Automated Assessment of Oral Diadochokinesis in Multiple Sclerosis Using a Neural Network Approach: Effect of Different Syllable Repetition Paradigms

Kris Rozenstoks, Michal Novotny, Dana Horakova, and Jan Rusz

***Abstract*—Slow and irregular oral diadochokinesis represents an important manifestation of spastic and ataxic dysarthria in multiple sclerosis (MS). We aimed to develop a robust algorithm based on convolutional neural networks for the accurate detection of syllables from different types of alternating motion rate (AMR) and sequential motion rate (SMR) paradigms. Subsequently, we explored the sensitivity of AMR and SMR paradigms based on voiceless and voiced consonants in the detection of speech impairment. The four types of syllable repetition paradigms including /ta/, /da/, /pa/-/ta/-/ka/, and /ba/-/da/-/ga/ were collected from 120 MS patients and 60 matched healthy control speakers. Our neural network algorithm was able to correctly identify the position of individual syllables with a very high average accuracy of 97.8 %, with the correct temporal detection of syllable position of 87.8 % for 10 ms and 95.5 % for 20 ms tolerance value. We found significantly altered diadochokinetic rate and regularity in MS compared to controls across all types of investigated tasks ($p < 0.001$). MS patients showed slower speech for SMR compared to AMR tasks, whereas voiced paradigms were more irregular. Objective evaluation of oral diadochokinesis using different AMR and SMR paradigms may provide important information regarding speech severity and pathophysiology of the underlying disease.**

***Index Terms*—Cerebellar, dysarthria, speech disorder, acoustic, deep learning.**

Manuscript received March 26, 2019; revised July 8, 2019 and August 20, 2019; accepted September 18, 2019. Date of publication September 23, 2019; date of current version January 8, 2020. This work was supported in part by the Czech Science Foundation under Grant 16-03322S and in part by the Research Center for Informatics under Grant CZ.02.1.01/0.0/0.0/16_019/0000765. *(Corresponding author: Jan Rusz.)*

K. Rozenstoks is with the Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, 160 00 Prague, Czech Republic, and also with the Institute of Biomedical Engineering and Nanotechnologies, Riga Technical University, 1048 Riga, Latvia (e-mail: k.rozenstoks@gmail.com).

M. Novotny is with the Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, 160 00 Prague, Czech Republic (e-mail: novotm26@fel.cvut.cz).

D. Horakova is with the Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine, Charles University, 120 00 Prague, Czech Republic (e-mail: dana.horakova@vfn.cz).

J. Rusz is with the Department of Circuit Theory, Faculty of Electrical Engineering, Czech Technical University in Prague, 160 00 Prague, Czech Republic, and also with the Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine, Charles University, 120 00 Prague, Czech Republic (e-mail: ruszjan@fel.cvut.cz).

Digital Object Identifier 10.1109/TNSRE.2019.2943064

I. Introduction

MULTIPLE sclerosis (MS) is the most common acquired demyelinating disease of the central nervous system occurring mainly in young and middle-aged adults and affecting about 0.1% to 0.2% of the population [1]. As a result of widespread brain atrophy affecting mostly white and gray matter, MS presents a wide range of neurological manifestations with motor, sensory and cognitive impairments. Motor speech impairment termed dysarthria is likely one of the least well-described clinical signs of MS (see [2] for review). Dysarthria in MS is typically mild with a reported prevalence of up to 60% [3]–[7]. Ataxic, spastic and mixed ataxic-spastic dysarthria are the most common dysarthria subtypes encountered in MS due to the involvement of cerebellum and pyramidal tract [8]. Darley *et al.* [9] first perceptually distinguished impaired loudness control, harshness, defective articulation, impaired emphasis, insufficient pitch control, hypernasality, inappropriate pitch level, breathiness, and articulatory breakdowns as most distinctive manifestations of dysarthria in MS. Only a few studies have verified or extended the perceptual observations of Darley *et al.* [9] by objective acoustic analyses and documented primarily phonatory abnormalities [10]–[12], as well as articulatory-prosodic disorder presenting by imprecise articulation, monopitch, articulatory decay, excess loudness variations, slow rate and various temporal deficits [5], [13]–[15]. Importantly, previous research has shown that the severity of dysarthria is attributed to the overall severity of neurological disease [4], [5], [15]. This observation provides an opportunity to consider objective speech evaluation as a potential biomarker for monitoring disease progression in MS. Speech assessment is fast, non-invasive, inexpensive, easy to apply and can be fully automated and monitored remotely, even by a smartphone application from the patient's home [16].

Oral diadochokinesis is a traditional component of motor speech assessment, which measures the motor abilities of the speech articulators and reveals their movement limitations [8]. It is considered an essential speech paradigm for the differential diagnosis of dysarthria as well as for determining the severity of speech motor control dysfunction [8]. Two basic measures quantify diadochokinetic (DDK) performance. DDK rate identifies articulatory velocity by analyzing the number of syllable vocalizations per time. DDK regularity refers to temporal irregularity by measuring the standard deviation of distances between following syllables. DDK rate tends to be slow in speakers with both spastic and ataxic dysarthria [17], [18], whereas diadochokinesis irregularity has been assumed to be most characteristic of ataxic dysarthria [19], [20]. Slow and irregular oral diadochokinesis has also been demonstrated in MS [21], [22], with DDK regularity reported as a sensitive measure for differentiation between MS and healthy controls [5].

Two types of DDK tasks are commonly used. Alternating motion rate (AMR) refers to the rapid repetition of single syllables such as /ta/ whereas sequential motion rate (SMR) indicates the rapid repetition of syllable sequences such as /pa/-/ta/-/ka/ [8]. Compared to AMR, SMR is more challenging to perform due to consecutive repetition of bilabial, alveolar and velar consonants. Therefore, SMR tasks are typically used to evaluate disorders with sequential planning and programming deficits such as apraxia of speech or cerebellar ataxia [8]. Indeed, previous research has already shown poorer performance of MS patients when performing SMR compared to the AMR task [22].

However, to the best of our knowledge, previous research in MS (or ataxic dysarthria in general) has only considered AMR/SMR tasks with voiceless consonants. Interestingly, a recent study revealed that patients with predominant cerebellar dysarthria secondary to multiple system atrophy manifested problems with the articulation of voiced consonants [23]. This articulatory undershoot of voiced consonants was characterized by the shortening of negative voice onset time duration until the voicing lead vanished and only short burst remained. Importantly, the extent of these deficits in the articulation of voiced consonants was related to the severity of cerebellar motor impairment. One might thus assume that modified AMR and SMR tasks based on voiced consonants such as /da/ or /ba/-/da/-/ga/ may be even more sensitive to cerebellar deficits, which may aid in the differential diagnosis of dysarthrias and monitoring the extent of speech impairment in cerebellar disorders.

To facilitate the use of oral diadochokinesis in common clinical practice, reliable and automatic methods for its accurate assessment are needed. Several methods for the automated assessment of oral diadochokinesis performance have already been developed. The Kay-Pentax Motor Speech Profile Program was developed for the analysis of oral diadochokinesis and tested using unvoiced AMR (/pa/, /ta/, and /ka/) collected from 21 speakers with ataxic dysarthria [24]. Although agreement across individual measures based on automated and hand labels was high, the detection accuracy of the algorithm was not tested. Another study designed an automated algorithm for the detection of events during the SMR paradigm based on /pa/-/ta/-/ka/ syllable repetitions of 27 Parkinson's disease patients [25]. The detection accuracy of this algorithm for voice onset time with 10 ms interval tolerance reached as high as 80%. Finally, likely the best performing automated algorithm for the evaluation of SMR performance based on /pa/-/ta/-/ka/ repetition of 24 Parkinson's disease patients was developed by Novotny *et al.* [26]. This algorithm reached an accuracy of up to 90% for the detection of vowel onset with 10 ms interval tolerance.

In summary, the accuracy of all previously developed algorithms was tested using only one type of DDK task based on voiceless consonants [24]–[26]. It is therefore unclear whether the accuracy of these algorithms would be sufficient across voiced DDK paradigms. Also, all previous algorithms were designed using traditional digital speech signal processing techniques [24]–[26]. The recent development and success in the field of object detection in images by convolutional neural networks (CNN) [27], [28] implies that a CNN-based approach could be beneficial for the precise detection of syllables during oral diadochokinesis. Indeed, a very recent study successfully used CNN to model transitions between voiced and unvoiced segments in dysarthric speech secondary to Parkinson's disease [29]. Nevertheless, the possible utility of CNN for accurate syllable detection during the DDK task in dysarthric speech has not yet been explored.

Therefore, the present study was designed to address the following aims:

(i) To develop a robust CNN-based segmentation algorithm allowing the accurate detection of syllables from different types of AMR and SMR paradigms. We hypothesized that a CNN-based approach would outperform traditional digital speech signal processing techniques.

(ii) To explore the sensitivity of AMR and SMR paradigms based on voiceless and voiced consonants in the detection of speech impairment in MS. We hypothesized that temporal irregularity of diadochokinesis in patients with MS would be greater for voiced than voiceless paradigms due to cerebellar involvement, whereas healthy controls would show similar DDK performance regularity across all paradigms.

(iii) To compare the performances in DDK tasks with the extent of neurological deficits to provide greater insight into the pathophysiology of dysarthria in MS. We hypothesized that performance in DDK rate would be correlated to overall disability while the extent of DDK irregularity would parallel the severity of cerebellar dysfunction.

II. Methods

A. Participants

A total of 180 consecutive Czech participants were recruited from 2016 to 2017. The study was approved by the Ethics Committee of the General University Hospital, Prague, Czech Republic and every participant provided written, informed consent. One hundred and twenty patients (89 women), mean age 44 (SD 11) years, were diagnosed with MS according to the revisited McDonald Criteria [30]. From this cohort, 94 patients were diagnosed with relapsing-remitting MS, 15 with secondary progressive MS, 3 with clinically isolated syndrome and 8 with primary-progressive MS. All patients were relapse-free for at least 30 days prior to testing. In addition, 60 volunteers (44 women), mean age 44 (SD 12) years, with no history of neurological or communication disorders were included as a healthy control group. The clinical severity of patients was estimated using the Expanded Disability Status Scale (EDSS) [31], which represents a widely used method for the quantification of disability in MS and monitors eight functional systems: pyramidal, cerebellar, brain stem, sensory, bowel and bladder, visual, mental, and other functions. Every domain is scored from 0 (no disability) to 6 (maximal disability); the pyramidal and cerebellar subscores were of interest in the present study due to the presence of spastic-ataxic dysarthria. An integrated EDSS score ranging from 0 (normal examination) to 10 (death due to MS) is obtained according to the score based upon each functional system and ambulation.

The investigated subjects also participated in a previous study focused on the detailed assessment of severity and patterns of dysarthria [5]; however, characteristics related to different AMR and SMR tasks were not previously investigated. The severity and type of dysarthria were based on perceptual estimation by two speech-language pathologists with experience in motor speech disorders using vocal paradigms of vowel prolongation, DDK task, and reading passage in according with the perceptual criteria outlined by Darley *et al.* [32]. Since the inter-rater reliability estimated using the two-way mixed single score intra-class correlation reached a relatively low value of 0.47, the final reported characteristics of dysarthria were based on consensus judgment of two speech-language pathologists. Although the dysarthria was imperceptible in 90 MS patients, acoustic analysis has the potential to reveal even subperceptual speech deviations [33]. The perceptible dysarthria in 30 MS patients mainly featured a combination of spastic and ataxic components with primary signs of slow rate, irregular speech timing, imprecise articulation, strained-strangled voiced and unnatural word stress expression. In addition, for MS patients with perceptible dysarthria, audio recordings were perceptually analyzed according to a five-dimensional scoring system (0 indicating no impairment and 4 indicating major impairment) relating to the major dimensions of dysarthria, i.e., respiration, voice quality, articulation, resonance, prosody, dysfluency, and naturalness. Clinical characteristics of the MS patients can be found in Table I.

B. Speech Examination

Speech recordings were performed in a quiet room with a low ambient noise level using a head-mounted condenser microphone (Beyerdynamic Opus 55, Heilbronn, Germany) situated approximately 5 cm from the mouth of each subject. Speech signals were sampled at 48 kHz with 16-bit resolution. Each participant was instructed to repeat the syllable /ta/, /da/, /pa/-/ta/-/ka/, and /ba/-/da/-/ga/ as quickly and accurately as possible at least seven times per one breath. Each of the four syllable repetition tasks was performed twice. Concerning previous research reporting that tongue function in MS was significantly more affected than lip function [13], the syllable /ta/ (and its voiced cognate /da/) was preferred as it best reflects the movement of the tongue. Also, from a phonetic point of view, /b/, /d/ and /g/ in the consonant-vowel context are usually pronounced as prevoiced in Czech (i.e., voiced during closure), while /p/, /t/ and /k/ are pronounced as voiceless and unaspirated [34].

TABLE I
CLINICAL CHARACTERISTICS OF MS PATIENTS. MS = MULTIPLE SCLEROSIS, EDSS =EXPANDED DISABILITY STATUS SCALE

	MS (n = 120)
Clinical characteristics (n = 120)	
Mean Age (years)	44 (SD 11, range 18–74)
Female (%)	74% (n = 89)
Mean disease duration (years)	14 (SD 8, range 2–37)
Mean EDSS score	3.6 (SD 1.4, range 1.0–6.5)
Mean EDSS pyramidal subscore	2.5 (SD 0.9, range 1.0–4.0)
Mean EDSS cerebellar subscore	1.4 (SD 1.2, range 0–4.0)
Dysarthria severity (n = 120)	
None	70% (n = 84)
Mild	27% (n = 33)
Moderate	3% (n = 3)
Severe	0% (n = 0)
Perceptible dysarthria type (n = 36)	
Spastic-ataxic dysarthria	24% (n = 29)
Spastic dysarthria	4% (n = 5)
Ataxic dysarthria	2% (n = 2)
Perceptible dysarthria characteristics (n = 36)	
Mean respiration	1.1 (SD 0.9, range 0–3)
Mean voice quality	2.4 (SD 0.6, range 1–4)
Mean articulation	1.8 (SD 0.9, range 0–3)
Mean resonance	0.9 (SD 0.7, range 0–2)
Mean prosody	1.8 (SD 0.8, range 0–3)
Mean dysfluency	0.1 (SD 0.3, range 0–1)
Mean naturalness	2.3 (SD 0.6, range 1–4)

C. Reference Hand Labels

For algorithm tuning and to obtain feedback for the evaluation of its reliability, manual syllable annotations for 15.6% of randomly selected utterances (i.e., for 224 utterances) from all types of AMR/SMR tasks and both MS and control groups were performed. In each syllable vocalization, the positions of two events including the initial burst of the consonant (/p/, /t/, /k/, /b/, /d/, or /g/) and occlusion of the vowel /a/ were annotated. This approach was preferred as it is difficult to hand-label the correct position of maximal energy during each syllable by visual inspection of speech waveforms. Previously defined rules were used as a foundation for our labeling criteria by Novotny *et al.* [26]. The time domain was preferred for the specification of burst onset due to its better resolution. In the case of multiple bursts, the initial burst was marked. In the frequency domain, the burst onset was characterized by the presence of moderate excitation of one or few time windows of spectrogram over the entire frequency range. The frequency domain was used for the identification of vowel occlusion, where the energy of fundamental, as well as the first three formant frequencies, slowly weakens. The second formant vowel offset was considered as the best indicator of occlusion onset. Figure 1 illustrates examples of voiced and unvoiced syllables with manually obtained labels.

D. CNN-Based Algorithm for Automatic Segmentation

The entire process of data pooling, CNN training and performance testing is illustrated in Figure 2. A subset of 144 utterances from 36 participants (10% of the entire dataset) with

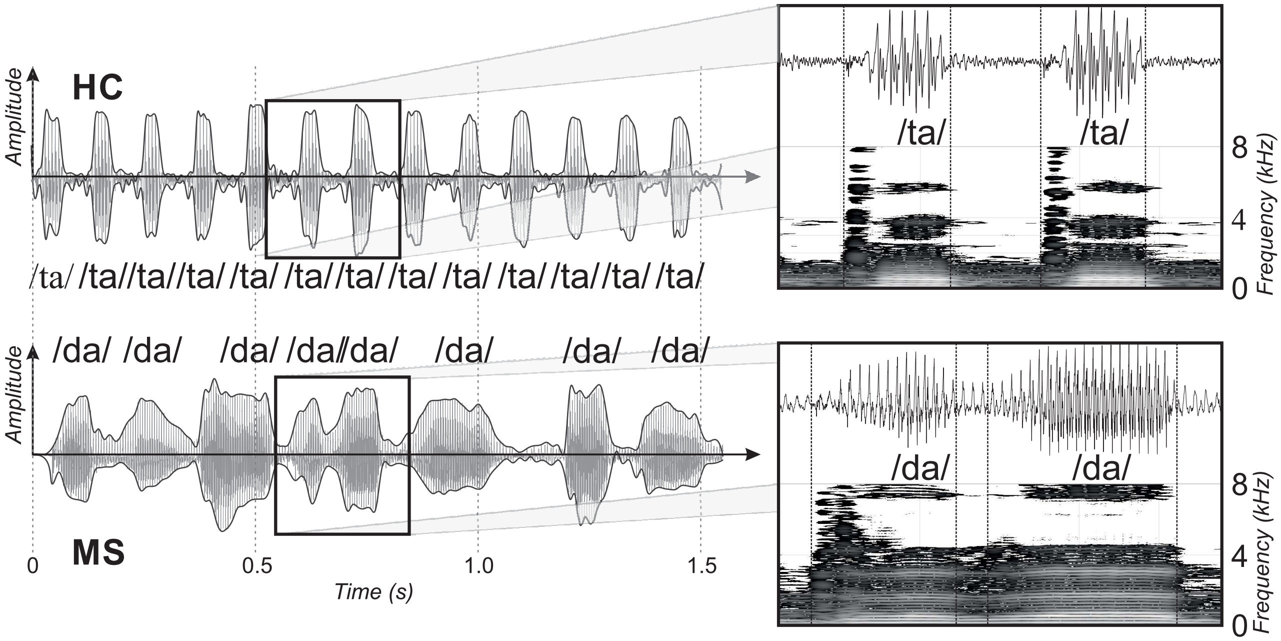

Fig. 1. Example of voiceless and (top) and voiced (bottom) AMR task with zoom to the oscillographic sound pressure signal and its respective spectrogram for two syllables with highlighted positions of burst and vowel occlusion by hand labels. An exemplary HC speaker (top) performed syllable repetition at a fast tempo (DDK rate = 8.1 syll/s) and stable regularity (DDK regularity = 3.7 ms), whereas the exemplary MS patient manifested slow velocity (DDK rate = 5.9 syll/s) and irregularity (DDK regularity = 60.5 ms). MS = multiple sclerosis, HC = healthy controls.

reference hand labels was selected as an independent testing dataset. Another 80 hand-labeled recordings from 20 participants (5.6% of the entire dataset) were used for CNN tuning. Both datasets were equally populated with recordings from HC speakers and participants diagnosed with MS. In addition, the training dataset consisted of 16 women (80%), 3 MS patients with perceptible dysarthria (30%), and 12 speakers with age < 50 years (60%; mean 48, SD 11 years), whereas testing dataset was composed of 28 women (78%), 5 MS patients with perceptible dysarthria (28%), and 26 speakers with age < 50 years (72%; mean 43, SD 12 years); there were no statistically significant differences between training and validation dataset for gender ($p = 0.86$), dysarthria severity ($p = 0.93$), or age ($p = 0.10$).The signals in the tuning dataset were downsampled to 16 kHz, and the spectrograms of the downsampled signals were estimated using a 1 ms window with 50% window overlap and zero padding to 256 samples. The obtained spectrograms were divided into 300 ms windows, and these windows were transformed into images 256 pixels wide and 192 pixels high. The corresponding hand labels were transformed into a PASCAL VOC style label file. Using this approach, 80 recordings designed for tunning were split into 2430 spectrogram images accompanied by the PASCAL VOC label file. These images were then divided into training (70%; 1700 images) and validation subsets (30%; 730 images).

Tensorflow implementation of Faster R-CNN in the Resnet-101 structure was used for syllable border detection [35]. The faster R-CNN was employed as a state-of-the-art object detection model which reduces the computational burden by using a separate neural network providing region proposals [28]. The Faster R-CNN model was used in the topology of the residual network Resnet with a bottleneck design and depth of 101 layers (Resnet-101) [27]. The advantage of the residual network is the signal shortcut which enables one to propagate input directly to the output of the net. Changes in the output of the net are realized by residual mapping according to equation 1:

$$H(x) = F(x) + x, \tag{1}$$

where x denotes the input of the net, $H(x)$ describes the output of the net and the $F(x)$ is the mapped residual. This approach helps to overcome the problem of the vanishing gradient in deep net structures. A version of the CNN pre-trained on the Common Objects in Context database was used to reduce training demands [36]. The training of the neural network was stopped after 1213 iterations when the loss function reached a value of 0.006.

The algorithm parsing audio recordings with the CNN input was designed to avoid confusion on the borders of the spectrogram window during the audio analysis phase (see Figure 2, section Audio analysis algorithm). In the first iteration, this algorithm generated a 300 ms spectrogram window and passed it as the CNN input. The algorithm detected syllables in the spectrogram and defined them by their beginning and end positions. The syllable positions found by CNN were saved in a.csv file. In the event that no syllable was detected, the active window was shifted by 10 ms and another attempt for CNN object detection was performed. If necessary, this approach was repeated until a syllable was found. If no new syllables were found for 2 seconds, the last selected frame was considered as the end. When a syllable was found, the last detected syllabic position was used to define the beginning of the new 300 ms window. This process was repeated until the end of the recording.

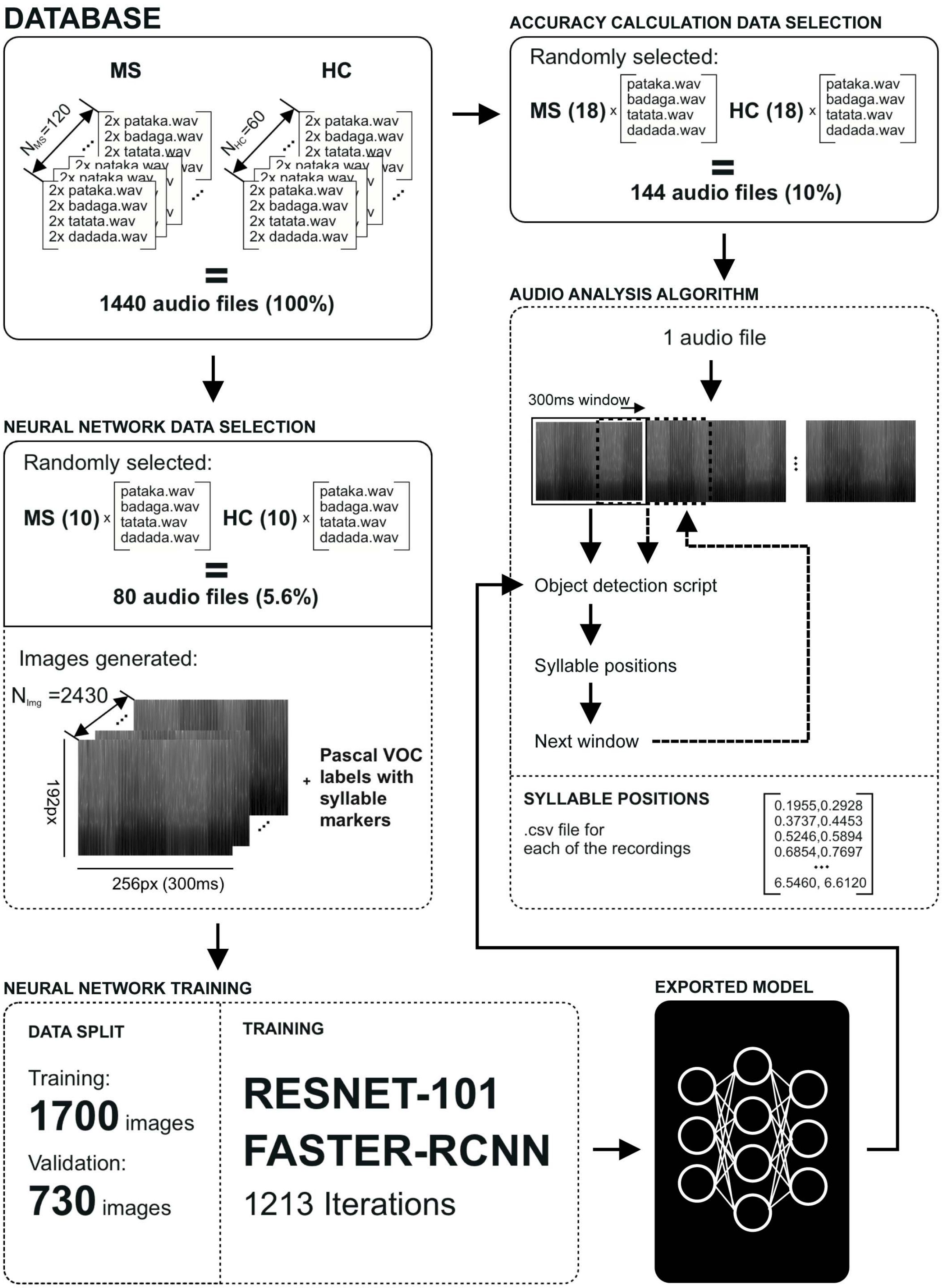

Fig. 2. Scheme depicting the process of data pooling, preparation of training/validation subsets and training of the CNN based on Restnet-101 topology with implemented Faster-RCNN data object model, as well construction of the CNN input and achieving the final classification result.

Using the obtained labels, syllabic centers were estimated as the mid time between detected consonant burst and vowel occlusion. This approach was preferred rather than selection of syllabic position with the highest energy, as the maximum energy is commonly at the beginning of voicing and such a label would not reflect the prolongation of vowels typical for ataxic dysarthria [8]. The DDK rate was calculated as the number of syllables per time and DDK regularity as the standard deviation of pause lengths between consecutive syllables. Figure 1 demonstrates the results of DDK rate and DDK regularity calculation across voiced and voiceless AMR task for an exemplary speech signal from MS and HC speaker.

E. Algorithm Performance Testing

The performance of the newly-developed CNN-based algorithm was compared with the algorithm based on traditional speech signal processing techniques by Novotny *et al.* [26], which is likely the most robust available from automated methods designed for the evaluation of oral diadochokinesis [24]–[26]. The method published by Novotny *et al.* [26] was designed for the detection of three events in each syllable including an initial burst of the stop consonant, the onset of voicing and occlusion. This method utilizes approaches of traditional signal processing including filtering the spectrogram for burst detection, Bayesian step change-point detection for voice onset detection and polynomial thresholding for the detection of an occlusion. The positions of the initial burst and occlusion were used for purposes of comparison.

To estimate the reliability of both automatic algorithms, each label obtained by the automatic algorithm was tested to determine whether it fits into the appropriate time interval between consonant burst and vowel occlusion, as was determined using manual labeling. A label that did not fit into an appropriate syllabic time interval was counted as an error. A syllabic time interval with no automatic label was counted as an error. Only one automatic label could be associated with one appropriate syllabic time interval, other automatic labels in the same interval were counted as errors. The overall percentage accuracy (ACC) of the algorithm for each utterance was calculated according to equation 2:

$$ACC = 100 - 100 \times \frac{N_{Err}}{N_{Manual}}, \tag{2}$$

where N_{Err} represents number of error detections by the algorithm and N_{Manual} represents number of syllables determined using manual annotation.

For the algorithm with the best ACC, three additional validation experiments were introduced and tested using a testing subset (i.e., 144 utterances from 36 participants representing 10 % of the entire dataset). First, the relationship between DDK features based on manual and automated labels was computed. Second, recordings were perceptually judged by the rate and regularity with a five-dimensional scoring system (0 indicating intact, rapid, and regular DDK performance and 4 indicating severe, slow, and irregular DDK performance). Third, we calculated the performance measure representing accuracy between manual and automated labels. This measure was defined as the percentage of cases in which the absolute temporal deviations between syllable positions obtained using manual and automated labels is lower than tolerance value; two tolerance values of 10 ms and 20 ms were considered. Additionally, inter-rater reliability based on the re-labeling of testing data by the second investigator was calculated using the same approach including two tolerance values of 10 ms and 20 ms; the correlations between resulting DDK features obtained by manual labels of two investigators were also performed.

TABLE II

ACCURACY OF ALGORITHMS ACROSS INDIVIDUAL GROUPS AND DDK PARADIGMS. MS = MULTIPLE SCLEROSIS, HC = HEALTHY CONTROLS, CNN = CONVOLUTIONAL NEURAL NETWORK, AMR = ALTERNATING MOTION RATES, SMR = SEQUENTIAL MOTION RATES

Accuracy across tasks	Unvoiced AMR	Voiced AMR	Unvoiced SMR	Voiced SMR	Average
	(%)	(%)	(%)	(%)	(%)
CNN-based approach					
MS	99.8	95.1	99.2	92.3	96.6
HC	99.3	99.4	100.0	96.9	98.9
All	99.6	97.3	99.6	94.6	97.8
Traditional digital speech signal processing approach (Novotny et al. [26])					
MS	79.3	64.1	83.4	37.0	66.0
HC	86.0	80.4	96.5	78.5	85.4
All	82.7	72.3	90.0	57.8	75.7

F. Statistics

To provide greater stability of speech assessment [37], final values of DDK features used for statistical analyses were calculated by averaging the data for each participant obtained in two vocal task runs. Both acoustic features were found to be normally distributed by the Kolmogorov-Smirnov test. Statistical analyses were performed using repeated measures analysis of variance (RM-ANOVA) with GROUP (MS vs. controls) treated as a between-group factor and TASK (/ta/ vs. /da/ vs. /pa/-/ta/-/ka/ vs. /ba/-/da/-/ga/) treated as a within-group factor. Bonferroni post-hoc significance was assessed for the effect of TASK. The Pearson correlation was applied to find relationships between speech variables. Relationships between speech and clinical parameters were tested using the Spearman correlation coefficient due to ordinal clinical scales; p-values were adjusted by False Discovery Rate correction for multiple comparisons by 16 correlations performed for each acoustic metric. The classification performance, including accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curve, between MS and HC group across individual DDK paradigms was calculated using binary logistic regression with leave-one-out cross-validation.

III. RESULTS

A. Algorithm Performance

Table II provides the overall classification accuracies of the tested algorithms across all investigated groups. The overall average classification accuracy of the CNN-based approach was 97.8%, with a performance of 96.6% for MS and 98.9% for HC. The average accuracy for voiceless AMR/SMR paradigms of 99.6% was slightly higher than the average accuracy of 96.0% for voiced AMR/SMR tasks. The algorithm by Novotny *et al.* [26] based on traditional digital speech signal processing reached an average accuracy of 75.7% with a performance of 66.0% for MS and 85.4% for HC. The

TABLE III
ACCURACY OF CNN-BASED ALGORITHM ACROSS INDIVIDUAL DDK PARADIGMS FOR GENDER, AGE, AND DYSARTHRIA SEVERITY. MS = MULTIPLE SCLEROSIS, AMR = ALTERNATING MOTION RATES, SMR = SEQUENTIAL MOTION RATES

Accuracy across tasks	Unvoiced AMR	Voiced AMR	Unvoiced SMR	Voiced SMR	Average
	(%)	(%)	(%)	(%)	(%)
Women	99.4	97.8	99.9	95.2	98.1
Men	100	95.3	98.5	92.5	96.5
Age $<$ 50 years	99.5	96.8	99.6	93.6	97.4
Age $>$ 50 years	99.8	98.5	99.5	97.2	98.8
None dysarthria in MS	99.5	98.2	99.5	95.2	98.1
Perceptible dysarthria in MS	100	91.7	100	90.8	95.6

average accuracy across voiceless AMR/SMR tasks of 86.4% was considerably higher than the average accuracy of 65.0% for voiced AMR/SMR paradigms. Table III lists the detailed classification accuracies for the CNN-based approach across gender, age, and dysarthria severity. It can be seen that algorithm performance is relatively consistent with average accuracy higher than 90% across all these different scenarios; the lowest scores were obtained for voiced paradigms in MS patients with perceptible dysarthria.

Considering additional validations for CNN-based algorithm across 10 % of testing data, correlation between features based on CNN and manual labels showed high reliability for DDK rate ($r = 0.94$, $p < 0.001$) as well as DDK regularity ($r = 0.77$, $p < 0.001$). In addition, we detected strong correlations between CNN-based features and perceptual judgments for DDK rate ($r = -0.66$, $p < 0.001$) with mean perceptible score of 0.88 (SD 0.84, range 0–3) for MS and 0.26 (SD 0.47, range 0–2) for HC as well as DDK regularity ($r = 0.63$, $p < 0.001$) with mean perceptible score of 1.00 (SD 0.90, range 0–3) for MS and 0.35 (SD 0.49, range 0–2) for HC. The algorithm performance in correct temporal detection of syllable position across all syllables was 87.8 % for 10 ms and 95.5 % for 20 ms tolerance value. Inter-rater reliability in correct temporal detection of syllable positions was 96.9 % for 10 ms and 99.3 % for 20 ms tolerance value; the correlations between DDK features based on manual labels of two investigators showed very high reliability for DDK rate ($r = 0.98$, $p < 0.001$) as well as DDK regularity ($r = 0.96$, $p < 0.001$).

B. Comparison of DDK Characteristics Between MS and HC Group

Figure 3 depicts the results of acoustic analyses for DDK rate and DDK regularity in MS and control subjects among different types of utterances. For DDK rate, RM-ANOVA showed a significant effect for GROUP ($F(1, 178) = 35.6$, $p < 0.001$, $\eta^2 = 0.17$), TASK ($F(3, 534) = 18.5$, $p < 0.001$, $\eta^2 = 0.09$), as well as GROUP × TASK ($F(1, 534) = 11.4$, $p < 0.001$, $\eta^2 = 0.06$). This interaction was different between MS and control groups with respect to various tasks performed. MS patients showed slower DDK rate for /ba/-/da/-/ga/ repetition compared to remaining tasks as well as slower SMR compared to AMR in general, whereas controls manifested only higher DDK rate for /pa/-/ta/-/ka/ repetition compared to remaining tasks. For DDK regularity, we detected significant effect for GROUP ($F(1, 178) = 27.2$, $p < 0.001$, $\eta^2 = 0.13$), TASK ($F(3, 534) = 116.0$, $p < 0.001$, $\eta^2 = 0.39$), as well as GROUP × TASK ($F(1, 534) = 2.9$, $p = 0.04$, $\eta^2 = 0.02$). This interaction was mainly associated with poorer performance in DDK regularity for /da/ compared to /pa/ in MS patients. Figure 4 shows ROC curves obtained by a combination of DDK rate and DDK regularity measures across different paradigms. The classification accuracy between MS and HC was 73.9% (sensitivity 65.9%, specificity 76.3%) for unvoiced AMR, 74.4% (sensitivity 65.9%, specificity 77.2%) for voiced AMR, 67.8% (sensitivity 52.8%, specificity 71.5%) for unvoiced SMR, 70.6% (sensitivity 60.0%, specificity 73.1%) for voiced SMR, and 73.9% (sensitivity 63.3%, specificity 77.9%) for combination of all DDK paradigms.

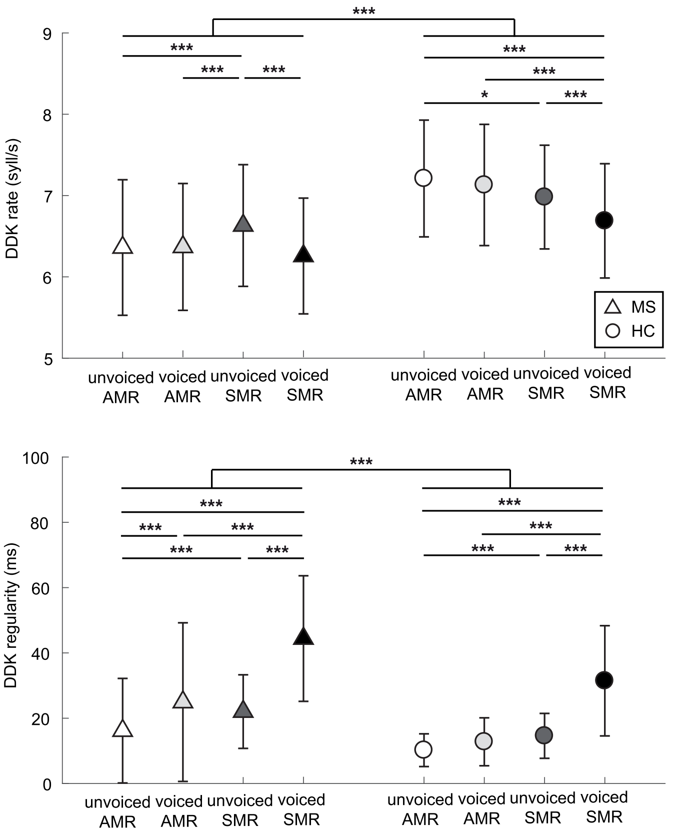

Fig. 3. Results of acoustic analyses across different DDK tasks. Group differences between MS and controls with * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, whereby the symbols represent mean values and error bars represent standard deviation. DDK = diadochokinetic, AMR = alternating motion rates, SMR = sequential motion rates, MS = multiple sclerosis, HC = healthy controls.

C. Association Between DDK Performance and Clinical Data in MS

Table IV shows the results of correlation analyses between DDK features, dysarthria severity, and neurological data. Weak to moderate significant negative correlations were observed between DDK rate across all types of utterances and EDDS

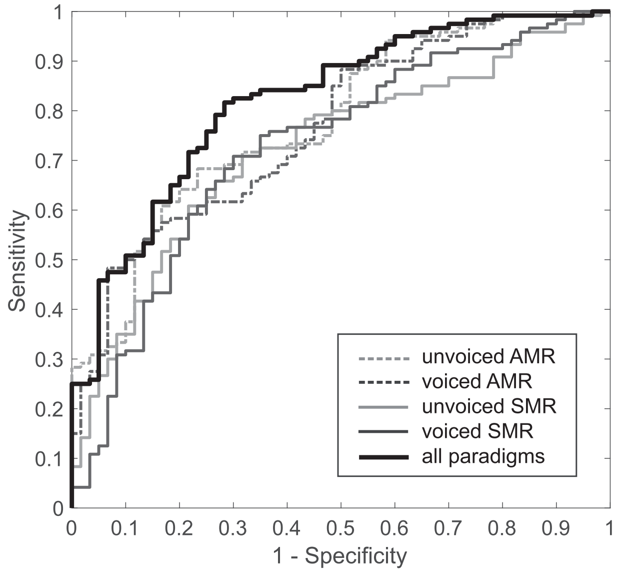

Fig. 4. ROC curves obtained between MS and HC groups using a combination of DDK rate and DDK regularity across different paradigms. AMR = alternating motion rates, SMR = sequential motion rates.

TABLE IV
Correlations Between Oral DDK Features, Dysarthria Severity and Neurological Data. Statistically Significant Differences:* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. EDSS = Expanded Disability Status Scale, AMR = Alternating Motion Rates, SMR = Sequential Motion Rates

	EDDS overall	EDDS pyramidal	EDDS cerebellar	Dysarthria severity
DDK rate				
Unvoiced AMR	-0.47***	-0.37***	-0.48***	-0.34***
Voiced AMR	-0.44***	-0.39***	-0.43***	-0.29**
Unvoiced SMR	-0.37***	-0.35***	-0.39***	-0.28**
Voiced SMR	-0.35***	-0.39***	-0.39***	-0.32***
DDK regularity				
Unvoiced AMR	0.18	0.07	0.21	0.25
Voiced AMR	0.11	0.07	0.08	0.16
Unvoiced SMR	0.14	0.08	0.17	0.14
Voiced SMR	0.19	0.14	0.18	0.12

score ($r < -0.35$, $p < 0.001$), EDSS pyramidal and cerebellar subscores ($r < -0.35$, $p < 0.001$), as well as dysarthria severity ($r < -0.28, p < 0.01$). DDK regularity showed only a trend toward correlation to the dysarthria severity, EDSS score and EDSS cerebellar subscore ($r > 0.17$, $p < 0.05$, uncorrected), while no trend toward relationship between DDK regularity and EDSS pyramidal subscore was observed. In addition, dysarthria severity was significantly correlated to all clinical measures including EDSS ($r = 0.49$, $p < 0.001$), EDSS pyramidal subscore ($r = 0.42$, $p < 0.001$), as well as EDSS cerebellar subscore ($r = 0.45$, $p < 0.001$).

IV. Discussion

The results of our work represent the first step toward the development of a fully CNN-based automated tool for robust task-independent evaluation of oral diadochokinesis. The algorithm we developed was able to correctly identify the position of individual syllables with a very high average accuracy of 97.8 %, with the performance in correct temporal detection of syllable position of 87.8 % for 10 ms and 95.5 % for 20 ms tolerance value. Although the severity of dysarthria was imperceptible in 70 % and mild to moderate in only 30 % of our MS cohort, we found significantly altered DDK rate and regularity across all different types of investigated tasks. To the best of our knowledge, this finding was based on the largest series of data concerning oral diadochokinesis in MS. The capability of the introduced models led to the classification accuracy of up to 74 % in discriminating between MS and HC subjects, confirming the importance of oral diadochokinesis for motor speech assessment in MS. Since the estimated prevalence of dysarthria in MS is 40-60 % [3]–[5], this classification accuracy is very promising and even comparable to 79 % accuracy obtained previously using multiple types of speaking tasks [5].

Our findings of slow DDK rate in MS are in general agreement with previous research demonstrating slow AMR and SMR in speakers with spastic and ataxic dysarthria [17], [18], [21], [22]. Interestingly, performance in DDK rate across all tasks showed weak to moderate correlations to both spastic and cerebellar EDSS subscores, supporting the hypothesis that damage to both the cerebellum and pyramidal tract may lead to slower speaking rate. In accordance with a previous study [22], no differences in DDK rate were generally found between AMR and SMR tasks for healthy controls, although they performed best on the unvoiced SMR paradigm. This may be due to the inclusion of all bilabial, alveolar and velar places of articulation compared to the same alveolar movements during single syllable /ta/ repetition, which requires more difficult tongue movement [26]. Conversely, MS patients were slower during the both SMR paradigms, but slowest in voiced SMR, indicating a possible divergent pattern of articulation abnormalities for voiced and unvoiced consonants.

Indeed, this phenomenon is more notable in DDK regularity, where MS patients manifested worse performance in both voiced tasks compared to their voiceless cognates. The average performance in voiced AMR was lower than in the voiceless SMR task in MS. Surprisingly, even the healthy control group showed poorer performance in DDK regularity for voiced SMR compared to remaining tasks, indicating more difficult timing control in general if the SMR paradigm is combined with voiced plosives. The accurate production of stop consonants requires close coordination between the larynx and the articulators. As the production of voiced consonants is characterized by voicing lead at the beginning followed by a period of articulatory closure, insufficient programming in advance of speech onset may lead to additional disruption of coordination between the larynx and the articulators, which has to be more precise in voiced compared to voiceless plosives. Our pilot analyses showed a trend toward correlation ($p < 0.05$, uncorrected) between the performance of DDK regularity and the extent of cerebellar but not pyramidal dysfunction, which may support the hypothesis that temporal irregularity of syllable repetition is primarily attributable to damage to the cerebellum [19], [20], [38]. However, cerebellar assessment in MS using the EDSS is a limited method as the scores are driven mostly by coarse tremor and gait ataxia while the cerebellar problems such as dysarthria are scored in the EDSS brainstem subscore. Therefore, future research based on more robust scoring of cerebellar dysfunction such as using the magnetic resonance imaging or the clinical Scale for the Assessment and Rating of Ataxia is warranted to ascertain the contribution of cerebellar atrophy to diadochokinetic irregularity.

Our newly designed CNN-based algorithm reached a very high average accuracy of 98.9 % for healthy controls and 96.6 % for MS subjects. In addition, the correlations between DDK features based on automated labels, manual labels, and perceptual judgments were sufficiently high. Thus, our algorithm substantially outperformed the conventional method based on traditional digital speech signal processing by Novotny et al. [26], which reached an average accuracy of only 85.4 % for healthy controls and 66.0 % for MS. Nevertheless, it is important to note that this algorithm was designed only for an unvoiced SMR paradigm and tested in patients with Parkinson's disease, which do not typically manifest substantial temporal oral diadochokinesis irregularity like speakers with ataxic dysarthria. Indeed, when considering the performance for unvoiced SMR and healthy speakers, the algorithm by Novotny *et al.* [26] reached a very high performance of 96.5 %. However, upon closer inspection, we found that even the accuracy of our new CNN-based algorithm reached a lower accuracy of 97.3 % and 94.6 % for voiced AMR and SMR paradigms compared to 99.6 % accuracy for both voiceless AMR and SMR tasks. This is likely due to the fact that the presence of unvoiced consonants makes the separation between consecutive syllables more obvious while the occurrence of voicing to a certain extent masks the beginning of consonant (see Fig. 1). Therefore, the segmentation of individual syllables becomes more challenging in voiced DDK tasks as there is not such a sharp transition between consecutive syllables in the spectrogram. Also, the severity of dysarthria likely plays a role in the algorithm performance as the lowest scores of 91.7 % and 90.8 % were found for voiced AMR and SMR paradigms in MS patients with perceptible dysarthria.

The current study has certain limitations. The observed correlations between DDK measures and clinical evaluation based on EDSS scores as well as dysarthria severity are rather weak. Our cohort was generally in lower stages of disease (EDSS mean 3.6, range 1–6.5) with 70 % of MS patients without the occurrence of perceptible dysarthria. In addition, from 36 MS individuals with some form of dysarthria, only 3 had moderate dysarthria while 33 manifested mild dysarthria, making it difficult to estimate the discrimination accuracy of DDK measures with respect to dysarthria severity. We may thus hypothesize that the magnitude of correlations would be greater in samples with a wider range of dysarthria severity. Such an assumption needs to be verified in future studies. Our findings related to voiced DDK paradigms may be language-specific as the articulation of voiced consonants is characterized by different voicing lead in various languages [39]. In most languages, voice onset times for voiced and voiceless stops are in discrete duration ranges that correspond to one of three voicing categories including long negative, short and long positive voice onset time [39]. Therefore, due to the presence of dysarthria, the plosives with short voice onset time duration may be unchanged or extended beyond normal, while the plosives with both positive and negative long voice onset time tend to be reduced [40]. As a result, languages where the voiced stops are characterized by long negative voice onset time, such as the Czech language [23], may be more sensitive to cerebellar dysfunction. However, voice onset time extracted from single syllables is particularly sensitive to changes due to cerebellar atrophy [23], [41], while no alterations were found in voice onset time obtained from more complex speech utterance [42], suggesting certain generalizability of our findings to different languages. Moreover, the generalization capability of oral diadochokinesis has already been verified in three different languages including German, Spanish, and Czech [43]. We did not employ the entire dataset for training/validation of our CNN-based algorithm. Nevertheless, even with only 80 utterances (5.6 %) used for training and fine-tuning, the accuracy of the algorithm significantly outperformed available state-of-the-art detector [26]. This generally supports the suitability of the proposed method for evaluation of oral diadochokinesis with the possibility to further strengthen the accuracy if necessary by including more training data. The accuracy of the algorithm was validated using 144 utterances (10 %), which represents the typical amount of data used for validation in similar applications [44] and approximately double the size used for validation in the previous studies (54 and 80 utterances) aimed at automated assessment of oral diadochokinesis [25], [26].

V. Conclusion

The present study provides a novel extension of available methods for the automatic evaluation of oral diadochokinesis. Objective evaluation of AMR and SMR tasks may aid in differential diagnosis and may provide markers of treatment efficacy and disease progression. We envisage that our results will support a continuum of technological solutions for the automated assessment of various dysarthric features, which may improve the quality of life of patients with neurodegenerative diseases. Further research is needed to elaborate on our findings in different languages and other types of dysarthria. Future studies are also encouraged to compare technologies designed specifically for the automatic evaluation of oral diadochokinesis with other already existing approaches to detect syllables with particular sounds (phonological units) [45], [46].

Acknowledgment

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] N. Koch-Henriksen and P. S. Sørensen, "The changing demographic pattern of multiple sclerosis epidemiology," *Lancet Neurol.*, vol. 9, no. 5, pp. 520–532, 2010. doi: 10.1016/S1474-4422(10)70064-8.

[2] G. Noffs *et al.*, "What speech can tell us: A systematic review of dysarthria characteristics in multiple sclerosis," *Autoimmunity Rev.*, vol. 17, pp. 1202–1209, Dec. 2018. doi: 10.1016/j.autrev.2018.06.010.

[3] R. M. Merson and M. I. Rolnick, “Speech-language pathology and dysphagia in multiple sclerosis,” *Phys. Med. Rehabil. Clinics*, vol. 9, pp. 631–641, Aug. 1998. doi: 10.1016/S1047-9651(18)30254-7.

[4] L. Hartelius, B. Runmarker, and O. Andersen, “Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: Relation to neurological data,” *Folia phoniatrica Logopaedica*, vol. 52, no. 4, pp. 160–177, 2000. doi: 10.1159/000021531.

[5] J. Rusz *et al.*, “Characteristics of motor speech phenotypes in multiple sclerosis,” *Multiple Sclerosis Related Disorders*, vol. 19, pp. 62–69, Jan. 2018. doi: 10.1016/j.msard.2017.11.007.

[6] D. R. Beukelman, G. H. Kraft, and J. Freal, “Expressive communication disorders in persons with multiple sclerosis: A survey,” *Arch. Phys. Med. Rehabil.*, vol. 66, no. 10, pp. 675–677, 1985.

[7] K. M. Yorkston *et al.*, “Characteristics of multiple sclerosis as a function of the severity of speech disorders,” *J. Med. Speech Lang. Pathol.*, vol. 11, no. 2, pp. 73–84, 2003.

[8] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 3rd ed. St. Louis, MO, USA: Elsevier, 2013.

[9] F. L. Darley, J. R. Brown, and N. P. Goldstein, “Dysarthria in multiple sclerosis,” *J. Speech Hearing Res.*, vol. 15, no. 2, pp. 229–245, 1972. doi: 10.1044/jshr.1502.229.

[10] L. Hartelius, E. H. Buder, and E. A. Strand, “Long-term phonatory instability in individuals with multiple sclerosis,” *J. Speech Lang. Hearing Res.*, vol. 40, pp. 1056–1072, Oct. 1997. doi: 10.1044/jslhr.4005.1056.

[11] A. V. Feijó, M. A. Parente, M. Behlau, S. Haussen, M. C. De Veccino, and B. C. de Faria Martignago, “Acoustic analysis of voice in multiple sclerosis patients,” *J. Voice*, vol. 18, no. 3, pp. 341–347, 2004. doi: 10.1016/j.jvoice.2003.05.004.

[12] K. Konstantopoulos, M. Vikelis, J. A. Seikel, and D. D. Mitsikostas, “The existence of phonatory instability in multiple sclerosis: An acoustic and electroglottographic study,” *Neurol. Sci.*, vol. 31, no. 3, pp. 259–268, 2010. doi: 10.1007/s10072-009-0170-3.

[13] L. Hartelius, B. Runmarker, P. Andersen, and L. Nord, “Temporal speech characteristics of individuals with multiple sclerosis and ataxic dysarthria: ’Scanning speech’revisited,” *Folia phoniatrica Logopaedica*, vol. 52, no. 5, pp. 228–238, 2000. doi: 10.1159/000021538.

[14] K. Tjaden and V. Martel-Sauvageau, “Consonant acoustics in Parkinson’s disease and multiple sclerosis: Comparison of clear and loud speaking conditions,” *Amer. J. Speech Lang. Pathol.*, vol. 26, no. 2S, pp. 569–582, 2017. doi: 10.1044/2017_AJSLP-16-0090.

[15] M. Fazeli *et al.*, “Dysphonia characteristics and vowel impairment in relation to neurological status in patients with multiple sclerosis,” *J. Voice*, to be published. doi: 10.1016/j.jvoice.2018.09.018.

[16] J. Rusz *et al.*, “Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson’s disease,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1495–1507, Aug. 2018. doi: 10.1109/TNSRE.2018.2851787.

[17] Y. Ozawa, O. Shiromoto, F. Ishizaki, and T. Watamori, “Symptomatic differences in decreased alternating motion rates between individuals with spastic and with ataxic dysarthria: An acoustic analysis,” *Folia Phoniatrica Logopaedica*, vol. 53, no. 2, pp. 67–72, 2001. doi: 10.1159/000052656.

[18] H. Ackermann, I. Hertrich, and T. Hehr, “Oral diadochokinesis in neurological dysarthrias,” *Folia Phoniatrica Logopaedica*, vol. 47, no. 1, pp. 15–23, 1995. doi: 10.1159/000266338.

[19] W. Ziegler and K. Wessel, “Speech timing in ataxic disorders: Sentence production and rapid repetitive articulation,” *Neurology*, vol. 47, no. 1, pp. 208–214, 1996. doi: 10.1212/WNL.47.1.208.

[20] H. Ackermann and I. Hertrich, “The contribution of the cerebellum to speech processing,” *J. Neurolinguistics*, vol. 13, pp. 95–116, Jul. 2000. doi: 10.1016/S0911-6044(00)00006-3.

[21] L. Hartelius and M. Lillvik, “Lip and tongue function differently affected in individuals with multiple sclerosis,” *Folia Phoniatrica Logopaedica*, vol. 55, no. 1, pp. 1–9, 2003. doi: 10.1159/000068057.

[22] K. Tjaden and E. Watling, “Characteristics of diadochokinesis in multiple sclerosis and Parkinson’s disease,” *Folia Phoniatrica Logopaedica*, vol. 55, no. 5, pp. 241–259, 2003. doi: 10.1159/000072155.

[23] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, and E. Ruzicka, “Distinct patterns of imprecise consonant articulation among Parkinson’s disease, progressive supranuclear palsy and multiple system atrophy,” *Brain Lang.*, vol. 165, pp. 1–9, Feb. 2017. doi: 10.1016/j.bandl.2016.11.005.

[24] Y.-T. Wang, R. D. Kent, J. R. Duffy, and J. E. Thomas, “Analysis of diadochokinesis in ataxic dysarthria using the motor speech profile program,” *Folia Phoniatrica Logopaedica*, vol. 61, no. 1, pp. 1–11, 2009. doi: 10.1159/000184539.

[25] D. Montaña, Y. Campos-Roca, and C. J. Pérez, “A Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson’s disease,” *Comput. Method Programs Biomed.*, vol. 154, pp. 89–97, Feb. 2018. doi: 10.1016/j.cmpb.2017.11.010.

[26] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, “Automatic evaluation of articulatory disorders in Parkinson’s disease,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014. doi: 10.1109/TASLP.2014.2329734.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[28] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. doi: 10.1109/TPAMI.2016.2577031.

[29] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, “Multimodal assessment of Parkinson’s disease: A deep learning approach,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019. doi: 10.1109/JBHI.2018.2866873.

[30] C. H. Polman *et al.*, “Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria,” *Ann. Neurol.*, vol. 69, no. 2, pp. 292–302, 2011. doi: 10.1002/ana.22366.

[31] J. F. Kurtzke, “Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS),” *Neurology*, vol. 33, no. 11, pp. 1444–1452, 1983. doi: 10.1212/WNL.33.11.1444.

[32] F. L. Darley, A. E. Aronson, and J. R. Brown, “Differential diagnostic patterns of dysarthria,” *J. Speech Hearing Res.*, vol. 12, no. 2, pp. 246–269, 1969. doi: 10.1044/jshr.1202.246.

[33] G. Weismer, “Motor speech disorders,” in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds., 3rd ed. Cambridge, MA, USA: Blackwell, 1997.

[34] S. Šimácková, V. J. Podlipský, and K. Chládková, “Czech spoken in Bohemia and Moravia,” *J. Int. Phonetic Assoc.*, vol. 42, no. 2, pp. 225–232, 2012. doi: 10.1017/S0025100312000102.

[35] M. Abadi *et al.*, “TensorFlow: A system for large-scale machine learning,” in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[37] A. P. Vogel, J. Fletcher, P. J. Snyder, A. Fredrickson, and P. Maruff, “Reliability, stability, and sensitivity to change and impairment in acoustic measures of timing and frequency,” *J. Voice*, vol. 25, no. 2, pp. 137–149, 2011. doi: 10.1016/j.jvoice.2009.09.003.

[38] K. A. Spencer and M. A. Rogers, “Speech motor programming in hypokinetic and ataxic dysarthria,” *Brain Lang.*, vol. 94, pp. 347–366, Sep. 2005. doi: 10.1016/j.bandl.2005.01.008.

[39] P. Auzou, C. Ozsancak, J. R. Morris, M. Jan, F. Eustache, and D. Hannequin, “Voice onset time in aphasia, apraxia of speech and dysarthria: A review,” *Clin. Linguistics Phonetics*, vol. 14, no. 2, pp. 131–150, 2000. doi: 10.1080/026992000298878.

[40] R. D. Kent, G. Weismer, J. F. Kent, H. K. Voperian, and J. R. Duffy, “Acoustic studies of dysarthric speech: Methods, progress, and potential,” *J. Commun. Disorders*, vol. 32, no. 3, pp. 141–186, 1999. doi: 10.1016/S0021-9924(99)00004-0.

[41] H. Ackermann and I. Hertrich, “Voice onset time in ataxic dysarthria,” *Brain Lang.*, vol. 56, pp. 321–333, Feb. 1997. doi: 10.1006/brln.1997.1740.

[42] H. Ackermann and I. Hertrich, “Dysarthria in Friedreich’s ataxia: Timing of speech segments,” *Clin. Linguistics Phonetics*, vol. 7, no. 1, pp. 75–91, 1993. doi: 10.3109/02699209308985545.

[43] J. R. Orozco-Arroyave *et al.*, “Automatic detection of Parkinson’s disease in running speech spoken in three different languages,” *J. Acoust. Soc. Amer.*, vol. 139, no. 1, pp. 481–500, 2016. doi: 10.1121/1.4939739.

[44] S. L. Salzberg, “On comparing classifiers: Pitfalls to avoid and a recommended approach,” *Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 317–328, 1997. doi: 10.1023/A:1009752403260.

[45] M. Cernak, B. Potard, and P. N. Garner, “Phonological vocoding using artificial neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2015, pp. 4844–4848.

[46] M. Cernak and P. N. Garner, “PhonVoc: A phonetic and phonological vocoding toolkit,” in *Proc. Interspeech*, 2016, pp. 988–992.