER: 78.59%) than for those with "no abnormalities" in their speech (Google WER: 16.11%; IBM WER: 14.89%; Microsoft WER: 23.16%) and the control group (Google WER: 3.95%; IBM WER: 5.26%; Microsoft WER: 6.94%; De Russis & Corno, 2019).

We also found a moderate linear relationship and a stronger quadratic relationship between Machine and Human WRRs. The association was weaker for speakers with poorer intelligibility than for speakers with better intelligibility, indicating a measurement floor for Machine WRR. Such a measurement floor is consistent with prior work that has found low accuracy for severely impaired speakers with ALS or CP (De Russis & Corno, 2019). Human WRR detects signs of dysarthria later than other measures of speech severity, including speaking rate (Rong, Yunusova, & Green, 2015). Therefore, speakers with normal speech intelligibility do not necessarily have typical speech. Machine WRR was not significantly different between speakers with normal and mildly impaired Human WRR (Groups 1–2), indicating that this OTS Machine WRR might be an even poorer detector of early speech decline than Human WRR. Overall, these results suggested that while Machine WRR and Human WRR are correlated, the tested OTS Machine WRR is a weak linear proxy for Human WRR.

(RQ2) Known-Groups Validity: Machine WRR Cannot Detect Mild Dysarthria Without Modification

Human WRR was different between nearly all severity groups, except between mild and moderate and between mild and normal. However, decision trees indicated that Human WRR could be used to successfully classify speakers in all severity groups except the moderate group. Machine WRR differed among most severity groups, except the two least and the two most severe groups. Decision tree analyses supported these results, as the tree that used Machine WRR classified speakers into just three groups: normal, severe, and profound, and performed more poorly than Human WRR (see Table 7). These results suggested that this OTS Machine WRR is not

Table 9. Summary of study findings.

Question	Conclusion	Recommendations
(RQ1) Convergent validity: Is the Off-the-Shelf Automatic Speech Recognition (OTS ASR) system output a valid representation of transcription speech intelligibility?	-There is a moderately strong correlation between Machine word recognition rate (WRR) and Human WRR. -Machine WRR is related, but not identical, to Human WRR. -Machine WRR and Human WRR have a quadratic, nonlinear relationship.	ASR as intelligibility proxy: One should not consider the tested Machine WRR a one-to-one substitute for Human WRR.
(RQ2) Known-groups validity: Is the OTS ASR output a valid proxy for clinician-rated severity groupings?	-Machine WRR has poor performance when discriminating between more mildly impaired groups. -While neither Human WRR nor Machine WRR could differentiate the moderate severity group, Human WRR outperformed Machine WRR for early detection. -When speech samples were first mixed with noise, Machine WRR improved at differentiating mild and normal severity groups.	ASR as a measure of clinician-rated severity: This OTS ASR system may work for coarse stratification, but not for fine-grained stratification, diagnosis, or early identification of speech changes. -With modification (i.e., adding noise to the signal), nASR may be used to differentiate normal and mild severities.
(RQ3) Internal reliability: Is the OTS ASR accuracy reliable across sentences of varying length and content?	-Increased sentence length was associated with decreased Human WRR and Machine WRR. -Machine WRR showed high variability for normal and mild speakers, indicating poor internal reliability. -Internal reliability was higher for Human WRR.	ASR for assessment: Poor internal reliability necessitates using the tested ASR on many samples for each speaker, which is unrealistic and might counter the benefits of ASR as a more time-efficient method than Human WRR. -Weak reliability especially for normal and mild speakers bodes poorly for ASR as a reliable early diagnosis tool. -High variability may render the tested ASR a poor tool for tracking precise speech changes, e.g., in response to medication or behavioral therapy. -Sentence length may impact both Human and Machine WRR.

appropriate for evaluating severity progression or early diagnosis. Prior work relating Google Speech recognition to clinician-rated severity for dysarthric speakers found a moderate relationship between the two (Pearson $r = .69$), which seems to generally align with our results (Tu et al., 2016). However, Machine WRR improved when the audio of normal and mild speakers was mixed with noise before speech recognition, thereby reducing an ASR ceiling effect and suggesting a possible route toward improved clinical applications of ASR (see Table 7). Notably, neither the Human WRR tree nor the Machine WRR tree correctly classified moderately impaired speech. It is possible that speech components that do not affect intelligibility influenced clinician severity ratings for this cohort; however, these findings may cast doubt on five-tiered severity stratification.

Overall, we observed results that corroborate prior work that Human WRR does not differ between mild and normal dysarthria severity groups (Allison et al., 2017; Rong, Yunusova, & Green, 2015; Stipancic et al., 2021). Crucially, while it is well known that Human WRR is an imperfect measure of clinician-rated speech severity, Machine WRR was no better in this study. Furthermore, our results suggested that noise augmentation (nASR) can mitigate a ceiling effect for Machine WRR, potentially enabling Machine WRR detection of early speech impairment. Additionally, thresholds provided in Table 6 demonstrate that Machine WRR has a dynamic range—or range of values over which a change can distinguish severity groups—of about 9.37%–55.80% (a 46.43%-wide interval), much lower than the Human WRR dynamic range 44.97%–94.17% (a 49.2%-wide interval). Although the magnitudes of the dynamic ranges are similar, the Machine WRR range is about 37 percentage points less than the Human WRR range. The similar magnitude of each measure's dynamic range may suggest that one could add 37% to a Machine WRR score to obtain a Human WRR. However, because we found a nonlinear relationship between Machine WRR and Human WRR, this linear transformation would not result in an accurate perceptual intelligibility score. These results further underscored that one should not directly substitute Machine WRR for Human WRR and that this OTS ASR cannot detect mild speech impairment in ALS.

(RQ3) Internal Reliability: Machine WRR Has High Intraspeaker Variability and Degrades on Longer Sentences

Our results indicated that a longer sentence length resulted in lower WRRs in our corpus. Only Human WRR, however, showed an interaction between sentence

length and severity, such that more severe groups had lower Human WRR for longer sentences. However, the floor effect for Machine WRR could be obscuring an effect of sentence length in these more impaired groups. It is notable that even in the normal group, Machine WER declined as a function of sentence length, suggesting that sentence length may be a confounding factor when using OTS ASR to grade severity. Because sentence length impacted both Human and Machine WRR, and earlier work has documented its effects on intelligibility (Allison et al., 2019), it seems likely that sentence length was impacting speaker production rather than the tools used to measure intelligibility. Moreover, we might expect the ASR language model to boost the recognition accuracy of longer sentences, because there would be more context to inform recognition; that we saw an opposite trend further indicated that sentence length was affecting the speaker's production.

While sentence length affected both Human and Machine WRR, intraspeaker variability analyses indicated that such variability had a limited functional impact on Human WRR, especially when compared with Machine WRR. Intraspeaker variability (*SD*) was greater for Machine WRR than for Human WRR for all speakers and for the normal and mild severity groups. In the profoundly impaired group, accuracy was near 0% for Machine WRR; thus, *SD* could not be accurately assessed.

Human WRR intraspeaker variations were within the measurement error, as the *SD* was lower than or equal to the MDC in all groups. Machine WRR intraspeaker variations, however, were outside the measurement error, as the *SD* was higher than the MDC in all groups except for the two most severe groups. The two most severe groups had the highest MDCs as well as a measurement floor for Machine WRR, possibly invalidating the use of *SD*. Wide variation in performance across sentences suggested that ASR performance is inconsistent and should be taken over many sentences to account for expected variation across sentences. Because variation across sentences was not observed for Human WRR scores, the variation seen in Machine WRR was likely due in large part to unreliable ASR performance. However, it may be that sentence length and content (including lexical characteristics like word frequency and neighborhood density) affect speech production in ways that influence ASR performance more than they do Human WRR (Goldwater et al., 2010).

These results do not preclude the use of a subset of sentences to assess with Machine WRR, but they do indicate that such a subset should be constructed carefully. Performances on sentences of different lengths cannot be considered interchangeable, as we found a significant effect of sentence length on both Human and Machine WRR. Furthermore, our findings showed that Machine WRR is imprecise and has large variability across sentences. In terms of Machine WRR applications, these results imply that multiple sentences should be used, and scores taken from varying stimuli should be considered with caution. Poor internal reliability might limit the interpretation of an OTS Machine WRR score for a single sentence, indicating that it is unreliable, or influenced by sentence length and content. OTS Machine WRR, notably, had an *SD* greater than the MDC for the two least impaired groups, indicating that it is unreliable for early stages of ALS dysarthria.

Limitations

For this study, we tested a single OTS ASR system, Google Cloud ASR. Because ASR systems use different underlying frameworks and test sets, we cannot necessarily generalize these findings to other OTS ASR systems.

Additionally, OTS ASR systems are constantly being updated. Therefore, even the performance of the system under consideration could change with respect to dysarthria ratings. Moreover, the development of dysarthria-specific ASR to boost recognition has become an active and promising area of research (Christensen et al., 2012; Green et al., 2021; Gupta et al., 2016; Keshet, 2018). Although such advancements would improve the usability of ASR systems for people with dysarthria, they could negatively impact their utility as diagnostic or treatment tools by reducing the performance differential among speakers with disordered speech.

Because this study tested clinical validations in speakers with ALS, we do not know how this ASR system may perform on other clinical populations. For example, Parkinson's disease is characterized by decreased vocal intensity and increased speaking rate, which have been shown to adversely affected ASR accuracy even in healthy speakers (Goldwater et al., 2010). We found that Machine WRR was not a linear proxy for Human WRR; therefore, its relationships to Human WRR and clinician-rated severity may need to be verified for each population and ASR system. Additionally, because our data set was cross-sectional, these findings cannot be directly applied to progress monitoring applications. Furthermore, our corpus did not enable us to compare the direct effects of linguistic content (e.g., syntactic complexity and word frequency) on either Human or Machine WRR. Additionally, due to limited research into MDC, we used values determined for speaking intelligibility groups rather than perceptual severity groups in particular. Finally, we considered sentence intelligibility only; OTS Machine WRR may perform differently at the word or discourse level, or when compared with alternate measures of speech severity, such as comprehensibility.

Conclusions

We conducted analyses to test the clinical validity of Google Cloud Speech Recognition, an OTS ASR system,

for evaluating dysarthria in people with ALS. Overall, our analyses indicated limited utility for OTS ASR within a clinical context for this population. OTS ASR (Machine WRR) performed well for coarser severity stratification (e.g., normal–mild, moderate, and severe–profound). However, Machine WRR differentiated typical and mildly impaired speakers only when noise was added to the speech signal (nASR), indicating that it is poorly suited for early speech impairment detection. Our analyses revealed shortcomings in the reliability of ASR across sentences, as we found significant intraspeaker variation in ASR performance. Both Human and Machine WRRs degraded with increased sentence length. However, overall intraspeaker variability was higher for Machine WRR than for Human WRR; intraspeaker variability in Machine WRR was especially high for the normal and mildly impaired severity groups. These results suggested that OTS Machine WRR measurements should be taken over multiple sentences, with scores derived from varying sentence sets considered cautiously. When comparing Machine and Human WRRs, we noted a floor effect for the ASR system in which WRR was consistently at 0% across sentences for the profoundly impaired group. Finally, while correlated with Human WRR, OTS Machine WRR did not provide a one-to-one mapping of Human WRR.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders (on Grants R01DC017291 [PIs: Yana Yunusova and Jordan R. Green], K24DC016312 [PI: Jordan R. Green], T32DC000038 [PI: Bertrand Delgutte], and F31DC019016[PI: Sarah E. Gutz].

References

Allison, K. M., Yunusova, Y., Campbell, T. F., Wang, J., Berry, J. D., & Green, J. R. (2017). The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 18*(5–6), 358–366. https://doi.org/10.1080/21678421.2017.1303515

Allison, K. M., Yunusova, Y., & Green, J. R. (2019). Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. *American Journal of Speech-Language Pathology, 28*(1), 96–107. https://doi.org/10.1044/2018_AJSLP-18-0049

Ballard, K. J., Etter, N. M., Shen, S., Monroe, P., & Tand, C. T. (2019). Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *American Journal of Speech-Language Pathology, 28*(2S), 818–834. https://doi.org/10.1044/2018_AJSLP-MSC18-18-0109

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10–11), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Blaney, B., & Hewlett, N. (2007). Dysarthria and Friedreich's ataxia: What can intelligibility assessment tell us? *International Journal of Language & Communication Disorders, 42*(1), 19–37. https://doi.org/10.1080/13682820600690993

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). *Journal of the Neurological Sciences, 169*(1–2), 13–21. https://doi.org/10.1016/S0022-510X(99)00210-5

Christensen, H., Cunningham, S. P., Fox, C. W., Green, P. D., & Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. *Interspeech 2012 Conference Proceedings,* 1776–1779.

Connaghan, K. P., & Patel, R. (2017). The impact of contrastive stress on vowel acoustics and intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research, 60*(1), 38–50. https://doi.org/10.1044/2016_JSLHR-S-15-0291

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*(2), 246–269. https://doi.org/10.1044/jshr.1202.246

De Russis, L., & Corno, F. (2019). On the impact of dysarthric speech on contemporary ASR cloud platforms. *Journal of Reliable Intelligent Environments, 5*(3), 163–172. https://doi.org/10.1007/s40860-019-00085-y

Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D., & Girardi, F. (2017). Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. *IEEE Access, 5,* 22199–22208. https://doi.org/10.1109/ACCESS.2017.2762475

Enderby, P. (1980). Frenchay dysarthria assessment. *British Journal of Disorders of Communication, 15*(3), 165–173. https://doi.org/10.3109/13682828009112541

Ferrier, L. J., Shane, H. C., Ballard, H. F., Carpenter, T., & Benoit, A. (1995). Dysarthric speakers intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication, 11*(3), 165–175. https://doi.org/10.1080/07434619512331277289

Goldsack, J. C., Coravos, A., Bakker, J. P., Bent, B., Dowling, A. V., Fitzer-Attas, C., Godfrey, A., Godino, J. G., Gujar, N., Izmailova, E., Manta, C., Peterson, B., Vandendriessche, B., Wood, W. A., Wang, K. W., & Dunn, J. (2020). Verification, analytical validation, and clinical validation (V3): The foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *npj Digital Medicine, 3*(1), 55. https://doi.org/10.1038/s41746-020-0260-4

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication, 52*(3), 181–200. https://doi.org/10.1016/j.specom.2009.10.001

Google LLC. (2020). *Speech-to-text: Automatic speech recognition, Google Cloud.* Retrieved January 18, 2022, from https://cloud.google.com/speech-to-text

Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M. A., Tobin, J., Brenner, M. P., Nelson, P. C., & Tomanek, K. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases.

Interspeech 2021 Conference Proceedings, 4778–4782. https://doi.org/10.21437/INTERSPEECH.2021-1384

Green, J. R., Yunusova, Y., Kuruvilla, M. S., Wang, J., Pattee, G. L., Synhorst, L., Zinman, L., & Berry, J. D. (2013). Bulbar and speech motor assessment in ALS: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 14*(7–8), 494–500. https://doi.org/10.3109/21678421.2013.817585

Gupta, R., Chaspari, T., Kim, J., Kumar, N., Bone, D., & Narayanan, S. (2016). Pathological speech processing: State-of-the-art, current challenges, and future directions. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6470–6474.

Gutz, S. E., Rowe, H. P., & Green, J. R. (2021). Speaking with a KN95 face mask: ASR performance and speaker compensation. *Interspeech 2021 Conference Proceedings,* 4798–4802.

Gutz, S. E., Wang, J., Yunusova, Y., & Green, J. R. (2019). Early identification of speech changes due to amyotrophic lateral sclerosis using machine classification. *Interspeech 2019 Conference Proceedings,* 604–608. https://doi.org/10.21437/Interspeech.2019-2967

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 51*(3), 562–573. https://doi.org/10.1044/1092-4388(2008/040)

Jacks, A., Haley, K. L., Bishop, G., & Harmon, T. G. (2019). Automated speech recognition in adult stroke survivors: Comparing human and computer transcriptions. *Folia Phoniatrica et Logopaedica, 71*(5–6), 286–296. https://doi.org/10.1159/000499156

Keshet, J. (2018). Automatic speech recognition: A primer for speech-language pathology researchers. *International Journal of Speech-Language Pathology, 20*(6), 599–609. https://doi.org/10.1080/17549507.2018.1510033

King, J. M., Watson, M., & Lof, G. L. (2012). Practice patterns of speech-language pathologists assessing intelligibility of dysarthric speech. *Journal of Medical Speech-Language Pathology, 20*(1), 1–16.

Krishna, G., Tran, C., Yu, J., & Tewfik, A. H. (2019). Speech recognition with no speech or with noisy speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1090–1094). https://doi.org/10.1109/ICASSP.2019.8683453

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology, 5*(1), 96–116. https://doi.org/10.1002/LIO2.354

Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., & Schuster, M. (2010). Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio, Speech, and Music Processing, 2010*(1), Article 24. https://doi.org/10.1186/1687-4722-2010-926951

MathWorks Audio Toolbox Team. (2022). *speech2text*. MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/65266-speech2text

McHenry, M. A., & Laconte, S. (2010). Computer speech recognition as an objective measure of intelligibility. *Journal of Medical Speech-Language Pathology, 18*(4), 99–103.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders, 48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061

Miracle Ear. (2018). *Online hearing test*. https://www.miracle-ear.com/online-hearing-test

Mulholland, M., Lopez, M., Evanini, K., Loukina, A., & Qian, Y. (2016). A comparison of ASR and human errors for transcription of non-native spontaneous speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5855–5859). https://doi.org/10.1109/ICASSP.2016.7472800

Mustafa, M. B., Rosdi, F., Salim, S. S., & Mughal, U. M. (2015). Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Systems with Applications, 42*(8), 3924–3932. https://doi.org/10.1016/j.eswa.2015.01.033

Park, D. S., Zhang, Y., Chiu, C.-C., Chen, Y., Li, B., Chan, W., Le, Q. V., & Wu, Y. (2020). SpecAugment on large scale datasets. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6897–6883.

Riedhammer, K., Stemmer, G., Haderlein, T., Schuster, M., Rosanowski, F., Nöth, E., & Maier, A. (2007). Towards robust automatic evaluation of pathologic telephone speech. *2007 IEEE Workshop on Automatic Speech Recognition and Understanding, Proceedings,* 717–722. https://doi.org/10.1109/asru.2007.4430200

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 1–8. https://doi.org/10.1186/1471-2105-12-77

Rong, P., Yunusova, Y., & Green, J. R. (2015). Speech intelligibility decline in individuals with fast and slow rates of ALS progression. *Interspeech 2015 Conference Proceedings,* 2967–2971.

Rong, P., Yunusova, Y., Wang, J., & Green, J. R. (2015). Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach. *Behavioural Neurology, 2015,* Article 183027. https://doi.org/10.1155/2015/183027

Rowe, H. P., Gutz, S. E., Maffei, M. F., & Green, J. R. (2020). Acoustic-based articulatory phenotypes of amyotrophic lateral sclerosis and Parkinson's disease: Towards an interpretable, hypothesis-driven framework of motor control. *Interspeech 2020 Conference Proceedings,* 4816–4820.

Stipancic, K. L., Palmer, K. M., Rowe, H. P., Yunusova, Y., Berry, J. D., & Green, J. R. (2021). "You say severe, I say mild": Toward an empirical classification of dysarthria severity. *Journal of Speech, Language, and Hearing Research, 64*(12), 4718–4735. https://doi.org/10.1044/2021_JSLHR-21-00197

Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2018). Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 61*(11), 2757–2771. https://doi.org/10.1044/2018_JSLHR-S-17-0366

Stratford, P. W., & Riddle, D. L. (2012). When minimal detectable change exceeds a diagnostic test–based threshold change value for an outcome measure: Resolving the conflict. *Physical Therapy, 92*(10), 1338–1347. https://doi.org/10.2522/ptj.20120002

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research, 55*(4), 1208–1219. https://doi.org/10.1044/1092-4388(2011/11-0048)

Therneau, T., & Atkinson, B. (2019). rpart: Recursive partitioning and regression trees. *R package version, 4,* 1–15. https://CRAN.R-project.org/package=rpart

Tjaden, K., & Liss, J. (1995). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics and Phonetics, 9*(2), 139–154. https://doi.org/10.3109/02699209508985329

Tjaden, K., & Watling, E. (2003). Characteristics of diadochokinesis in multiple sclerosis and Parkinson's disease. *Folia Phoniatrica et Logopaedica, 55*(5), 241–259. https://doi.org/10.1159/000072155

Tomik, B., & Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotrophic Lateral Sclerosis, 11*(1–2), 4–15. https://doi.org/10.3109/17482960802379004

Toth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., Pakaski, M., & Kalman, J. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research, 15*(2), 130–138. https://doi.org/10.2174/1567205014666171121114930

Tu, M., Wisler, A., Berisha, V., & Liss, J. M. (2016). The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance. *The Journal of the Acoustical Society of America, 140*(5), EL416–EL422. https://doi.org/10.1121/1.4967208

Vásquez-Correa, J. C., Orozco-Arroyave, J. R., Bocklet, T., & Nöth, E. (2018). Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *Journal of Communication Disorders, 76,* 21–36. https://doi.org/10.1016/j.jcomdis.2018.08.002

Weismer, G., Jeng, J.-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica, 53*(1), 1–18. https://doi.org/10.1159/000052649

Yorkston, K., Beukelman, D., & Hakel, M. (2007). *Speech Intelligibility Test (SIT) for Windows.* Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital.

Zhang, A. (2017). *Speech recognition* (Version 3.8) [Software]. https://github.com/Uberi/speech_recognition

Ziegler, W. (2002). Task-related factors in oral motor control: Speech and oral diadochokinesis in dysarthria and apraxia of speech. *Brain and Language, 80*(3), 556–575. https://doi.org/10.1006/brln.2001.2614