frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 23 July 2024
DOI 10.3389/fdgth.2024.1440986

OPEN ACCESS

EDITED BY
Graham Jones,
Tufts Medical Center, United States

REVIEWED BY
Daragh Heitzman,
Texas Neurology, United States
Dale Joachim,
Sonde Health, Inc., United States

*CORRESPONDENCE
Johannes Tröger
johannes.troeger@ki-elements.de

RECEIVED 30 May 2024
ACCEPTED 05 July 2024
PUBLISHED 23 July 2024

CITATION
Tröger J, Dörr F, Schwed L, Linz N, König A, Thies T, Orozco-Arroyave JR and Rusz J (2024) An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders. Front. Digit. Health 6:1440986. doi: 10.3389/fdgth.2024.1440986

An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders

Johannes Tröger[1*], Felix Dörr[1], Louisa Schwed[1], Nicklas Linz[1], Alexandra König[1,2,3], Tabea Thies[4,5], Juan Rafael Orozco-Arroyave[6,7] and Jan Rusz[8]

[1]ki elements GmbH, Saarbrücken, Germany, [2]Cobtek (Cognition-Behaviour-Technology) Lab, University Côte d'azur, Nice, France, [3]Centre de Mémoire de Ressources et de Recherche, Centre Hospitalier Universitaire Nice (CHUN), Nice, France, [4]Department of Neurology, Faculty of Medicine and University Hospital Cologne, Cologne, Germany, [5]IfL Phonetics, Faculty of Arts and Humanities, University of Cologne, Cologne, Germany, [6]GITA Lab, Faculty of Engineering, University of Antioquia, Medellín, Colombia, [7]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, [8]Department of Circuit Theory, Czech Technical University in Prague, Prague, Czechia

Introduction: Dysarthria, a motor speech disorder caused by muscle weakness or paralysis, severely impacts speech intelligibility and quality of life. The condition is prevalent in motor speech disorders such as Parkinson's disease (PD), atypical parkinsonism such as progressive supranuclear palsy (PSP), Huntington's disease (HD), and amyotrophic lateral sclerosis (ALS). Improving intelligibility is not only an outcome that matters to patients but can also play a critical role as an endpoint in clinical research and drug development. This study validates a digital measure for speech intelligibility, the ki: SB-M intelligibility score, across various motor speech disorders and languages following the Digital Medicine Society (DiMe) V3 framework.
Methods: The study used four datasets: healthy controls (HCs) and patients with PD, HD, PSP, and ALS from Czech, Colombian, and German populations. Participants' speech intelligibility was assessed using the ki: SB-M intelligibility score, which is derived from automatic speech recognition (ASR) systems. Verification with inter-ASR reliability and temporal consistency, analytical validation with correlations to gold standard clinical dysarthria scores in each disease, and clinical validation with group comparisons between HCs and patients were performed.
Results: Verification showed good to excellent inter-rater reliability between ASR systems and fair to good consistency. Analytical validation revealed significant correlations between the SB-M intelligibility score and established clinical measures for speech impairments across all patient groups and languages. Clinical validation demonstrated significant differences in intelligibility scores between pathological groups and healthy controls, indicating the measure's discriminative capability.
Discussion: The ki: SB-M intelligibility score is a reliable, valid, and clinically relevant tool for assessing speech intelligibility in motor speech disorders. It holds promise for improving clinical trials through automated, objective, and scalable assessments. Future studies should explore its utility in monitoring disease progression and therapeutic efficacy as well as add data from further dysarthrias to the validation.

KEYWORDS

amyotrophic lateral sclerosis (ALS), Huntington's disease (HD), Parkinson's disease (PD), progressive supranuclear palsy (PSP), speech analysis, intelligibility, digital biomarkers

Introduction

Dysarthria is a motor speech disorder resulting from weakness or paralysis of speech-related muscles (1). It leads to decreased speech intelligibility, frequent communication breakdowns, and a reduced quality of life. Speech intelligibility is reduced in many types of dysarthria, including typical Parkinson's Disease (PD) (2–5), atypical parkinsonism such as progressive supranuclear palsy (PSP) (4, 6, 7), Huntington's disease (HD) (8, 9), amyotrophic lateral sclerosis (ALS) (1, 10), and multiple sclerosis (MS) (11, 12).

Reduced intelligibility of patients' speech often leads to communication difficulties and affects social participation and quality of life in general (13, 14). Hence, communication deficits and perceived intelligibility of their speech represents a major concern for patients with motor speech disorders (15, 16). Speech intelligibility is a construct depending on (a) a speaker (sender) who produces an acoustic signal within, e.g., conversational speech, and (b) a listener (receiver) who receives the signal and interprets it; the success of the interpretation is a direct function of the intelligibility (17) (see also Figure 1). Although a major concern, speech intelligibility is not necessarily dependent on disease severity, duration, or motor phenotype and patients' own perceptions of the severity do not necessarily reflect objective measures (18). Improved intelligibility is often a primary goal of speech therapy, especially for individuals with dysarthria, and can be a valuable endpoint for clinical research and drug development (19).

Accordingly, measuring speech intelligibility is a clinically relevant assessment for monitoring a dysarthric patient's status and tracking the effectiveness of treatments (20). The common method for assessing speech intelligibility is perceptual evaluation by trained personnel—often clinicians. Standard clinical assessments for disorders associated with dysarthria, such as the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) (21), the Unified Huntington's Disease Rating Scale (UHDRS), and the revised amyotrophic lateral sclerosis functional rating scale (ALSFRS-R) (22), are based on clinician-rated questionnaires and assess, among other symptoms, speech intelligibility. However, these assessments require patient and clinician presence and can be subject to observer bias, pointing to a need for more objective automated methods for assessing speech disorders.

As the field of automated speech analysis is growing in clinical research and healthcare applications, there is increasing potential for digital automatic assessments of speech-related symptoms in motor speech disorders (23, 24). Digital dysarthria assessments are better suited for automated patient-administered screening or stratification at low cost to accelerate clinical trials (24–26). Furthermore, a high level of automation can easily scale up outreach to draw unbiased and representative trial populations beyond established clinical sites and hospital networks. In addition, within clinical trials, digital markers deliver objective high-frequency data to guide interventional clinical trial decision-making and make evaluation more efficient (27).

Previous studies have demonstrated how commercially available automatic speech recognition (ASR) systems could be a feasible platform for automatic measures of intelligibility in patients with motor speech disorders (19, 28). As commercial ASR systems are developed majorly on typical—presumably non-dysarthric—speech, the recognition accuracy of such a system should be an inverse model of the intelligibility of the speaker (29–31).

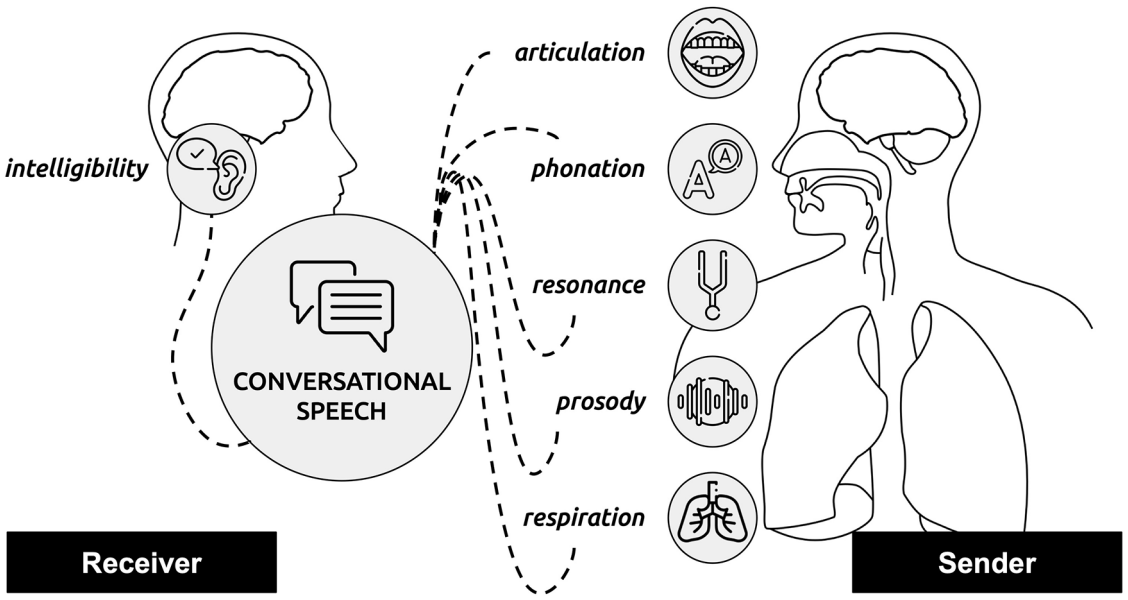

FIGURE 1
Conceptual model of intelligibility; being a receiver/listener-focused measure and being affected by impaired speech subsystems underlying dysarthrias within the sender: articulation, phonation, resonance, prosody, and respiration.

However, although promising results have been published in feasibility studies, there has not been any comprehensive validation work including multiple pathologies and multiple languages and following a systematic validation framework. The Digital Medicine Society (DiMe) V3 framework (verification, analytical validation, and clinical validation) (32–34) defines validation cases that digital measures should comply with to be considered fit-for-purpose for clinical trials and eventually medical devices, such as digital diagnostics. This framework has gained in importance in recent years and can be regarded as an industry standard for digital measures in this field.

In this study, we present a validation following the DiMe V3 framework for a digital measure for intelligibility, the ki: speech biomarker score for motor speech disorders intelligibility (ki: SB-M intelligibility score). We validate the SB-M intelligibility score in individuals with motor speech disorders, including PD, PSP, HD, and ALS, in multiple languages, including German, Czech, and Colombian Spanish, representing the Germanic, Slavic, and Romance language families.

Methods

Data

Four different datasets were used in the analysis: (1) Czech data from $N = 39$ patients with HD (35), $N = 43$ patients with PD (36), $N = 16$ patients with ALS (37), $N = 17$ patients with PSP (6), and $N = 46$ healthy controls (HCs); (2) Colombian data from $N = 50$ HCs and $N = 50$ patients with PD (38); and (3) German data (39) from $N = 98$ patients with PD. For detailed information on the initial cohorts, reading texts, and data collection process, we refer to the initial publications cited; however, for better readability for this manuscript, a short description will be given in the following sections. Compare also Table 1.

Czech data

Participants read an 80-word long paragraph in the respective language, which was phonemically balanced and well-established in clinical research (3). Recordings were conducted in a quiet room with low ambient noise, using a condenser microphone placed approximately 15 cm from the subject's mouth. Each participant had one recording session with the speech-language pathologist, without time limits. Participants were briefed on the speaking tasks and recording process. Each participant provided written informed consent. The collection of the Czech data was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic (approval number 6/15 Grant GACR VFN).

Colombian data

Participants read 10 sentences of increasing complexity (38). Recordings were collected in a soundproof booth at the Clinica Noel in Medellin, Colombia, using a dynamic omnidirectional microphone and a professional audio card. This study was in compliance with the Helsinki Declaration and was approved by the ethics committee of the Clinica Noel in Medellin, Colombia. Written informed consent was signed by each participant.

German data

Participants read an 80-word long paragraph in the respective language, which was phonemically balanced, well-established, and taken from the German protocol version of the Dysarthria Analyzer (40). Speech data were collected in the Department of Neurology of the University Hospital Cologne in a room with low ambient noise using a condenser microphone headset to keep the mouth-to-microphone distance constant at approximately 7 cm from the mouth. Each participant provided written informed consent. The data collection was approved by the local ethics committee (protocol code: 23-1461-retro).

After the reading task, patients in all three cohorts underwent a range of clinical assessments (different for each study and cohort), of which the following are important for this study: the MDS-UPDRS (21), UHDRS (41), Natural History and Neuroprotection in Parkinson Plus Syndromes—Parkinson Plus Scale (NNIPPS) (42), and ALSFRS-R (22).

Automatic speech recognition and intelligibility score

To calculate the automatic intelligibility scores, we first ran the audios from the reading passage and reading sentences (in Colombian Spanish) through SIGMA the ki: proprietary speech processing library, which—besides other preprocessing and feature extraction steps—also interfaces with commercially available ASR systems; for verification, we selected two different

TABLE 1 Demographic information of the samples and as essential clinical information.

	German	Colombian		Czech				
	PD DE	PD CO	HCs CO	PD CZ	HD CZ	PSP CZ	ALS CZ	HCs CZ
N	98 (32 F)	50 (25 F)	50 (25 F)	43 (19 F)	39 (20 F)	17 (6 F)	16 (11 F)	46 (21 F)
Age (years)	62.7 (±8.23)	61.02(±9.44)	60.98 (±9.46)	63.0 (±9.92)	48.28 (±13.4)	66.76 (±4.8)	60.0 (±10.66)	51.54 (±14.05)
MDS-UPDRS, UHDRS, NNIPPS, ALSFRS-R	37.43 (±10.89)	37.66 (±18.32)	—	20.88 (±10.92)	26.51 (±11.47)	67.12 (±26.7)	35.06 (±6.97)	—
Clinical scale speech items	0.80 (±0.90)	1.34 (±0.82)	—	0.81 (±0.63)	0.81 (±0.46)	1.88 (±0.7)	2.75 (±0.86)	—
ki: SB-M intelligibility score	0.82 (±0.18)	0.73 (±0.18)	0.86 (±0.11)	0.81 (±0.07)	0.67 (±0.17)	0.54 (±0.28)	0.58 (±0.29)	0.85 (±0.04)

CO, Colombian Spanish; CZ, Czech; DE, German.
ALSFRS-R: note that ALSFRS-R has an inverse relationship to disease severity, unlike the other scales where higher scores mean greater severity. Clinical scale speech items: MDS-UPDRS item 3.1, UHDRS dysarthria score, NNIPPS speech item, ALSFRS-R speech item from the bulbar score.

providers: Google Speech API (43) and Amazon Transcribe (44). Based on the transcripts and the target reading texts, we calculated the word error rate (WER, error between the number of target words in the reading text and that in the ASR transcripts) and word accuracy (WA, similar to [28]). From those raw measures, we then derived an automatic proxy for the intelligibility of the speech—the ki: SB-M intelligibility score.

V3 framework

The V3 framework established by the DiMe Society (32) provides a unified evaluation framework for digital measures. V3 includes three distinct phases in sequential order: verification, analytical validation, and clinical validation. For all the three phases, different data have to be collected and statistically analyzed to provide the necessary results.

Verification

Verification entails the systematic evaluation of sample-level sensor outputs against prespecified criteria. The ki: SB-M intelligibility score relies on ASR. Therefore, the most critical part of the sensor output and preprocessing pipeline is the automatic transcription of speech. The ki: SB-M intelligibility score uses a proprietary speech processing pipeline leveraging commercial ASR providers. To verify the performance at this stage, we calculated intraclass correlation coefficients (ICCs) for the WER and SB-M intelligibility score between Google and Amazon ASR. Previous studies and our own work have shown that error rates on a low level, such as phoneme error rate, do not necessarily model losses of perceptual intelligibility (45). We performed verification across the whole data sets except for the German PD data due to a lack of consent from patients.

In addition, we computed ICCs between repeated tests for data sets in which participants performed two repeated reading passages (all CZ data sets). Although tests are executed in quick succession, this can provide first insights into the retest reliability of the measures. Based on the current state of the art in the field, we considered an ICC of 0.40 (fair correlation) acceptable for verification (46).

Analytical validation

Analytical validation evaluates performance to measure a certain concept of interest (similar to construct validity). The ki: SB-M intelligibility score is related to speech impairments resulting in reduced speech intelligibility. For the analytical validation, we compared the ki: SB-M intelligibility score against established clinical anchor measures for speech impairments or dysarthria in the respective populations. Depending on the pathology, these measures differ: PD → MDS-UPDRS → speech item, HD → UHDRS → dysarthria item, PSP → NNIPPS → speech item, and ALS → ALSFRS-R → speech item (please note that in direct comparison with the other clinical scales, the ALSFRS-R has an inverse relationship to disease severity, meaning patients lose points as the disease progresses). For the comparison with the clinical anchors, we computed Spearman's rank correlation coefficient between the ki: SB-M intelligibility score and the respective speech impairment measure.

Clinical validation

Clinical validation evaluates the ability to validly measure clinically meaningful change within an intended scenario, including a specified clinical population. The ki: SB-M intelligibility score is built to measure clinically meaningful change in the intelligibility of speech in dysarthrias. To cover a significant range of dysarthrias, we included clinical validation on the following pathologies: PD, HD, PSP, and ALS.

We performed Kruskal–Wallis test group comparisons in the ki: SB-M intelligibility score between the different diagnostic groups (HC vs. pathology). In addition, we analyzed Spearman's rank correlation between the ki: SB-M intelligibility score and the respective global clinical staging measure: MDS-UPDRS, UHDRS, NNIPPS, and ALSFRS-R.

Results

Verification

For verification of the SB-M intelligibility score, we report reliability between the SB-M intelligibility score based on two different ASR methods and reliability between successive performances of the reading task and calculation of the SB-M intelligibility score.

Inter-rater reliability for ASRs

We compared different ASRs (Google and Amazon) as the basis for the SB-M intelligibility score. For most of the pathological groups, the ICC between both ASR methods showed a good to excellent performance (ICC equal or above 0.30). However, for Colombian PD data, the ICC was only fair and for Czech PD poor; both were still highly significant. The overall HC ICC (across all languages) was also only poor. For details, compare Table 2. WERs showed similar trends to the final intelligibility score, with the following pattern: HCs < PD < HC, PSP = ALS.

Consistency

Consistency over a short period of time (i.e., the same day in the same assessment reading the paragraph twice) was calculated based on repeated paragraph reading in all groups except the Colombian group, which read multiple sentences of increasing difficulty and not one overall homogenous paragraph. The ICCs for consistency were above 0.70, representing a good to excellent agreement. Compare also Table 2.

Analytical validation

For the analytical validation, we compared the ki: SB-M intelligibility score against established clinical anchor measures for speech impairments or dysarthria in the respective

TABLE 2 Agreement between two different ASR methods—Google Speech API and Amazon Transcribe—and the resulting SB-M intelligibility score and raw word error rate.

	HC overall	HC CZ	HC CO	PD CO	PD CZ	HD CZ	PSP CZ	ALS CZ
Google SB-M intelligibility score	0.862 (0.182)	0.853 (0.039)	0.859 (0.200)	0.733 (0.273)	0.810 (0.073)	0.675 (0.173)	0.537 (0.281)	0.590 (0.283)
Amazon SB-M intelligibility score	0.968 (0.088)	0.900 (0.041)	0.980 (0.090)	0.917 (0.177)	0.882 (0.050)	0.775 (0.126)	0.666 (0.28)	0.714 (0.238)
ICC SB-M intelligibility score	0.295 (0.0)	0.180 (0.008)	0.283 (0.0)	0.486 (0.0)	0.290 (0.0)	0.702 (0.0)	0.841 (0.0)	0.869 (0.0)
Google word error rate	0.167 (0.184)	0.238 (0.038)	0.160 (0.2)	0.303 (0.276)	0.288 (0.084)	0.437 (0.154)	0.540 (0.231)	0.479 (0.237)
Amazon word error rate	0.058 (0.113)	0.198 (0.042)	0.032 (0.106)	0.121 (0.202)	0.22 (0.066)	0.372 (0.143)	0.425 (0.228)	0.364 (0.193)
ICC consistency	—	—	—	—	0.75	0.858	0.955	0.982

CO, Colombian Spanish; CZ, Czech.

populations. We found significant correlations between the intelligibility score and the respective dysarthria anchor score for DE PD ($r = -0.46$, $p < 0.01$, $d = 1.03$), CO PD ($r = -0.39$, $p < 0.01$, $d = 0.85$), CZ PD ($r = -0.32$, $p < 0.05$, $d = 0.67$), and CZ HD ($r = -0.37$, $p < 0.05$, $d = 0.80$). Probably owing to the small sample size, statistically we only found a trend in CZ PSP ($r = -0.42$, $p < 0.10$, $d = 0.92$) and CZ ALS ($r = 0.32$, $p = 0.21$, $d = 0.68$), although effect sizes were medium to large. Compare also Figure 2.

Clinical validation

For the group comparisons, we found significant differences, with the ki: SB-M intelligibility score being significantly lower for the respective pathological group for all cohorts: HC CO > PD CO ($H = 17.425$, $p < 0.001$, $\eta^2 = 0.17$), HC CZ > PD CZ ($H = 13.304$, $p < 0.001$, $\eta^2 = 0.14$), CZ HC > CZ HD ($H = 44.437$, $p < 0.001$, $\eta^2 = 0.52$), CZ HC > CZ PSP ($H = 29.696$, $p < 0.001$, $\eta^2 = 0.46$), and CZ HC > CZ ALS ($H = 18.565$, $p < 0.001$, $\eta^2 = 0.29$). For description, please see Table 2, and a graphical overview of the group differences is provided in Figure 3.

Post hoc group comparisons revealed that the intelligibility scores were comparable for the CZ HD, PSP, and ALS groups, and the CZ PD and CO PD groups. However, German PD showed significantly better intelligibility than the other patient groups, actually performing on a par with the other language HC groups.

Discussion

This study aimed to validate the ki: speech biomarker for motor speech disorders intelligibility score (ki: SB-M

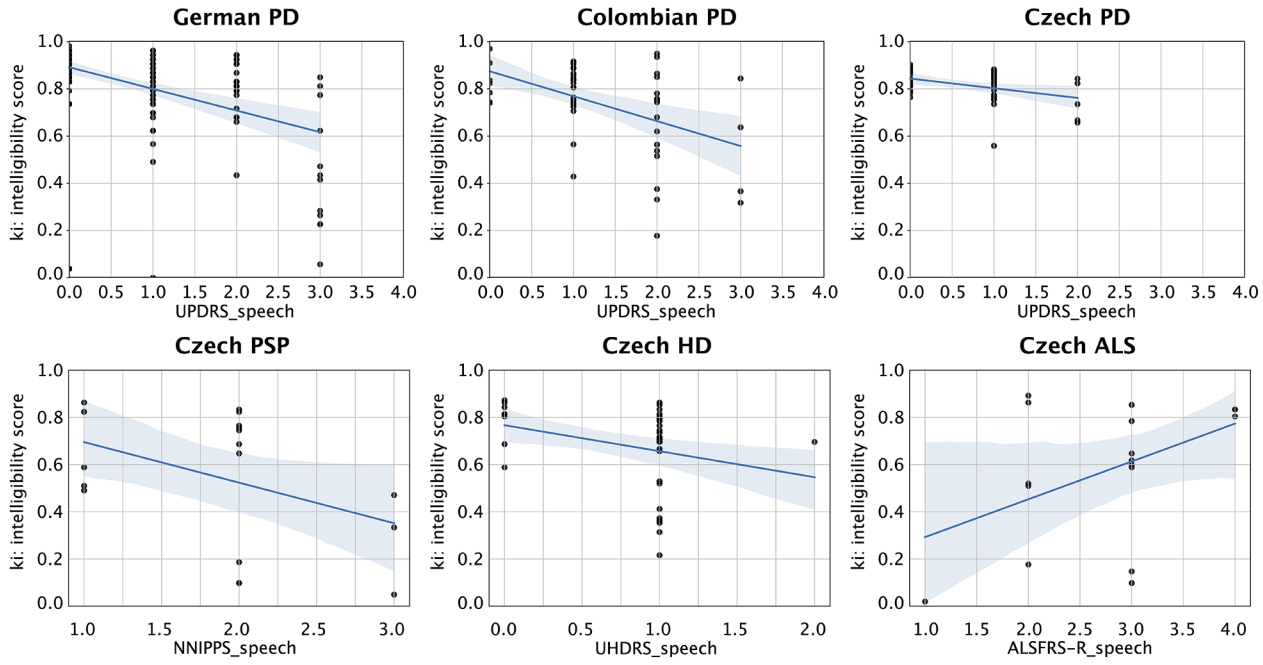

FIGURE 2
Scatter plots for the correlations between the intelligibility score and respective speech dysarthria clinical anchor score. From upper left to lower right: DE PD correlation with the MDS-UPDRS speech item; CO PD correlation with the MDS-UPDRS speech item; CZ PD correlation with the MDS-UPDRS speech item; CZ PSP DE PD correlation with the NNIPPS speech item; CZ HD correlation with the UHDRS dysarthria score; and CZ ALS correlation with the ALSFRS-R speech item (note that ALSFRS-R has an inverse relationship to disease severity, unlike the other scales in which higher scores mean greater severity). DE, German; CO, Colombian Spanish; CZ, Czech.

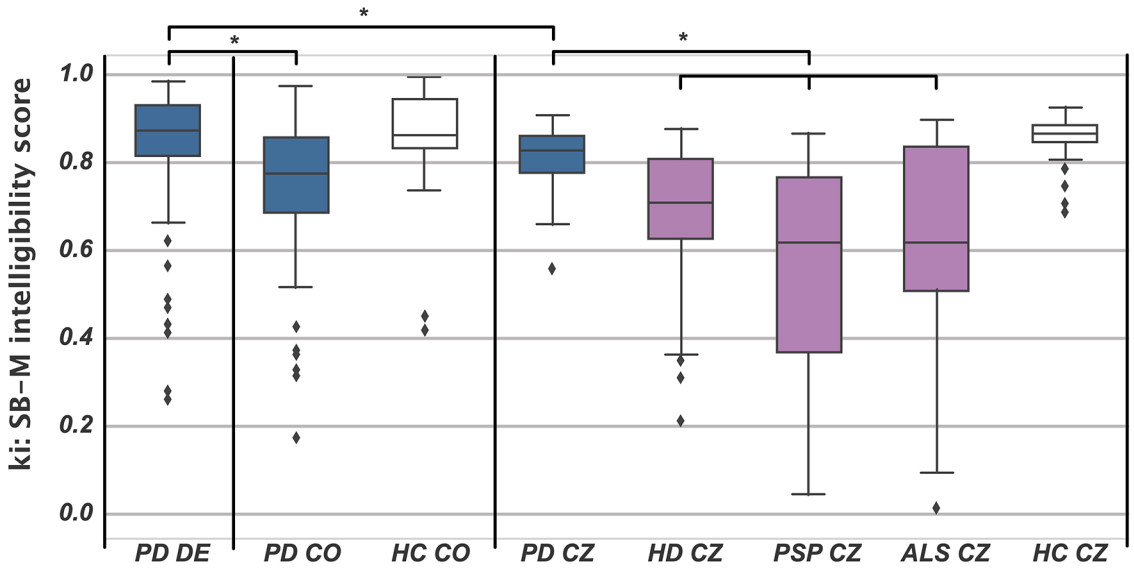

FIGURE 3
Boxplots of the SB-M intelligibility score for all groups. Blue, PD; white, HC; purple, HD, PSP, and ALS. Asterisks denote a significant *post hoc* group comparison. CO, Colombian Spanish; CZ, Czech.

intelligibility score) using the DiMe V3 framework, covering verification, analytical validation, and clinical validation across multiple languages and dysarthria pathologies. Making use of off-the-shelf ASR systems, we took a state-of-the-art approach to automatically measure speech intelligibility in dysarthrias (19, 28, 47). On a conceptual level, we went beyond the aforementioned studies, as we followed the DiME society V3 framework for assessing the readiness of digital measures for clinical research and also included multiple pathologies from the dysarthria spectrum as well as two different ASR systems.

We ran verification on the SB-M intelligibility score, calculating it based on two different automatic speech recognition systems: Google Speech API and Amazon Transcribe. Overall, the ICC indicated good to excellent agreement between the two ASR systems for most pathological groups. However, discrepancies were noted in the Colombian PD and Czech PD data, in which the ICC was only fair to poor, respectively. Poor stability of ASR-based intelligibility measures has been reported previously, especially for typical and mildly impaired severity groups, specifically decreasing their ability to measure changes in the early phases of motor speech disorders (19). The discrepancy might be due to the rather small variance and very good speech recognition, performing almost at an HC level of 0.80, whereas HD, PSP, and ALS have intelligibility scores of 0.70–0.50, with much bigger variances. In these cases, we assume that already small word-level differences inflate discrepancies between ASRs and might cause low ICCs. Especially with the advent of ever-improving ASRs, which also push the needle in dysarthric speech recognition alongside other underrepresented groups, this issue has to be watched closely.

The validity of Google and Amazon ASRs as commercial products naturally extends beyond pathological groups. Both ASR systems have shown high accuracy in recognizing speech from healthy individuals, providing a strong benchmark for comparison (48). However, ensuring robust performance for underrepresented groups remains crucial for the broad applicability and reliability of ASR systems in clinical and everyday settings. On the level of ASR performance in dysarthric speakers, our results compare well with other studies in the field. Gutz et al. (19) found WERs of 10% for mild ALS-related dysarthria to approximately 50% for moderate cases and approximately 80% for severe cases. This is in line with our results for the Czech ALS population, which can be classified as moderately dysarthric based on the ALSFRS-R speech item and shows a 40%–50% WER depending on the ASR system.

Consistency was assessed by comparing the intelligibility scores obtained from repeated paragraph readings. Overall, the ICC values indicated good to excellent consistency. This is an encouraging result but has to be further investigated for repeated measurements of the SB-M intelligibility score assessed longer timeframes apart, such as a couple of days or weeks.

Analytical validation compared the SB-M intelligibility score against established clinical anchor measures for dysarthrias derived from the respective gold standard clinical staging scale. Significant correlations were observed between the SB-M intelligibility score and the respective dysarthria anchor scores for the German PD, Colombian PD, Czech PD, and Czech HD groups. Although specific items are not designed as stand-alone assessments of dysarthria and even less as assessments of intelligibility in principle, we could still demonstrate correlations between the ki: SB-M intelligibility score and those measures. These findings support the SB-M intelligibility score's validity as a measure of perceived speech intelligibility being associated with dysarthria on the speaker side, as confirmed by traditional

clinical assessments such as the MDS-UPDRS, NNIPPS, and ALSFRS-R speech items or the UHDRS dysarthria score. Despite medium to large effect sizes, statistical significance was not achieved for the Czech PSP and Czech ALS groups, likely due to smaller sample sizes. Future studies should aim to include larger cohorts to increase statistical power and provide more robust analytical validation.

Our approach to measuring speech intelligibility differs from other research by using a direct measure based on ASR performance, rather than classifying speech into different states/classes of intelligibility. This research is sometimes carried out using machine learning techniques (49, 50). This line of research frames intelligibility as a classification problem, requiring labeled training data to categorize speech into predefined stages. By contrast, our method leverages the continuous output of ASR systems as a proxy for intelligibility, offering multiple benefits. This continuous measure might provide finer granularity and sensitivity to subtle changes in speech quality over time or between groups. In addition, using an off-the-shelf ASR approach eliminates the need for additional machine learning training, making it more accessible and easier to implement in various clinical and research settings.

One of the major limitations of the analytical validation we performed is that we cannot prove this further by comparing with manual intelligibility ratings by either trained professionals or human raters in general, as has been carried out by Gutz et al. in ALS (19). Future studies should add this piece of analytical validation, leveraging existing methods to rate intelligibility by multiple trained and/or untrained raters (51). In addition, our approach presents, in some respect, a black box approach that directly evaluates dysarthria based on intelligibility as perceived by a somehow non-transparent ASR black box. There is a whole research tradition on using carefully crafted acoustic features to estimate dysarthria and different subsystems, as mentioned in the introduction. Pursuing a hybrid approach that taps into ASR-based intelligibility and traditional acoustic analysis features (e.g., pause rate, articulation rate, pitch instability, or monotonicity) to evaluate patients' dysarthrias would increase the impact of such research and be an important next step.

Clinical validation demonstrated significant differences in SB-M intelligibility scores between healthy controls and pathological groups across all cohorts. This finding underscores the potential of the SB-M intelligibility score as a discriminative tool for identifying and quantifying speech impairments in individuals with motor speech disorders. The consistent pattern of lower intelligibility scores in pathological groups compared with healthy controls across different languages and disorders further supports the robustness and generalizability of the measure. Nevertheless, the experiments presented here still only cover a fraction of the total spectrum of motor-speech-disorder-related dysarthrias or dysarthrias in general. However, our data set of more than 250 patients across four different pathologies and three languages covers a significant amount in this field of research; for rare diseases such as ALS or atypical PD in particular, datasets of that size are rarely reported. In addition, we acknowledge that we did not perform specific testing for cognitive involvement, as the primary aim was to investigate motor speech deviations that are the main contributors to reduced intelligibility. Furthermore, we did not measure the vital capacity of our patients; cohorts such as ALS and PSP may have respiratory impairments that could significantly contribute to reduced intelligibility.

In general, we observed better speech intelligibility in patients with PD than in patients with HD, PSP, or ALS. One reason could be that in the earlier stages of PD, articulation impairment is not as pronounced, allowing for relatively clearer speech. Conversely, HD is characterized by hyperkinetic irregular articulation, and ALS and PSP are associated with hypertonia, leading to imprecise consonant production (52). These speech deficits in HD, ALS, and PSP significantly contribute to reduced intelligibility. These imprecise consonant and uncontrolled (sometimes spastic) irregularities in speech are known to hamper speech intelligibility a lot more than monopitch and monoloudness, which are typically observed in early PD. In addition, the spread in intelligibility scores was a lot greater for HD, PSP, and ALS than for PD, which was also in line with studies on those diseases showing more heterogeneity in their behavioral and speech impairment phenotype.

Between the separate PD groups (DE, CO, and CZ), we observed comparable intelligibility scores in CO and CZ but the German PD group was significantly more intelligible—actually performing on a par with the other language HC groups. This could be related to different recording setups in each study or a general language difference in the underlying ASR performance.

ASR and the measures derived from it exhibit considerable variability when applied to different types of dysarthria (53). Articulatory precision has been identified as the most critical factor influencing speech intelligibility, surpassing the impact of prosody (54).

Finally, another limitation to this study is that we compared intelligibility for audios collected from different studies with different audio recording settings. Although all studies used state-of-the-art microphones for audio recording and professional recording setups—as recommended by recent guidelines (5)—differences in audio recording setups can always play a role in head-to-head comparisons; this is especially the case when comparing our results from CZ directly with CO and DE. Eventually, the accuracy of an automatic speech intelligibility measure is highly dependent on recording conditions. Poor recording environments, such as those with high background noise or subpar microphone quality, can introduce significant bias, leading to artificially low intelligibility ratings. This may result in the erroneous classification of normal speech as dysarthric. Furthermore, different recording devices and handling methods introduce substantial variance, which can confound the measurements and reduce their sensitivity to detect small changes over time or differences between low dysarthria groups. However, one of the most promising scenarios in which to deploy this kind of technology is in at-home environments, where the patient is monitored in everyday life, always using the same device and with similar acoustic conditions. This approach

has shown promising results (55). Future studies in this field should adhere even closer to a standardized recording setup or record with multiple devices—one being a standardized microphone setup next to others.

Conclusion

Overall, this study provides a comprehensive validation of the ki: SB-M intelligibility score for assessing speech intelligibility in motor speech disorders across multiple languages and pathologies. The findings support its reliability, validity, and clinical relevance, highlighting its potential as a standardized tool for clinical and research applications. Automated objective measures of speech intelligibility, such as the SB-M intelligibility score, can increase the efficiency and accuracy of dysarthria assessments, reduce observer bias, and facilitate remote monitoring. This is particularly advantageous for large-scale international clinical trials, in which high-frequency data collection and scalability are critical.

Future efforts should complement validation by investigating the SB-M intelligibility score's ability to monitor disease progression and treatment efficacy. Longitudinal studies assessing changes in the intelligibility score over time and in response to therapeutic interventions could provide valuable insights into the clinical utility of this digital measure.

Data availability statement

The data analyzed in this study are subject to the following licenses/restrictions: the speech data can be accessed from the respective cohort-associated author upon reasonable request. Please navigate through the linked reference for each study in the Methods section. Requests to access these datasets should be directed to rafael.orozco@udea.edu.co, rusz.mz@gmail.com, and tabea.thies@uk-koeln.de.

Ethics statement

The studies involving humans were approved by local ethics committees in Cologne, Germany, Prague, Czech Republic, and Medellín, Colombia. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JT: Writing – original draft, Writing – review & editing. FD: Writing – original draft, Writing – review & editing. LS: Writing – original draft, Writing – review & editing. NL: Writing – original draft, Writing – review & editing. AK: Writing – original draft, Writing – review & editing. TT: Writing – original draft, Writing – review & editing. JO-A: Writing – original draft, Writing – review & editing. JR: Writing – original draft, Writing – review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article.

JR received grant funding from The Czech Ministry of Health (NW24-04-00211) and National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107)—Funded by the European Union—Next Generation EU. The work on the Colombian data was partially funded by CODI at UdeA (PI2023-58010).

Conflict of interest

JT, FD, LS, NL, and AK are employed by the speech biomarker company ki:elements GmbH. JT and NL also hold shares in the speech biomarker company ki:elements GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. *J Speech Hear Res.* (1969) 12(2):246–69. doi: 10.1044/jshr.1202.246

2. Amato F, Borzì L, Olmo G, Artusi CA, Imbalzano G, Lopiano L. Speech impairment in Parkinson's disease: acoustic analysis of unvoiced consonants in Italian native speakers. *IEEE Access.* (2021) 9:166370–81. doi: 10.1109/ACCESS.2021.3135626

3. Rusz J, Cmejla R, Tykalova T, Ruzickova H, Klempir J, Majerova V, et al. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. *J Acoust Soc Am.* (2013) 134(3):2171–81. doi: 10.1121/1.4816541

4. Rusz J, Tykalová T, Salerno G, Bancone S, Scarpelli J, Pellecchia MT. Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy. *J Neurol.* (2019) 266(6):1394–404. doi: 10.1007/s00415-019-09271-7

5. Rusz J, Tykalova T, Ramig LO, Tripoliti E. Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. *Mov Disord.* (2021) 36(4):803–14. doi: 10.1002/mds.28465

6. Daoudi K, Das B, Tykalova T, Klempir J, Rusz J. Speech acoustic indices for differential diagnosis between Parkinson's disease, multiple system atrophy and progressive supranuclear palsy. *NPJ Parkinsons Dis*. (2022) 8(1):1–13. doi: 10.1038/s41531-021-00272-w

7. Kim Y, Kent RD, Kent JF, Duffy JR. Perceptual and acoustic features of dysarthria in multiple system atrophy. *J Med Speech-Lang Pathol*. (2010) 18(4):66–71.

8. Diehl SK, Mefferd AS, Lin YC, Sellers J, McDonell KE, De Riesthal M, et al. Motor speech patterns in Huntington disease. *Neurology*. (2019) 93(22):E2042–52. doi: 10.1212/WNL.0000000000008541

9. Kouba T, Frank W, Tykalova T, Mühlbäck A, Klempíř J, Lindenberg KS, et al. Speech biomarkers in Huntington disease: a cross-sectional study in pre-symptomatic, prodromal and early manifest stages. *Eur J Neurol*. (2023) 1262–71. doi: 10.1111/ene.15726

10. Tomik B, Guiloff RJ. Dysarthria in amyotrophic lateral sclerosis: a review. *Amyotroph Lateral Scler*. (2010) 11(1–2):4–15. doi: 10.3109/17482960802379004

11. Darley FL, Brown JR, Goldstein NP. Dysarthria in multiple sclerosis. *J Speech Hear Res*. (1972) 15(2):229–45. doi: 10.1044/jshr.1502.229

12. Rusz J, Benova B, Ruzickova H, Novotny M, Tykalova T, Hlavnicka J, et al. Characteristics of motor speech phenotypes in multiple sclerosis. *Mult Scler Relat Disord*. (2018) 19:62–9. doi: 10.1016/j.msard.2017.11.007

13. Van Uem JMT, Marinus J, Canning C, Van Lummel R, Dodel R, Liepelt-Scarfone I, et al. Health-related quality of life in patients with Parkinson's disease—a systematic review based on the ICF model. *Neurosci Biobehav Rev*. (2016) 61:26–34. doi: 10.1016/j.neubiorev.2015.11.014

14. Chu SY, Tan CL. Subjective self-rated speech intelligibility and quality of life in patients with Parkinson's disease in a Malaysian sample. *Open Public Health J*. (2018) 11(1):485–93. doi: 10.2174/1874944501811010485

15. McAuliffe MJ, Baylor CR, Yorkston KM. Variables associated with communicative participation in Parkinson's disease and its relationship to measures of health-related quality-of-life. *Int J Speech Lang Pathol*. (2017) 19(4):407–17. doi: 10.1080/17549507.2016.1193900

16. Schrag A, Jahanshahi M, Quinn N. How does Parkinson's disease affect quality of life? A comparison with quality of life in the general population. *Mov Disord*. (2000) 15(6):1112–8. doi: 10.1002/1531-8257(200011)15:6<1112::AID-MDS1008>3.0.CO;2-A

17. Yorkston KM, Yorkston KM. *Management of motor speech disorders in children and adults*. 2nd ed. Austin, TX: Pro-Ed (1999). p. 618.

18. Miller N, Allcock L, Jones D, Noble E, Hildreth AJ, Burn DJ. Prevalence and pattern of perceived intelligibility changes in Parkinson's disease. *J Neurol Neurosurg Amp Psychiatry*. (2007) 78(11):1188–90. doi: 10.1136/jnnp.2006.110171

19. Gutz SE, Stipancic KL, Yunusova Y, Berry JD, Green JR. Validity of off-the-shelf automatic speech recognition for assessing speech intelligibility and speech severity in speakers with amyotrophic lateral sclerosis. *J Speech Lang Hear Res*. (2022) 65 (6):2128–43. doi: 10.1044/2022_JSLHR-21-00589

20. Kent RD, Weismer G, Kent JF, Rosenbek JC. Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord*. (1989) 54(4):482–99. doi: 10.1044/jshd.5404.482

21. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results: MDS-UPDRS: clinimetric assessment. *Mov Disord*. (2008) 23(15):2129–70. doi: 10.1002/mds.22340

22. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J Neurol Sci*. (1999) 169(1–2):13–21. doi: 10.1016/S0022-510X(99)00210-5

23. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit Biomark*. (2021) 5 (1):78–88. doi: 10.1159/000515346

24. Espay AJ, Hausdorff JM, Sánchez-Ferro Á, Klucken J, Merola A, Bonato P, et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. *Mov Disord*. (2019) 34(5):657–63. doi: 10.1002/mds.27671

25. Salmon K, Genge A. "Clinical trials in ALS—current challenges and strategies for future directions". In: Shaw CA, Morrice JR, editors. *Spectrums of Amyotrophic Lateral Sclerosis*. 1st ed. Hoboken, NJ: John Wiley & Sons, Inc. (2021). p. 161–80.

26. Bowden M, Beswick E, Tam J, Perry D, Smith A, Newton J, et al. A systematic review and narrative analysis of digital speech biomarkers in motor neuron disease. *NPJ Digit Med*. (2023) 6(1):228. doi: 10.1038/s41746-023-00959-9

27. Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The first frontier: digital biomarkers for neurodegenerative disorders. *Digit Biomark*. (2017) 1(1):6–13. doi: 10.1159/000477383

28. Vásquez-Correa JC, Orozco-Arroyave JR, Bocklet T, Nöth E. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *J Commun Disord*. (2018) 76:21–36. doi: 10.1016/j.jcomdis.2018.08.002

29. De Russis L, Corno F. On the impact of dysarthric speech on contemporary ASR cloud platforms. *J Reliab Intell Environ*. (2019) 5(3):163–72. doi: 10.1007/s40860-019-00085-y

30. Mustafa MB, Rosdi F, Salim SS, Mughal MU. Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Syst Appl*. (2015) 42(8):3924–32. doi: 10.1016/j.eswa.2015.01.033

31. Keshet J. Automatic speech recognition: a primer for speech-language pathology researchers. *Int J Speech Lang Pathol*. (2018) 20(6):599–609. doi: 10.1080/17549507.2018.1510033

32. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). *NPJ Digit Med*. (2020) 3(1):1–15. doi: 10.1038/s41746-020-0260-4

33. Coravos A, Doerr M, Goldsack J, Manta C, Shervey M, Woods B, et al. Modernizing and designing evaluation frameworks for connected sensor technologies in medicine. *NPJ Digit Med*. (2020) 3(1):37. doi: 10.1038/s41746-020-0237-3

34. Goldsack JC, Dowling AV, Samuelson D, Patrick-Lake B, Clay I. Evaluation, acceptance, and qualification of digital measures: from proof of concept to endpoint. *Digit Biomark*. (2021) 5(1):53–64. doi: 10.1159/000514730

35. Rusz J, Klempíř J, Tykalová T, Baborová E, Čmejla R, Růžička E, et al. Characteristics and occurrence of speech impairment in Huntington's disease: possible influence of antipsychotic medication. *J Neural Transm*. (2014) 121 (12):1529–39. doi: 10.1007/s00702-014-1229-8

36. Hlavnička J, Čmejla R, Tykalová T, Šonka K, Růžička E, Rusz J. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Sci Rep*. (2017) 7(1):12. doi: 10.1038/s41598-017-00047-5

37. Novotny M, Melechovsky J, Rozenstoks K, Tykalova T, Kryze P, Kanok M, et al. Comparison of automated acoustic methods for oral diadochokinesis assessment in amyotrophic lateral sclerosis. *J Speech Lang Hear Res*. (2020) 63(10):3453–60. doi: 10.1044/2020_JSLHR-20-00109

38. Orozco-Arroyave JR, Arias-Londoño JD, Vargas-Bonilla JF, González-Rátiva MC, Nöth E. *New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease*. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 26–31 May 2014; Reykjavik, Iceland*. Paris: European Language Resources Association (ELRA) (2014). p. 342–7.

39. Thies T, Mücke D, Lowit A, Kalbe E, Steffen J, Barbe MT. Prominence marking in parkinsonian speech and its correlation with motor performance and cognitive abilities. *Neuropsychologia*. (2020) 137:107306. doi: 10.1016/j.neuropsychologia.2019.107306

40. Hlavnička J. The dysarthria analyzer. Available online at: https://www.dysan.cz/ (accessed June 25, 2024).

41. Kieburtz K, Penney JB, Como P, Ranen N, Shoulson I, Feigin A, et al. Unified Huntington's disease rating scale: reliability and consistency. *Mov Disord*. (1996) 11 (2):136–42. doi: 10.1002/mds.870110204

42. Payan CAM, Viallet F, Landwehrmeyer BG, Bonnet AM, Borg M, Durif F, et al. Disease severity and progression in progressive supranuclear palsy and multiple system atrophy: validation of the NNIPPS—Parkinson plus scale. *PLoS One*. (2011) 6(8):e22293. doi: 10.1371/journal.pone.0022293

43. Google LLC. Google speech-to-text V2.25.1 (2023). Available online at: https://cloud.google.com/speech-to-text/ (Accessed July 29, 2022).

44. Amazon Web Services Inc. Amazon Transcribe V1.20.15 (2023). Available online at: https://aws.amazon.com/pm/transcribe/ (Accessed April 11, 2024).

45. Van Nuffelen G, Middag C, De Bodt M, Martens J. Speech technology-based assessment of phoneme intelligibility in dysarthria. *Int J Lang Commun Disord*. (2009) 44(5):716–30. doi: 10.1080/13682820802342062

46. Ratitch B, Trigg A, Majumder M, Vlajnic V, Rethemeier N, Nkulikiyinka R. Clinical validation of novel digital measures: statistical methods for reliability evaluation. *Digit Biomark*. (2023) 7:74–91. doi: 10.1159/000531054

47. Dimauro G, Di Nicola V, Bevilacqua V, Caivano D, Girardi F. Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. *IEEE Access*. (2017) 5:22199–208. doi: 10.1109/ACCESS.2017.2762475

48. Zhang Y, Han W, Qin J, Wang Y, Bapna A, Chen Z, et al. Google USM: scaling automatic speech recognition beyond 100 languages. arXiv [preprint]. (2023). Available online at: https://arxiv.org/abs/2303.01037 (accessed July 1, 2024).

49. Kim J, Kumar N, Tsiartas A, Li M, Narayanan SS. Automatic intelligibility classification of sentence-level pathological speech. *Comput Speech Lang*. (2015) 29 (1):132–44. doi: 10.1016/j.csl.2014.02.001

50. Huang A, Hall K, Watson C, Shahamiri SR. *A review of automated intelligibility assessment for dysarthric speakers*. In: *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD); 13–15 Oct 2021; Bucharest, Romania*. New York, NY: IEEE (2021). p. 19–24.

51. Stipancic KL, Tjaden K, Wilding G. Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *J Speech Lang Hear Res*. (2016) 59(2):230–8. doi: 10.1044/2015_JSLHR-S-15-0271

52. Campi M, Peters GW, Toczydlowska D. Ataxic speech disorders and Parkinson's disease diagnostics via stochastic embedding of empirical mode decomposition. *PLoS One*. (2023) 18(4):e0284667. doi: 10.1371/journal.pone.0284667

53. Rowe HP, Gutz SE, Maffei MF, Tomanek K, Green JR. Characterizing dysarthria diversity for automatic speech recognition: a tutorial from the clinical perspective. *Front Comput Sci*. (2022) 4:770210. doi: 10.3389/fcomp.2022.770210

54. Tu M, Wisler A, Berisha V, Liss JM. The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance. *J Acoust Soc Am*. (2016) 140(5):EL416–22. doi: 10.1121/1.4967208

55. Arias-Vergara T, Vásquez-Correa JC, Orozco-Arroyave JR, Nöth E. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. *Speech Commun*. (2018) 101:11–25. doi: 10.1016/j.specom.2018.05.007