ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Validating Automatic Diadochokinesis Analysis Methods Across Dysarthria Severity and Syllable Task in Amyotrophic Lateral Sclerosis

Chelsea Tanchip,[a] Diego L. Guarin,[b] Scotia McKinlay,[a] Carolina Barnett,[c] Sanjay Kalra,[d,e] Angela Genge,[f] Lawrence Korngut,[g] Jordan R. Green,[h] James Berry,[i] Lorne Zinman,[j,k] Azadeh Yadollahi,[l,m] Agessandro Abrahao,[j,k] and Yana Yunusova[a,k,l]

[a] Department of Speech-Language Pathology, Rehabilitation Sciences Institute, University of Toronto, Ontario, Canada [b] Department of Biomedical Engineering, Florida Institute of Technology, Melbourne [c] Division of Neurology, Department of Medicine, University of Toronto and University Health Network, Ontario, Canada [d] Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Canada [e] Division of Neurology, University of Alberta, Edmonton, Canada [f] Clinical Research Unit, Montreal Neurological Institute & Hospital, and Department of Neurology and Neurosurgery, McGill University, Québec, Canada [g] Department of Clinical Neurosciences, Hotchkiss Brain Institute, University of Calgary, Alberta, Canada [h] Department of Communication Sciences and Disorders, MGH Institute of Health Professions, Boston, MA [i] Department of Neurology, Massachusetts General Hospital, Boston [j] Division of Neurology, Department of Medicine, Sunnybrook Health Sciences Centre, University of Toronto, Ontario, Canada [k] Hurvitz Brain Sciences Program, Sunnybrook Research Institute, Toronto, Ontario, Canada [l] KITE, Toronto Rehabilitation Institute, University Health Network, Ontario, Canada [m] Institute of Biomedical Engineering, University of Toronto, Ontario, Canada

ARTICLE INFO

Article History:
Received September 21, 2021
Revision received November 19, 2021
Accepted November 20, 2021

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

https://doi.org/10.1044/2021_JSLHR-21-00503

ABSTRACT

Purpose: Oral diadochokinesis (DDK) is a standard dysarthria assessment task. To extract automatic and semi-automatic DDK measurements, numerous DDK analysis algorithms based on acoustic signal processing are available, including amplitude based, spectral based, and hybrid. However, these algorithms have been predominantly validated in individuals with no perceptible to mild dysarthria. The behavior of these algorithms across dysarthria severity is largely unknown. Likewise, these algorithms have not been tested equally for various syllable types. The goal of this study was to evaluate the performance of five common DDK algorithms as a function of dysarthria severity, considering syllable types.
Method: We analyzed 282 DDK recordings of /ba/, /pa/, and /ta/ from 145 participants with amyotrophic lateral sclerosis. Recordings were stratified into mild, moderate, or severe dysarthria groups based on individual performance on the Speech Intelligibility Test. Analysis included manual and automatic estimation of the number of syllables, DDK rate, and cycle-to-cycle temporal variability (cTV). Validation metrics included Bland–Altman mixed-effects limits of agreement between manual and automatic syllable counts, recall and precision between manual and automatic syllable boundary detection, and Kendall's tau-b correlations between manual and algorithm-detected DDK rate and cTV.
Results: The amplitude-based algorithm (absolute energy) yielded the strongest correlations with manual analysis across all severity groups for DDK rate (τ_b = 0.7–0.84) and cTV (τ_b = 0.7–0.84) and the narrowest limits of agreement (–5.92 to 7.12 syllable difference). Moreover, this algorithm also provided the highest mean recall and precision across severity groups for /ba/ and /pa/, but with significantly more variation for/ta/.
Conclusions: Algorithms based on signal energy analysis appeared to be the most robust for DDK analysis across dysarthria severity and syllable types; however, it remains prone to error against severe dysarthria and alveolar syllable context. Further development is needed to address this important issue.

Correspondence to Yana Yunusova: yana.yunusova@utoronto.ca. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease that affects the upper and lower motor neurons in the brain, brainstem, and spinal cord (Kiernan et al., 2011). One of the most devastating consequences of the brainstem involvement in ALS is bulbar weakness resulting in the deterioration of speech function. Bulbar weakness commonly results in various degrees of spastic and flaccid dysarthria (Yorkston et al., 1995, 2007) characterized by reduced speaking rate, consonant imprecision, hypernasality, and harsh/strained and strangled or breathy voice quality. Dysarthria affects speech intelligibility and can eventually lead to the complete loss of oral communication (Chen, 2005; Hillel, 1999). Because bulbar dysfunction in ALS is associated with worsened prognosis, its assessment requires validated measures, sensitive to the various stages of bulbar disease (Yunusova et al., 2019).

Oral diadochokinesis (DDK) is a standard clinical dysarthria assessment task. DDK entails the rapid repetition of a single syllable (e.g., /pa/, also known as sequential motion rate [SMR]) or a stream of alternating syllables (e.g., /pa-ta-ka/, also known as alternating motion rate [AMR]). The task is linguistically and cognitively simple yet physically demanding, and DDK metrics showed great potential utility for detection and progression tracking of bulbar dysfunction in ALS (Rong et al., 2016). DDK performance is typically evaluated based on rate, calculated as the number of syllables per unit time. DDK rate has been used to detect early declines in speech motor performance and to measure the severity of bulbar dysfunction in ALS (Rong et al., 2016). This measure has also been used to stratify patients into slow and fast bulbar disease progressors (Rong et al., 2015). In addition to the basic estimate of rate, DDK performance has been quantified in terms of the variability of syllable durations, captured by metrics such as cycle-to-cycle temporal variability (i.e., cycle-to-cycle temporal variability [cTV], or average temporal difference between DDK cycles) or components of the syllables. Previous research found that variability measures were able to predict declines in bulbar function earlier than rate in ALS (Rong, 2020).

The "gold standard" for objective DDK assessment is manual acoustic analysis, where a clinician segments an acoustic recording to obtain the number and duration of syllables as well as syllable boundary locations. Manual acoustic analysis is, however, laborious and time-consuming (Kent, 1996), necessitating the development and implementation of automatic tools based on acoustic signal processing methods.

One traditional method for syllable segmentation involves using the speech signal's energy envelope based on its amplitude or spectral properties. An energy-based algorithm based on amplitude has been implemented in the commercially available Diadochokinetic Rate Analysis tool part of the Motor Speech Profile tool (KayPENTAX). The tool calculates the speech signal energy envelope and applies a threshold to identify syllable endpoints, which are manually adjusted (Wang et al., 2009). Expanding on this basic method, Novotny et al. (2014) developed a fully automated DDK algorithm, a multistep procedure combining peak elimination and spectral energy thresholding. Using advanced statistical techniques, this algorithm detects certain landmarks within the syllable, including the consonant burst, vowel onset, and occlusion (vowel offset). When validated on individuals with Parkinson's disease (PD) with no or mild dysarthria, the algorithm achieved nearly 80% accuracy in detecting syllable onsets and offsets. Later, the authors validated their algorithm on 36 AMR recordings from 18 individuals with ALS whose ALS Functional Rating Score–Revised (ALSFRS-R) speech subscores ranged from very severe (1 point) to mild dysarthria (4 points; mean subscore = 2.9), but the number of participants in each category was not specified (Novotny et al., 2020). Within a 10-ms tolerance window, their algorithm achieved 74.4% detection accuracy, surpassing the performance of a deep learning model used as a comparator (Rozenstoks et al., 2020).

Neurospeech's (Orozco-Arroyave et al., 2018) DDK assessment module also used the spectral energy to perform DDK segmentation. Unlike Novotny et al.'s (2014) algorithm, Neurospeech extracts the pitch contour from the signal via Praat's voice activity detector (VAD) function (Boersma & Weenink, 2001), differentiating speech intervals from silence. To the best of our knowledge, the VAD has not been formally evaluated in terms of its DDK segmentation accuracy but has been able to discriminate PD from healthy controls with over 70% accuracy for AMR tasks and over 60% for individual SMR tasks using various temporal and spectral features (e.g., DDK rate, F0 statistics; Vásquez-Correa et al., 2019).

The Diadochokinetic Tasks Analysis (DTA) tool applies Hilbert transformation—a spectral analysis technique—to identify envelope peaks instead of endpoints (Smékal et al., 2013). To the best of our knowledge, only one published study (Wang et al., 2019) has assessed the performance of the DTA, while comparing it to a deep learning–based DDK model (DeepDDK) on 17 individuals (306 recordings) from an unspecified clinical population using several SMR tasks. A cumulative distribution of syllable count difference was presented on a range of 0–5, representing the syllable count difference (number of syllables) between the algorithm and manual methods. A score of 0 would mean there was no difference in the syllables counted (perfect algorithm performance) and 5 would mean a total lack of agreement. DTA achieved a count difference of zero for 3.70% of recordings and a count difference of 5 for 51.85% of recordings.

Novotny et al. (2020) compared their 2014 algorithm with one developed by Rong (2020), another peak-detecting algorithm that applied Hilbert transformation to the signal envelope. Unlike the DTA, Rong's algorithm

was designed for ALS and featured a more sophisticated signal-filtering technique that simulated cochlear processing based on studies of the auditory perception of speech. Rong's algorithm was originally designed to be semi-automatic, requiring manual readjustment of peaks that exceeded the 95% confidence interval of height and/or distance distributions. Novotny et al.'s (2020) study featured a fully automatic version of Rong's algorithm, replacing the manual adjustment option for spurious peaks with a modified peak selection procedure, where thresholds were set based on peak prominence and the final set of peaks was selected through k-means clustering. According to the results of Novotny et al., Rong's modified algorithm achieved 45.8% detection accuracy with a 10-ms tolerance.

A notable gap in previous DDK algorithm validation studies concerns the lack of recordings of individuals with notable (e.g., moderate to severe) dysarthria. Dysarthria impairs the function of speech articulators (e.g., lips, tongue, velum) and respiratory/laryngeal control as severity increases, which can yield irregularities in the acoustic signal and cause significant problems for automatic acoustic analysis. Dysarthria in ALS is known to reduce consonant precision, producing continuous voicing energy even during voiceless intervals (Ackermann & Ziegler, 1991) and distorted spectral properties (e.g., Mel-Spectrum Fourier Coefficients; Kent et al., 1999). In the analysis of DDK, these effects may yield less precisely detected boundaries and more commonly missed or falsely detected syllables. Furthermore, speakers with dysarthria tend to struggle more with certain types of syllables due to the different articulatory mechanisms involved (e.g., alveolar and voiceless /ta/ becomes particularly difficult to articulate and perceive compared to the labial and voiced /ba/; Ackermann & Ziegler, 1991; Kent et al., 1991). The effect of dysarthria severity on algorithm performance may, therefore, vary between different syllable types. With respect to these shortcomings, the behavior of current DDK algorithms across dysarthria severity and syllable types remains largely unexplored.

This study had two aims. The primary aim was to examine the concurrent validity of five available DDK algorithms as a function of dysarthria severity (mild to severe) by examining correlations with clinical DDK metrics (e.g., syllable rate and variability) as well as changes in boundary detection accuracy and total syllable count differences relative to the manual "gold standard." The secondary aim was to examine potential confounding effects of syllable type (e.g., voiced /ba/ vs. voiceless /pa/; labial /pa/ versus alveolar /ta/) on the algorithmic methods. The following questions were addressed: (a) Which algorithm performs the best against the manual gold standard across dysarthria severity (and is therefore the most valid)? and (b) Does syllable type influence algorithm performance? We hypothesized, based on clinical studies of ALS, that all algorithms show declines in performance due to dysarthria severity and show variable performance across different syllable types.

Method

Participants

DDK recordings (817 total) were selected from three previous longitudinal prospective studies (Green et al., 2013; Kalra et al., 2020; Rong et al., 2015) that administered DDK tasks as part of a broader assessment of bulbar dysfunction. The primary objective of these studies was to identify speech or oro-motor biomarkers for early detection and disease progression. Recordings significantly corrupted by background chatter or noise as well as those without bulbar disease severity scores as determined by the Sentence Intelligibility Test (SIT; Yorkston et al., 2007) were excluded. Furthermore, session recordings belonging to individuals at the severe range of the bulbar disease continuum (as described below) were counted. The recordings at the mild disease stage, which were much larger in number, were selected based on preselected intelligible speaking rate criteria, and a subset of those participants were extracted randomly per stratification group to produce comparable sample sizes per group.

The final data set consisted of 282 DDK recordings from 145 individuals with ALS aged 41–81 years (M = 57.8, SD = 8.0), including 88 men and 57 women. Participants were diagnosed with ALS based on the El Escorial criteria from the World Federation of Neurology (Brooks et al., 2000). Thirty-three participants had the site of onset in the bulbar region and 139 in the spinal region. Some individuals (n = 76) contributed more than one recording (M = 1.9, SD = 1.3). The time between recordings was spaced between 2 weeks and 12 months (M = 4.4 months, SD = 3.6). The participants' mean disease duration was 34.94 months (SD = 31.69).

Stratification by Bulbar Disease Severity

In this study, bulbar disease severity was determined by the SIT (Yorkston et al., 1995), which measured speaking rate and speech intelligibility. In the SIT, participants read 11 randomly generated sentences containing five to 15 words. One naïve adult listener who was unfamiliar with the stimuli and speech profile of the participants orthographically transcribed the sentences (see the work of Stipancic et al., 2018). Speaking rate was measured as the average number of words produced per minute (words per minute [WPM]), and intelligibility was measured as the proportion of correctly transcribed words to the total amount of words across sentences.

Table 1. The summary of participant and session characteristics by severity group.

Measure		Mild	Moderate	Severe
Recording sessions		100	100	82
Participants		60	45	40
Sex	Male	35	24	29
	Female	25	21	11
Mean age, years (*SD*)		57.45 (10.18)	58.44 (9.60)	59.60 (10.34)
Onset	Bulbar	11	14	8
	Spinal	57	38	44
Mean disease duration from onset, months (*SD*)		39.69 (18.81)	29.18 (40.50)	35.69 (31.88)
Mean speech intelligibility, % (*SD*)		98.48 (2.06)	98.14 (4.36)	96.20 (28.29)
Mean speaking rate, WPM (*SD*)		196.70 (25.39)	133.90 (14.09)	85.49 (21.29)
Mean intelligible rate, WPM (*SD*)		193.34 (26.20)	128.64 (13.40)	65.06 (29.11)

Participants in this study were stratified into severity groups based on intelligible speaking rate (in WPM) calculated as (*intelligibility / 100*) × *rate*—a communication efficiency metric that combines two measures of speech disorder severity (Yorkston & Beukelman, 1981). The stratification scheme was built from visualizing the spread of intelligible rate via histogram and consulting previous ALS literature on speaking rate and speech intelligibility (Barnett et al., 2020; Shellikeri et al., 2016; Stipancic et al., 2018; Wang et al., 2016). The stratification scheme became as follows: (a) *mild* = intelligible rate of ≥ 150 WPM; (b) *moderate* = intelligible rate between 100 and 149 WPM; (c) *severe* = intelligible rate of < 100 WPM. Table 1 summarizes the demographic and clinical information for each stratification group.

Speech Samples

Participants repeated the syllables /ba/, /pa/, or /ta/ as fast as possible on one breath. The syllables were selected to allow the voice—voiceless (ba vs. pa) and bilabial—lingual (pa vs. ta) contrast evaluations. Each task was repeated up to 2 times; only the second repetition was used for analysis. The type of tasks performed per subject varied, leading to an uneven sample. Overall, there were 103 recordings of /ba/, 68 of /pa/, and 111 of /ta/. Due to the differing protocols of the original studies, participants did not perform the same set of tasks; some performed only one of the three tasks; others repeated both /ba/ and /ta/, or both /pa/ and /ta/. Thirty-eight participants performed only /ba/, 45 only /pa/, 40 only /ta/, 21 /ba/ and /ta/, and 12 /pa/ and /ta/. Table 2 lists the distribution of recordings across syllable types and severity groups.

The speech samples were recorded in a clinical laboratory setting with high-quality recording equipment. The first and second studies from which the data were obtained used a Marantz PMD660 compact flash recorder with an accompanying Countryman E6 omnidirectional microphone. The third study used an Olympus WS-853 recorder with an accompanying ME52W unidirectional microphone. Microphones were distanced 8–10 cm from the mouth. Sampling rates varied across the data set: 22.1 kHz (129 recordings), 32 kHz (eight recordings), 44.1 kHz (94 recordings), and 48 kHz (85 recordings), with 16-bit resolution.

Manual Segmentation

The first author (Rater 1) and two trained research assistants (graduate speech-language pathology students; Raters 2 and 3) manually annotated all 282 recordings using Praat. The files were randomly split for each annotator, but Rater 1 and Rater 2 jointly annotated 11% of the data set and Rater 1 and Rater 3 jointly annotated a separate 10% of the data set to calculate interrater reliability. The first 10 s of each recording were annotated; recordings of shorter length were fully annotated. The 10-s marker was determined from the onset of speech. Syllable start and end points were defined as the initial consonant burst and vowel occlusion/offset, respectively, using Novotny et al.'s (2014) labeling criteria as a reference. Vowel onset locations were also obtained for comparison to the nuclei/peak-detecting algorithms. The frequency domain was primarily examined. The initial burst was characterized as the abrupt noise over the entire frequency range on the spectrogram. The vowel onset was marked by the first appearance of the second formant frequency. The vowel occlusion was characterized as the last glottal pulse on the spectrogram.

Table 2. Number of recordings per syllable type and severity group.

Syllable	Mild	Moderate	Severe
/ba/	44	40	19
/pa/	28	25	15
/ta/	28	35	48

Note. WPM = words per minute.

Algorithmic Segmentation: Description of Algorithms

Five algorithms used for automated segmentation of DDK data were evaluated in this study. All but Energy were selected based on their prior usage in the clinical assessment of neurodegenerative disorders (for practice and/or research) and use of different signal processing methods. The Energy algorithm was a slightly modified version of Colonna et al.'s (2015) open-source bioacoustics signal segmentation algorithm on MATLAB. This algorithm, originally used to analyze animal vocalizations, calculated the absolute (sum-of-squares) energy envelope of a signal and applied a static threshold (0.3) for segmentation. In this study, the threshold was modified to be the moving average of the signal, similar to Novotny's method (2014). The computational framework underlying each algorithm is summarized in Table 3 and visualized in Figure 1.

Algorithmic Segmentation: Performance Evaluation

To determine the concurrent validity of the algorithms of interest, we compared the detection of the five relevant measures extracted algorithmically and manually. The first three measures are common algorithm performance evaluation metrics (see the works of Novotny et al., 2020; Rozenstoks et al., 2020; Wang et al., 2019). They include the following:

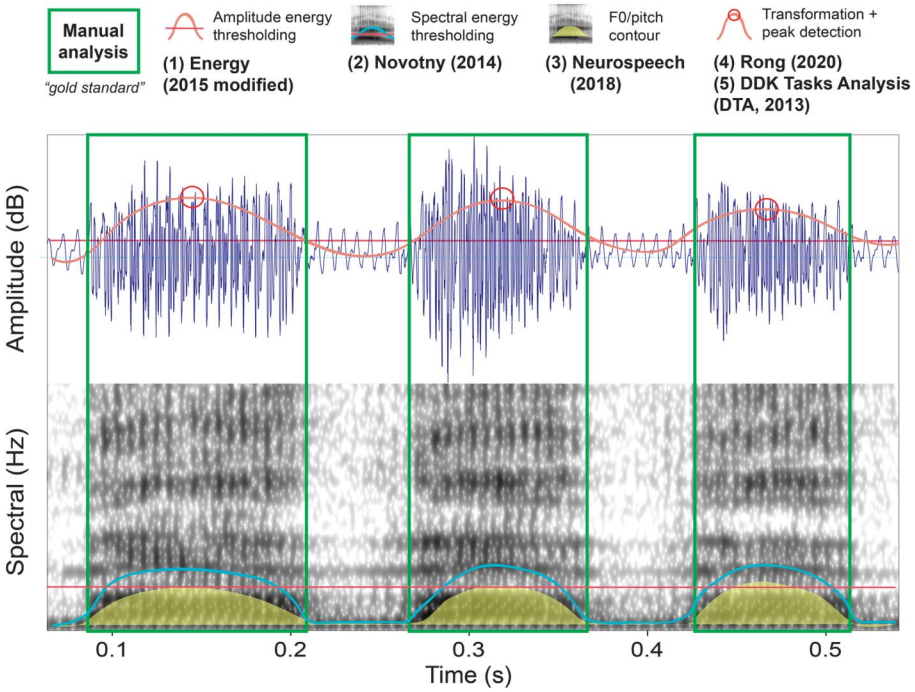
1. Total syllable count agreement was defined as the absolute difference in the number of syllables uttered in the DDK sequence.

Table 3. Summary of computational methods underlying DDK segmentation algorithms.

Method	Preprocessing	Algorithm segmentation summary
Energy *Simple energy envelope estimation*	1) DC offset removal 2) Scaling by maximum value 3) Zero-padding 4) Low-pass FIR with 200 Hz cut-off frequency	1) Calculated sum of squares (absolute energy) from 20 ms frames to create an envelope; 2) Created a threshold by taking the moving average over the signal with a sliding 20 ms window. Points crossing the threshold were considered syllable boundaries
Novotny *Multistep procedure involving peak elimination and spectral energy envelope estimation*	1) Low-pass FIR with a 1000 Hz cut-off frequency 2) DC offset removal 3) Scaling by maximum value 4) Downsampling to 20 kHz	1) Squared and smoothed the signal with a moving average filter to estimate the position of each syllabic nucleus (peak); 2) Extracted the signal's spectrogram and applied a weighted mean spectral energy threshold to eliminate detected spurious peaks; 3) Generated two spectral energy envelopes and their centroid positions to eliminate remaining false detections; 3) Detected bursts (syllable onset) with a Bayesian step-change detector (BSCD); 4) Detected vowel offsets with an inverted polynomial energy threshold
Neurospeech *F0 (pitch) contour estimation*	1) Set frequency range for pitch estimation to 75–600 Hz	F0 values were estimated from windows of 20 ms with 10 ms time step using the external PRAAT VAD
DTA *Peak identification with static threshold*	1) DC offset removal 2) Scaling by maximum value	1) Applied Hilbert transformation to signal to create an envelope 2) Obtained first set of peaks with the minimum distance threshold set to *sampling rate/20* and minimum height threshold to the *envelope mean amplitude/15*. 3) Obtained second set of peaks with the minimum distance threshold set to *sampling rate/25* and same height threshold as in (2). 4) The standard deviations of inter-peak differences were calculated. The set of differences with a smaller standard deviation was chosen as the final set of peaks.
Rong *Peak identification supplemented with acoustic filtering designed to simulate human hearing*	1) DC offset removal 2) Bandpass filter applied by splitting acoustic waveform into 64 frequency bands 3) low-pass filter the frequency envelope at 10 Hz	1) Calculated sum of absolute values of Hilbert-transformed signal over 64 frequency bands to create an envelope; 2) Calculated and eliminated envelope peaks if their height and distance exceeded the 95% confidence interval of the overall peak heights and inter-peak distances

Note. *Energy*: modified version of Colonna et al.'s (2015) algorithm; *Novotny*: Novotny et al.'s (2014) algorithm, obtained from the authors with permission to use in this study; *Neurospeech*: Orozco-Arroyave et al.'s (2018) software; *DTA*: Diadochokinetic Tasks Analysis, Smékal et al.'s (2013) software; *Rong*: Rong (2020)'s algorithm, obtained from the author with permission to use in this study. DC = direct current voltage; FIR = finite impulse response filter; PRAAT VAD = Voice Activity Detector module in the acoustic software PRAAT.

Figure 1. Sample segmentation of a speech signal produced by a patient with ALS and mild dysarthria, demonstrating the principles of analyses used in the five selected algorithms. ALS = amyotrophic lateral sclerosis; DDK = diadochokinesis; DTA = Diadochokinetic Tasks Analysis.



2. Recall (of boundary detection, also known as sensitivity) referred to the proportion of true syllable boundaries that are correctly identified relative to the ground truth (manual segmentation). A high recall value indicated that a large amount of syllable boundaries was correctly identified by an algorithm.
3. Precision (of boundary detection, also known as positive predictive value) referred to the proportion of true syllable boundaries that were detected by the algorithm relative to the total boundaries detected. A high precision value signified that more true boundaries were detected than false ones.

To enhance clinical interpretation of the results, clinical measures such as

4. DDK rate, defined as the number of syllables divided by the duration of the DDK sequence, and
5. TV, defined as the mean absolute difference in duration between consecutive cycles of the acoustic envelope, were also calculated.

Statistical Analyses

The interrater reliability of manual measurements was determined by calculating the intraclass correlation coefficient (ICC) between the syllable count data obtained across the two sets of raters.

Syllable count distributions per algorithm were reviewed graphically and using descriptive statistics (e.g., mean, standard deviation). Outliers were identified as points at least 2.5 × interquartile range of the data and removed (Ben-Gal, 2010). The effects of dysarthria severity and syllable type on the total syllable count agreement between the manual versus each algorithmic method were evaluated via the Bland–Altman method, a well-established technique for measuring agreement between two raters or systems. A variation of this method based on a mixed-effects analysis, as described in the work of Parker et al. (2016), was used to determine the limits of agreement (LoA), while accounting for potential covariates. To achieve this, two regression models were created, where the first model calculated a crude (unadjusted) mean bias (difference) estimate and the second model (mixed) calculated the same metric, adjusted for severity and syllable type (as fixed effects). To account for the uneven sample, participants were modeled as a random effect. A regression model with an interaction term between severity and syllable types was also evaluated. The LoA were calculated as $\textit{mean bias} \pm 2 \times \textit{SD}$. Model assumptions were assessed using plots of the (a) standardized residuals against fitted values, (b) Q-Q plots of residuals, and (c) Q-Q plots of the random effect predictions. Violation of assumptions (e.g., normality and homoscedasticity of residuals, normality of within participant effects) resulted in the log-transformation of data. The

log-transformation was required for the syllable count data obtained with the Neurospeech algorithm only. As recommended by Parker et al. (2016), both original and transformed statistics were reported for this method. The results were evaluated via the Bland–Altman plots showing the groups stratified by severity and syllable types separately, and regression coefficients for each variable were reported as effect estimates.

Recall and precision were obtained by comparing syllable boundary distances via graphical examination and Python script using tolerance windows of 30 and 60 ms between algorithmically and manually determined boundaries (see the work of Wang et al., 2019). For algorithms that only produced peak locations instead of end points (i.e., DTA and Rong), the boundary comparison was done with respect to the manually determined vowel onset, chosen to approximate the point of maximal lip opening (Rong, 2020). If multiple boundaries belonging to different syllables were detected within a manually determined syllabic interval, the syllable with the boundary closest to the onset (or vowel) was counted as the correct and the others as errors. Recall and precision values ≥ 0.8 were considered as strong (Saito & Rehmsmeier, 2015).

Concurrent validity of the automated tools was established for DDK rate and cTV measures using the Spearman rho correlations and their adjusted p values between manually and automatically detected DDK rate and cTV. Correlations of .7 or higher were considered strong (Akoglu, 2018).

All analyses were conducted on R (RStudio Team, 2020). For the Bland–Altman analysis, an open-source R script (with the *nlme* package) made available by Parker et al. (2016) was used and customized to fit the data in this study.

Results

Interrater Reliability

Ten thousand four hundred nineteen syllables were manually identified and segmented across 282 recordings (4,427 /ba/, 3,219 /ta/, and 2,773 /pa/). The ICC obtained between syllable counts was 0.99 (95% CI [0.97, 0.99]) between both Raters 1 and 2 as well as between Raters 2 and 3, indicating excellent reliability for the manual procedure.

Syllable Counts Across Severity Groups and Syllables for Each Measurement Method

Upon examination of outliers, four points each were removed from syllable counts estimated by Energy and Novotny; 2 points were removed from Neurospeech, 10 points, including two extremely deviant points (> 200-syllable difference between manual and algorithm), were removed from DTA; lastly, 3 points were removed from Rong. After applying these changes, all model assumptions were met. Table 4 shows the mean and standard deviation of manually and algorithmically obtained syllable counts per each recording—with outliers removed—grouped by severity level and syllable type.

Syllable Count Agreement: Severity and Syllable Analyses

Figures 2 and 3 show the Bland–Altman plots of overall agreement between hand-measured and algorithmic syllable detection, where groups—by severity (see Figure 2) and syllable type (see Figure 3)—are presented in different colors. A solid line represents the mean bias, while the dashed-dotted lines represent the mixed 95% LoA.

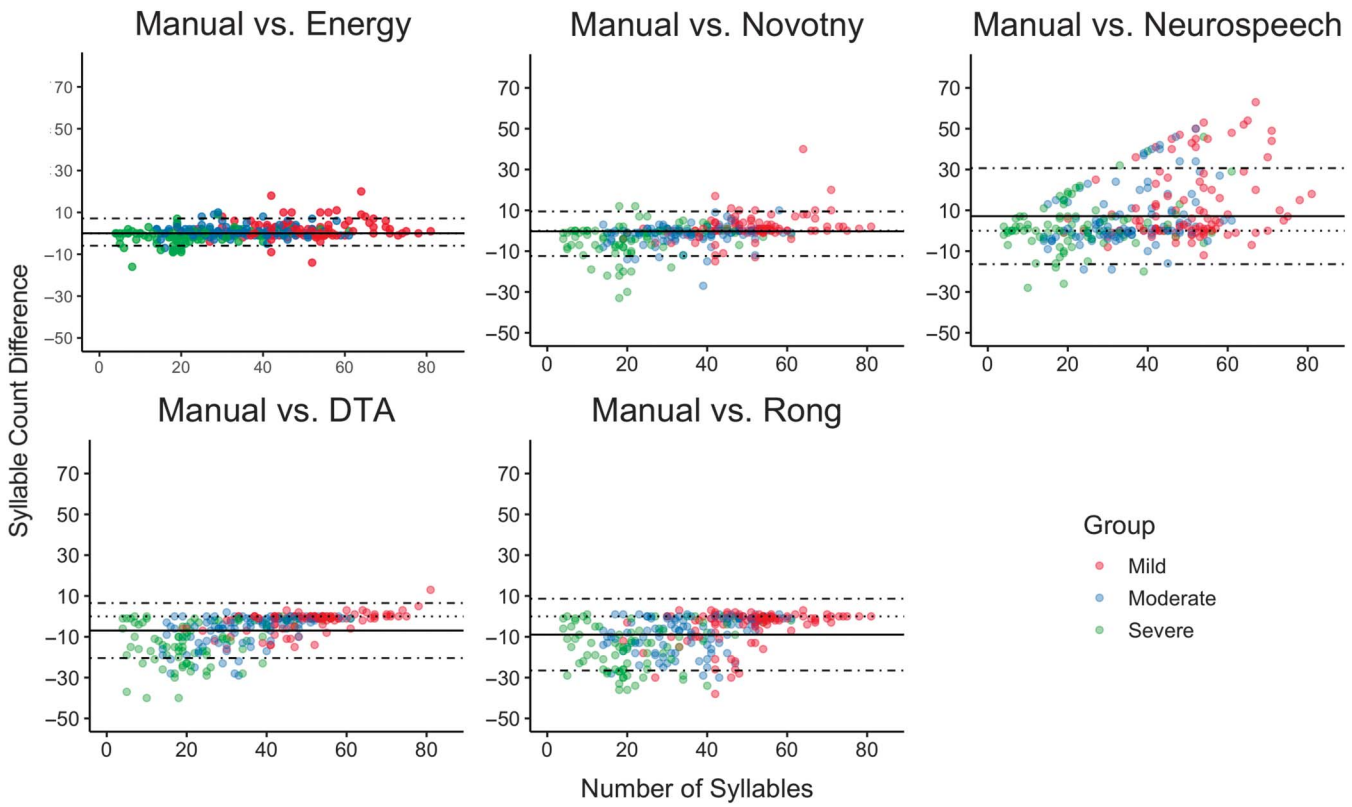
Table 5 shows the results of the regression model for the independent severity and syllable types (fixed) effects. The model LoAs, accounting for both effects, revealed that Energy showed the narrowest LoA, while Rong yielded the widest LoA. Severity had a significant effect on syllable count agreement for all algorithms ($p < .001$). All five algorithms significantly overestimated syllable count in the severe stage of dysarthria as compared to the mild stage, as indicated by the negative differences between manual and algorithmic counts. All algorithms, except Energy, also significantly overestimated syllable counts in the moderate stage of dysarthria severity relative to the mild stage.

Table 4. Summary statistics (mean, *SD*) for syllable counts by severity groups and syllable type per each recording.

Method	Severity			Syllable		
	Mild	Moderate	Severe	/ba/	/pa/	/ta/
Manual	50.37 (12.44)	35.08 (11.70)	22.85 (13.54)	42.98 (16.22)	40.78 (15.61)	29 (14.61)
Energy	48.62 (12.9)	34.14 (11.94)	24.32 (13.11)	42.59 (15.3)	40.18 (14.67)	28.39 (14)
Novotny	48.89 (12.07)	36.86 (11.7)	27.39 (14.24)	42.94 (13.93)	43.29 (13.89)	31.12 (14.53)
Neurospeech	37.78 (19.08)	28.78 (14.87)	20.7 (15.44)	23.14 (18.78)	39.32 (15.94)	29.69 (15.6)
DTA	59.49 (56.63)	41.59 (10.38)	35.59 (14.36)	53.45 (57.06)	45.75 (11.9)	39.73 (14.22)
Rong	55.04 (12.41)	43.58 (12.37)	37.38 (15.5)	53.98 (13.32)	47.22 (12.46)	37.44 (13.95)

Note. DTA = Diadochokinetic Tasks Analysis.

Figure 2. Bland–Altman plots showing the raw/untransformed paired differences between the manual and algorithmic syllable counts for five algorithms, colored by severity group. DTA = Diadochokinetic Tasks Analysis.



The effect of syllable type on syllable count agreement was also significant ($p < .01$). Among algorithms, only Novotny did not show significant differences in the counts between tasks. The lingual syllable /ta/ was more challenging to count for all remaining algorithms; Energy underestimated counts of /ta/, while the rest of the algorithms overestimated them. The performance of Neurospeech was affected both by severity and task, with a significant interaction effect found between these two factors ($p < .001$), suggesting that severity affected algorithm performance differently by task.

Boundary Detection Accuracy

Figures 4 and 5 show the mean recall (top row) and precision (bottom row) for each algorithm across severity and syllable type, respectively, for thresholds of 30 and 60 ms (light and dark gray bars, respectively). Figure 4 demonstrated that most of the algorithms yielded a recall and precision below 0.8 for the 30-ms window across the severity range, except Novotny. For a 60-ms window, Energy, Novotny, and, to some extent, Rong produced good results for mild and moderate severity groups. Inspecting results by syllable type, Energy, and Novotny achieved sufficient recall and precision for a 60-ms window, with somewhat lower values for /ta/ as compared to /ba/ and /pa/ for Energy.

Correlations Between Algorithm-Based and Manual DDK Rate and cTV Across Severity and Syllable Type

Table 6 lists the τ_b correlations between manually and algorithmically determined DDK rates and cTV values for each severity group. Overall, Energy performed the best, yielding correlations over .7 with manual analysis.

Table 7 lists the individual τ_b correlations (all reached statistical significance as well, $p < .01$) between manually and algorithmically determined DDK rates and cTV values by syllable type. Overall, Energy yielded the most consistent and strongest correlations with the manual analysis across tasks, consistently producing correlations over .7.

Discussion

This study investigated the validity of algorithmic DDK tools in segmenting increasingly dysarthric speech and various syllable types. As anticipated, the performance of all algorithms declined as dysarthria progressed in

Figure 3. Bland–Altman plots showing the raw/untransformed paired differences between the manual and algorithmic syllable counts for five algorithms, colored by syllable type. DTA = Diadochokinetic Tasks Analysis.

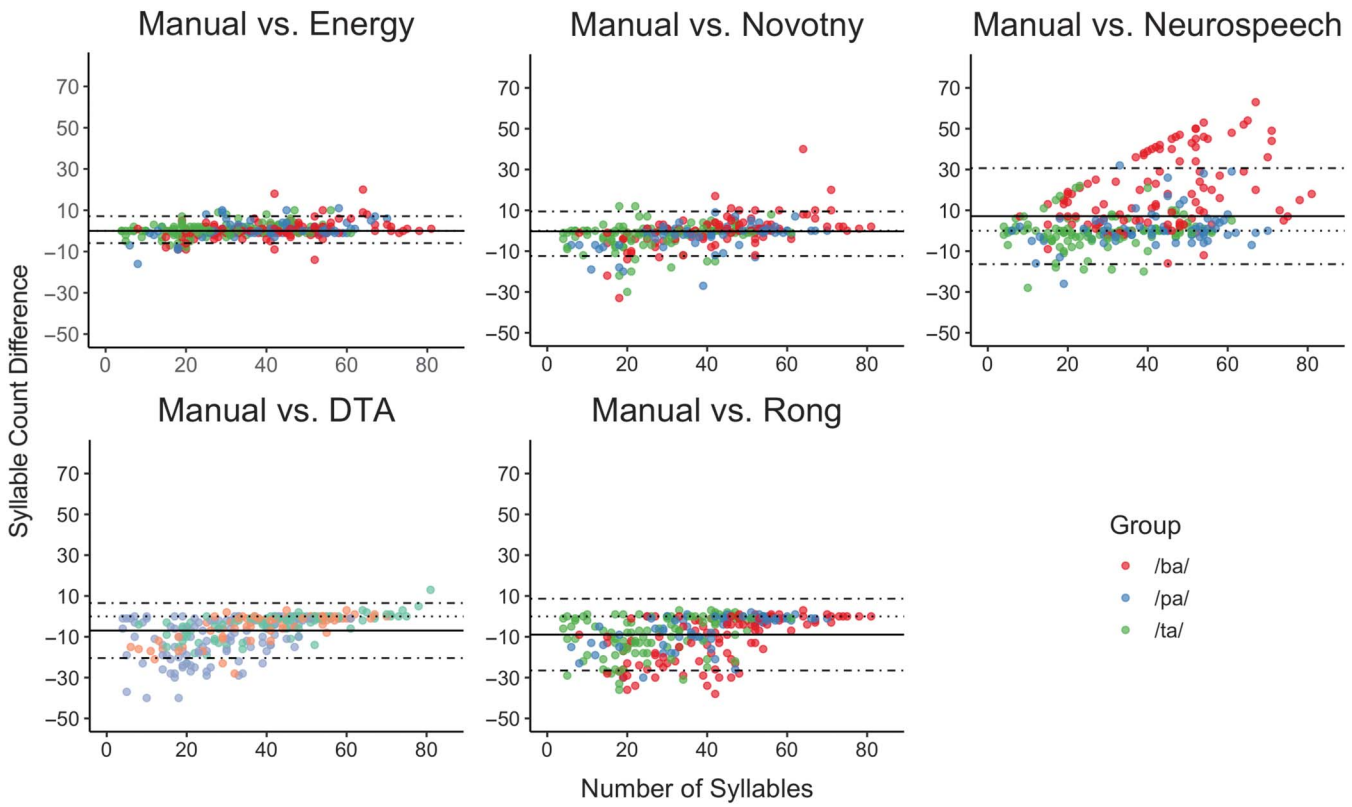


Table 5. Summary statistics of the regression model in the Bland–Altman analysis for the total syllable count agreement between the manual versus each algorithmic methods, with dysarthria severity and syllable type as fixed effects.

Method	Mean bias (95% LoA)	Within-participant *SD*	Effect estimates after adjustment (baseline = severity – mild, syllable – /ba/)
Energy	0.2 (–5.92, 7.12)	0.64	(Intercept: 0.27) SEVERITY: Moderate: –0.12, Severe: –0.87*** SYLLABLE: /pa/: 0.3, /ta/: 0.33**
Novotny	–1.47 (–12.4, 9.46)	4.35	(Intercept: 1.41) SEVERITY: Moderate: –2.63**, Severe: –5.06*** SYLLABLE: /pa/: –1.29, /ta/: –0.33
Neurospeech	7.54 (–16.72, 31.79)	8.94	(Intercept: 23.44) SEVERITY: Moderate: –5.12*, Severe: –7.86*** SYLLABLE: /pa/: –17.57***, /ta/: –17.97***
Neurospeech (log-transformed)	0.81 (–2.08, 3.7)	1.07	(Intercept: 2.45) SEVERITY: Moderate: –0.42*, Severe: –0.75*** SYLLABLE: /pa/: –1.89***, /ta/: –1.87***
DTA	–6.86 (–20.42, 6.53)	4.76	(Intercept: –1.36) SEVERITY: Moderate: –2.83**, Severe: –8.04*** SYLLABLE: /pa/: –0.05, /ta/: –5.61***
Rong	–8.94 (–26.53, 8.66)	6.41	(Intercept: –9.04) SEVERITY: Moderate: –2.72**, Severe: –10.24*** SYLLABLE: /pa/: 6.64***, /ta/: 5.53**

Note. LoA = limits of agreement; DTA = Diadochokinetic Tasks Analysis.
*$p < .05$. **$p < .01$. ***$p < .001$.

Figure 4. Mean recall and precision expressed as proportions for each algorithm for each dysarthria severity groups. DTA = Diadochokinetic Tasks Analysis.

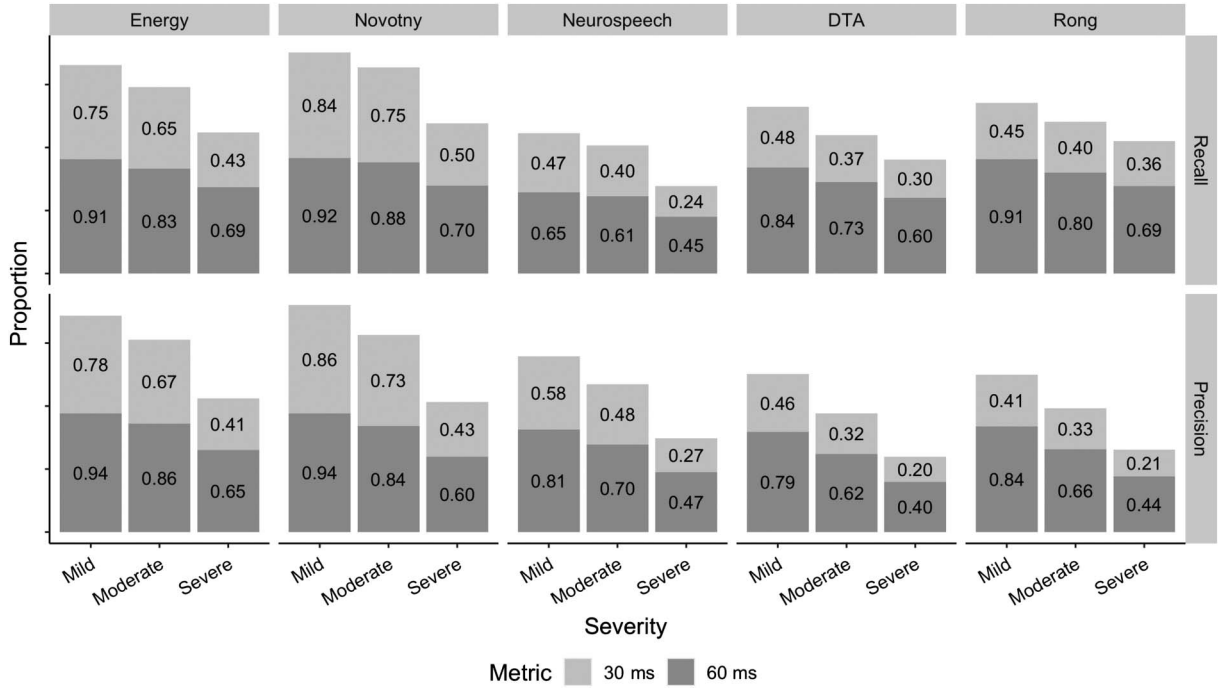


Figure 5. Mean recall and precision expressed as proportions for each algorithm between syllable types. DTA = Diadochokinetic Tasks Analysis.

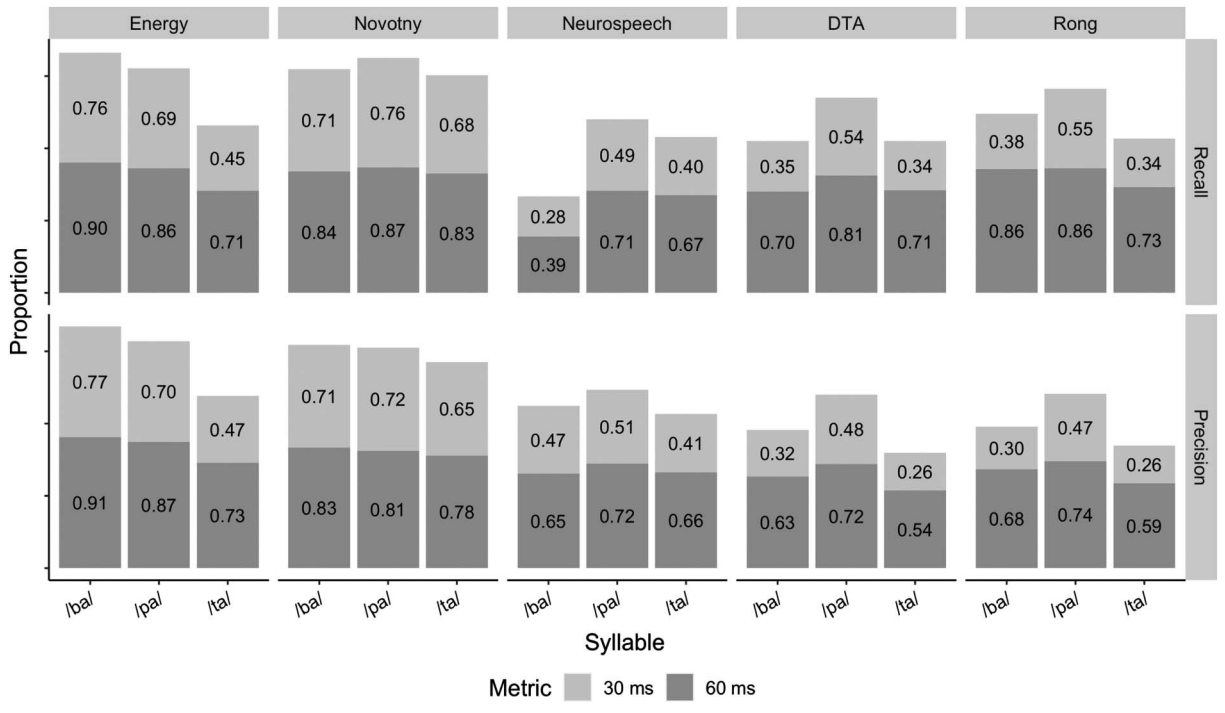


Table 6. Kendall tau-b correlations between manually and algorithmically calculated DDK rate and cTV, by severity groups.

Algorithm	DDK rate			cTV		
	Mild	Moderate	Severe	Mild	Moderate	Severe
Energy	**.76**	**.87**	**.70**	**.76**	**.88**	**.70**
Novotny	.69	**.75**	.49	.68	**.74**	.49
DTA	.65	.51	.27	.67	.53	.28
Neurospeech	.31	.32	.38	.26	.33	.42
Rong	.54	.47	.24	.51	.46	.22

Note. Correlations that exceeded the .7 threshold (strong) are in bold. DDK = diadochokinesis; cTV = cycle-to-cycle temporal variability; DTA = Diadochokinetic Tasks Analysis.

severity, measured primarily by mixed LoA, boundary recall and precision, and correlations with manual measurements. All algorithms, except Energy, began to show significant discrepancies in syllable count agreement with manual analysis by the moderate stage of dysarthria (intelligible speaking rate between 100 and 149 WPM). At the most severe stage (intelligible speaking rate < 100 WPM), all five algorithms counted syllables relatively poorly with a tendency to overestimate the syllable counts relative to the manual method. Furthermore, boundary detection recall and precision deviated progressively from manual analysis as severity worsened. However, for these measures, Energy and Novotny deviated the least among all algorithms, at least in the mild and moderate dysarthria cases. In addition, the syllable type affected syllable count agreement for most of the methods. Novotny yielded the highest recall and precision (0.78–0.87) across all tasks. Overall, among the algorithms, Energy and Novotny methods proved the most robust in conducting automatic DDK analysis, either in terms of variation across the severity range or syllable types.

Similarities and differences in the algorithms' performance can be explained by differences in their computational frameworks. The Energy and Novotny methods primarily use a dynamic envelope thresholding approach (e.g., moving average), while DTA and Rong relay on static peak identification-based methods. Novotny, as compared to Energy, offered a more complicated thresholding strategy using a Bayesian step-change detector and polynomial thresholding, yet did not perform substantially better than a simpler absolute energy-based method across the range of dysarthria severity. The step-change detector appeared more sensitive to small fluctuations in noisy data, which typically become more prominent in more severe speech.

Table 7. Kendall tau-b correlations between manually and algorithmically calculated DDK rate and cTV, by syllable type.

Algorithm	DDK rate			cTV		
	/ba/	/pa/	/ta/	/ba/	/pa/	/ta/
Energy	**.82**	**.81**	**.87**	**.82**	**.80**	**.87**
Novotny	.69	**.73**	**.74**	.68	**.71**	**.72**
DTA	**.78**	.65	.40	**.79**	.67	.41
Neurospeech	.29	.59	.64	.15	.59	.65
Rong	.43	.62	.47	.43	.59	.47

Note. Correlations that exceeded the .7 threshold (strong) are in bold. DDK = diadochokinesis; cycle-to-cycle temporal variability; DTA = Diadochokinetic Tasks Analysis.

Novotny showed less varied results across the syllable types. In general, the variation in algorithmic performance was expected across tasks because of the differences in the effect of dysarthria on the voicing and place of articulation features of the consonants, which posit varying degrees of acoustic degradation on these features (Ackermann & Ziegler, 1991; Kent et al., 1991; Riddel et al., 1995). Based on clinical observations, it was expected that the detection of syllable boundaries would be more challenging in /ta/ than in /ba/ and /pa/. Energy appeared consistent with this expectation, as both recall and precision were worse for the alveolar as compared to the bilabial task. Recall and precision values for both Energy and Novotny methods were well matched, meaning that their ability to detect correct boundaries (recall) and minimize the number of false boundaries (precision) was similar.

DTA, Rong, and Neurospeech performed relatively poorly across severity and syllable types in this study. Neurospeech, like Energy and Novotny, used thresholding to segment the signal. However, unlike the other two methods, it used a VAD, which ended up being highly sensitive to natural voicing fluctuation. Meanwhile, DTA and Rong—the peak-based algorithms—used predefined heights and distance thresholds for all speakers (e.g., sampling frequency/20) and were not sufficiently sensitive to the random fluctuations in the signal due to the aperiodicities present in the voice of severely impaired speakers (Herzel et al., 1995; Zhang et al., 2005). Rong's original study (2020) addressed the issue of fluctuation due to intersyllable voicing via implementing manual adjustments (e.g., manually identifying if a detected peak exceeded the threshold). Since we did not employ manual adjustment, relatively poor performance of this method was unsurprising. Although the DTA framework is fully automatic, its poor performance suggested that automatic peak-based algorithms might be inferior to thresholding-based methods.

Reference norms for raw syllable counts in speakers with bulbar disease are extremely scarce in the literature, where the derived measures like DDK rate are instead reported. Eshghi et al. (2019) found approximately a 10-syllable difference in the mean raw syllable count between healthy controls and a group of individuals in the early stage of ALS. Based on this reference, we inferred a clinically meaningful LoA to be ± 10 syllables. Energy was the only algorithm to produce LoA within this range, and although Novotny (LoA = −12.4, 9.46) was a close contender, Energy was the most relatively well-rounded and consistent algorithm, showing a reasonable ability to capture clinical metrics such as syllable rate and variability. The findings of this study thus warrant a recommendation to use the basic Energy algorithm for DDK analysis. However, manual intervention would still be required with severely dysarthric samples and alveolar syllables.

Based on the results of this study, further development of automatic methods for DDK analysis is warranted. In recent years, machine learning (ML) methods have emerged to improve automatic DDK analysis (Rozenstoks et al., 2020; Wang et al., 2019). These methods have shown superior performance to non-ML ones but require substantial user data for training to ensure high accuracy (Rozenstoks et al., 2020). Having a limited sample size across severity range for training would bias the algorithms, as Novotny et al. (2020) reported. This study did not include ML tools in order to keep all algorithms disease-agnostic and avoid potential training data bias.

The clinical recommendations made in this study cannot be fully embraced without the availability of these algorithms. Unlike Energy, Neurospeech, and DTA, Novotny is not yet open source and cannot be used for research and clinical practice. In comparison, Neurospeech and DTA are publicly available but are not recommended for use with individuals with bulbar ALS, based on the current study findings. The software developers should expedite the development of automatic methods to support clinical assessment of ALS.

Study Limitations

Several limitations befell this study. First, an unbalanced data set with uneven samples per severity group and task was used, which warranted the use of the mixed Bland–Altman LoA. We expect that a more balanced data set would strengthen the results and emphasize the variability of algorithm measurements due to both severity and task. This expectation remains a question to be explored in future studies. Nevertheless, results of this study only reported an overall LoA across groups, without considering that the LoA may need to be narrower for severe dysarthria counts since said counts are somewhat lower than those of the other groups. Therefore, more group-sensitive Bland–Altman statistics should be used to improve the clinical interpretability of the reported LoA (see the work of Parker et al., 2020, for a summary). Another limitation concerned the use of intelligible speaking rate to measure bulbar disease progression, as previous studies typically stratified individuals into severity groups based on other metrics, for example, the ALSFRS-R total or bulbar score (see the work of Rong et al., 2020). The results may differ to some extent with different stratification methods. Lastly, boundary recall and precision evaluation methods differed between the thresholding- and peak-based algorithms as different parts of the syllable were used for manual reference, which may have affected the validity of the results. Future studies should focus on algorithms that can output the same reference points.

Conclusions

Dysarthria severity and syllable type affected the performance of current automatic DDK analysis algorithms to varying degrees. Algorithms based on or derived from energy envelope segmentation (Energy and Novotny) outperformed peak-based ones in measuring syllable count, boundary detection recall and precision, and clinical metrics like DDK rate. Between the thresholding-based algorithms, Energy showed superior ability with regard to syllable count detection and correlations with clinical metrics across dysarthria severities but is still not sufficient by itself for clinical use without some degree of manual intervention in severe samples and the alveolar context. An analysis with a more balanced data set and novel ML methods are needed for a more advanced, comprehensive evaluation of automatic methods for DDK analysis.

Acknowledgments

This research was supported by National Institute on Deafness and Other Communication Disorders Grants R01DC009890, R01DC013547, and R01DC017291 awarded to Yana Yunusova and Jordan Green, as well as ALS Canada and Brain Canada for the Canadian ALS Neuroimaging Consortium (https://calsnic.org) grants awarded to Yana Yunusova. We extend our deepest gratitude to the participants and their families for their contribution to this study. We also thank Madhura Kulkarni and Farah Wehbe for their assistance with the project.

References

Ackermann, H., & Ziegler, W. (1991). Articulatory deficits in Parkinsonian dysarthria: An acoustic analysis. *Journal of*

Neurology, Neurosurgery, & Psychiatry, 54(12), 1093–1098. https://doi.org/10.1136/jnnp.54.12.1093

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Barnett, C., Green, J. R., Marzouqah, R., Stipancic, K. L., Berry, J. D., Korngut, L., Genge, A., Shoesmith, C., Briemberg, H., Abrahao, A., Kalra, S., Zinman, L., & Yunusova, Y. (2020). Reliability and validity of speech & pause measures during passage reading in ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 21*(1–2), 42–50. https://doi.org/10.1080/21678421.2019.1697888

Ben-Gal, I. (2010). Outlier Detection. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 117–130). Springer. https://doi.org/10.1007/978-0-387-09823-4_7

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Brooks, B. R., Miller, R. G., Swash, M., & Munsat, T. L. (2000). El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders, 1*(5), 293–299. https://doi.org/10.1080/146608200300079536

Chen, A. (2005). Otolaryngologic presentations of amyotrophic lateral sclerosis. *Otolaryngology—Head & Neck Surgery, 132*(3), 500–504. https://doi.org/10.1016/j.otohns.2004.09.092

Colonna, J. G., Cristo, M., Salvatierra, M., & Nakamura, E. F. (2015). An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications: An International Journal, 42*(21), 7367–7374. https://doi.org/10.1016/j.eswa.2015.05.030

Eshghi, M., Stipancic, K. L., Mefferd, A., Rong, P., Berry, J. D., Yunusova, Y., & Green, J. R. (2019). Assessing oromotor capacity in ALS: The effect of a fixed-target task on lip biomechanics. *Frontiers in Neurology, 10,* Article 1288. https://doi.org/10.3389/fneur.2019.01288

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

Herzel, H., Berry, D., Titze, I., & Steinecke, I. (1995). Nonlinear dynamics of the voice: Signal analysis and biomechanical modeling. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 5*(1), 30–34. https://doi.org/10.1063/1.166078

Hillel, A. (1999). Presentation of ALS to the otolaryngologist/head and neck surgeon: Getting to the neurologist. *Neurology, 53*(8), S22.

Kalra, S., Khan, M., Barlow, L., Beaulieu, C., Benatar, M., Briemberg, H., Chenji, S., Clua, M. G., Das, S., Dionne, A., Dupré, N., Emery, D., Eurich, D., Frayne, R., Genge, A., Gibson, S., Graham, S., Hanstock, C., Ishaque, A., . . . for the Canadian ALS Neuroimaging Consortium.. (2020). The Canadian ALS Neuroimaging Consortium (CALSNIC)—A multicentre platform for standardized imaging and clinical studies in ALS. *MedRxiv*. https://doi.org/10.1101/2020.07.10.20142679

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23. https://doi.org/10.1044/1058-0360.0503.07

Kent, R. D., Sufit, R. L., Rosenbek, J. C., Kent, J. F., Weismer, G., Martin, R. E., & Brooks, B. R. (1991). Speech deterioration in amyotrophic lateral sclerosis: A case study. *Journal of Speech and Hearing Research, 34*(6), 1269–1275. https://doi.org/10.1044/jshr.3406.1269

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech. *Journal of Communication Disorders, 32*(3), 141–186. https://doi.org/10.1016/S0021-9924(99)00004-0

Kiernan, M. C., Vucic, S., Cheah, B. C., Turner, M. R., Eisen, A., Hardiman, O., Burrell, J. R., & Zoing, M. C. (2011). Amyotrophic lateral sclerosis. *The Lancet, 377*(9769), 12–18. https://doi.org/10.1016/S0140-6736(10)61156-7

Novotny, M., Melechovsky, J., Rozenstoks, K., Tykalova, T., Kryze, P., Kanok, M., Klempir, J., & Rusz, J. (2020). Comparison of automated acoustic methods for oral diadochokinesis assessment in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 63*(10), 3453–3460. https://doi.org/10.1044/2020_JSLHR-20-00109

Novotny, M., Rusz, J., Čmejla, R., & Ružička, E. (2014). Automatic evaluation of articulatory disorders in Parkinson's disease. *IEEE/ACM Transactions on Audio, Speech, Language Processing, 22*(9), 1366–1378. https://doi.org/10.1109/TASLP.2014.2329734

Orozco-Arroyave, J. R., Vásquez-Correa, J. C., Vargas-Bonilla, J. F., Arora, R., Dehak, N., Nidadavolu, P. S., Christensen, H., Rudzicz, F., Yancheva, M., Chinaei, H., Vann, A., Vogler, N., Bocklet, T., Cernak, M., Hannink, J., & Nöth, E. (2018). NeuroSpeech. *SoftwareX, 8,* 69–70. https://doi.org/10.1016/j.softx.2017.08.004

Parker, R. A., Scott, C., Inácio, V., & Stevens, N. T. (2020). Using multiple agreement methods for continuous repeated measures data: A tutorial for practitioners. *BMC Medical Research Methodology, 20*(1), 154. https://doi.org/10.1186/s12874-020-01022-x

Parker, R. A., Weir, C. J., Rubio, N., Rabinovich, R., Pinnock, H., Hanley, J., McCloughan, L., Drost, E. M., Mantoani, L. C., MacNee, W., & McKinstry, B. (2016). Application of mixed effects limits of agreement in the presence of multiple sources of variability: Exemplar from the comparison of several devices to measure respiratory rate in COPD patients. *PLOS ONE, 11*(12), Article e0168321. https://doi.org/10.1371/journal.pone.0168321

Riddel, J., McCauley, R. J., Mulligan, M., &Tandan, R. (1995). Intelligibility and phonetic contrast errors in highly intelligible speakers with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Research, 38*(2), 304–314. https://doi.org/10.1044/jshr.3802.304

Rong, P. (2020). Automated acoustic analysis of oral diadochokinesis to assess bulbar motor involvement in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 63*(1), 59–73. https://doi.org/10.1044/2019_JSLHR-19-00178

Rong, P., Yunusova, Y., Eshghi, M., Rowe, H. P., &Green, J. R. (2020). A speech measure for early stratification of fast and slow progressors of bulbar amyotrophic lateral sclerosis: Lip movement jitter. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 21*(1–2), 34–41. https://doi.org/10.1080/21678421.2019.1681454

Rong, P., Yunusova, Y., & Green, J. R. (2015). Speech intelligibility decline in individuals with fast and slow rates of ALS progression. In P. Bell & B. Ramabhadran (Eds.), *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association* (pp. 2967–1971). International Speech Communication Association.

Rong, P., Yunusova, Y., Wang, J., Zinman, L., Pattee, G. L., Berry, J. D., Perry, B., & Green, J. R. (2016). Predicting speech intelligibility decline in amyotrophic lateral sclerosis

based on the deterioration of individual speech subsystems. *PLOS ONE, 11*(5), Article e0154971. https://doi.org/10.1371/journal.pone.0154971

Rozenstoks, K., Novotny, M., Horakova, D., & Rusz, J. (2020). Automated assessment of oral diadochokinesis in multiple sclerosis using a neural network approach: Effect of different syllable repetition paradigms. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28*(1), 32–41. https://doi.org/10.1109/TNSRE.2019.2943064

RStudio Team. (2020). *RStudio: Integrated development for R.* http://www.rstudio.com/

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Shellikeri, S., Green, J. R., Kulkarni, M., Rong, P., Martino, R., Zinman, L., & Yunusova, Y. (2016). Speech movement measures as markers of bulbar disease in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 59*(5), 887–899. https://doi.org/10.1044/2016_JSLHR-S-15-0238

Smékal, Z., Mekyska, J., Rektorová, I., & Faúndez-Zanuy, M. (2013). Analysis of neurological disorders based on digital processing of speech and handwritten text. *International Symposium on Signals, Circuits and Systems (ISSCS),* 1–6. https://doi.org/10.1109/ISSCS.2013.6651178

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

Vásquez-Correa, J. C., Rios-Urrego, C. D., Rueda, A., Orozco-Arroyave, J. R., Krishnan, S., & Nöth, E. (2019). Articulation and empirical mode decomposition features in diadochokinetic exercises for the speech assessment of Parkinson's disease patients BT. In I. Nyström, Y. H. Heredia, & V. M. Núñez (Eds.), *Progress in pattern recognition, image analysis, computer vision, and applications* (pp. 688–696). Springer.

Wang, Y., Gao, K., Kloepper, A., Zhao, Y., Kuruvilla-Dugdale, M., Lever, T., & Bunyak, F. (2019). DeepDDK: A deep learning based oral-diadochokinesis analysis software. In J. Liang, & D. I. Fotiadis (Eds.), *IEEE-EMBS International Conference on Biomedical and Health Informatics* (Vol. 2019). IEEE. https://doi.org/10.1109/BHI.2019.8834506

Wang, Y. T., Kent, R. D., Duffy, J. R., & Thomas, J. E. (2009). Analysis of diadochokinesis in ataxic dysarthria using the Motor Speech Profile program. *Folia Phoniatrica et Logopaedica, 61*(1), 1–11. https://doi.org/10.1159/000184539

Wang, J., Kothalkar, P. V., Kim, M., Yunusova, Y., Campbell, T. F., Heitzman, D., & Green, J. R. (2016). Predicting intelligible speaking rate in individuals with amyotrophic lateral sclerosis from a small number of speech acoustic and articulatory samples. *Workshop on Speech and Language Processing for Assistive Technologies, 2016,* 91–97. https://doi.org/10.21437/SLPAT.2016-16

Yorkston, K. M., & Beukelman, D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders, 46*(3), 296–301. https://doi.org/10.1044/jshd.4603.296

Yorkston, K. M., Miller, R., & Strand, E. (1995). *Management of speech and swallowing in degenerative diseases*. Communication Skill Builders.

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

Yunusova, Y., Plowman, E. K., Green, J. R., Barnett, C., & Bede, P. (2019). Clinical measures of bulbar dysfunction in ALS. *Frontiers in Neurology, 10,* 106. https://doi.org/10.3389/fneur.2019.00106

Zhang, Y., Jiang, J., & Rahn, D. A. (2005). Studying vocal fold vibrations in Parkinson's disease with a nonlinear model. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 15*(3), Article 033903. https://doi.org/10.1063/1.1916186