ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Automatic Speech Recognition for Intelligibility Assessment in Children With Dysarthria

Jiyoung Choi,[a] Gemma Moya-Galé,[a] KyungHae Hwang,[b] Julia Hirschberg,[c] and Erika S. Levy[a]

[a] Department of Biobehavioral Sciences, Teachers College, Columbia University, New York, NY [b] Department of Communicative Sciences and Disorders, New York University, NY [c] Department of Computer Science, Columbia University, New York, NY

ARTICLE INFO

Article History:
Received July 17, 2025
Revision received October 26, 2025
Accepted November 25, 2025

Editor-in-Chief: Jessica E. Huber
Editor: Kristen M. Allison

https://doi.org/10.1044/2025_JSLHR-25-00562

ABSTRACT

Purpose: Accurate assessment of speech intelligibility is critical for children with dysarthria secondary to cerebral palsy. Traditional assessment methods, such as human listeners' orthographic transcription and perceptual ratings (e.g., of ease of understanding [EoU]), are time consuming or subjective. Automatic speech recognition (ASR) may provide a more efficient, objective alternative, but its use for assessing intelligibility in this population is unexamined. This study evaluated the potential of ASR for intelligibility assessment in children with dysarthria and identified the most appropriate ASR systems for approximating human listeners' judgments.
Method: Five ASR systems transcribed speech samples from 20 children with dysarthria. Additionally, 168 adult listeners provided orthographic transcriptions and EoU ratings. Word recognition rate (WRR) was used as the metric for calculating ASR and human listeners' transcription accuracy. Spearman correlations were used to assess the relationship between ASR WRR and human WRR, as well as between ASR WRR and human EoU ratings.
Results: The WRR yielded by four ASR systems (WhisperX-small, WhisperX-medium, WhisperX-large, and Google Cloud) showed strong correlations with human WRR, with WhisperX-medium demonstrating the strongest correlation. These four systems' WRRs also exhibited moderate-to-strong correlations with EoU ratings, with Google Cloud ASR showing the strongest correlation. In contrast, the WRR of Wav2Vec2 demonstrated a weak correlation with both human WRR and EoU ratings.
Conclusions: ASR shows promise for use in intelligibility assessment in children with dysarthria. Of the tested ASR systems, WhisperX-medium appears most promising for approximating human transcription accuracy, whereas Google Cloud ASR aligns best with perceptual ratings. Such differences in ASR performance highlight the need for careful system selection in clinical applications.
Supplemental Material: https://doi.org/10.23641/asha.31397457

Cerebral palsy (CP) is the most common neuromotor disorder in children, affecting approximately one in 350 children (Centers for Disease Control and Prevention, 2020). The motor speech disorder of dysarthria is present in the majority of these children and is associated with reduced speech intelligibility (Allison & Hustad, 2018; Hustad, 2008; Mei et al., 2020). Because intelligibility impairments can negatively impact communication in children with dysarthria due to CP (henceforth, children with dysarthria) in ways that may limit their social and educational engagement, increasing speech intelligibility is a fundamental treatment goal for this population (Hustad, 2008; Levy et al., 2021; Mei et al., 2020). Given the pivotal role of intelligibility in determining successful communication, accurate measures of children's intelligibility are essential for assessment and for determining any treatment-related gains (Miller, 2013).

Speech intelligibility in dysarthria is typically assessed through human perceptual evaluation (Allison, 2020; Hustad, 2008). Two primary human assessment measures often implemented are orthographic word transcription accuracy and

Correspondence to Jiyoung Choi: jyc2173@tc.columbia.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

perceptual ratings. Human listeners' transcription accuracy, considered the gold standard and most objective measure of intelligibility (Hustad, 2006), requires listeners to write the words they believe they heard. Intelligibility is generally measured in terms of the percentage of words correctly transcribed (Hustad, 2008; Levy et al., 2017). Perceptual ratings, in contrast, involve listeners judging the ease of understanding (EoU) or related qualities of speech (Allison, 2020; Fletcher et al., 2017; Landa et al., 2014; Levy et al., 2017, 2021). Additionally, human judgments of intelligibility, including visual analog scale (VAS) ratings and percent estimates, are widely used in both clinical and research settings (Abur et al., 2019; Hustad, 2006; Levy et al., 2017, 2021; Stipancic et al., 2016; Tjaden et al., 2014). Advantages of ratings and percent estimates include their ease of administration and minimal time requirements (Abur et al., 2019; Hustad, 2006). Human listeners' transcription accuracy and perceptual ratings can complement each other, as transcription accuracy quantifies the percentage of words correctly understood, while ratings reflect broader listener experiences, such as the perceived effort needed to decode the speech.

Despite their respective advantages, both measures have significant drawbacks. In research, human perceptual evaluation entails significant time spent on preparing speech samples for listeners, recruiting and testing listeners, and analyzing their ratings and/or transcriptions (Allison, 2020; Atkins et al., 2019; Jiao et al., 2019). Additionally, factors such as listeners' familiarity with the speaker, dysarthria, or test materials, as well as their age, hearing, and attitude, can impact both transcriptions and ratings (Borrie et al., 2012; Miller, 2013). Furthermore, determining transcription accuracy can require testing numerous listeners to minimize learning effects from repeated exposure to the same utterances (Tu et al., 2016). In fact, the extensive time and resources required for transcription and its analysis may be prohibitive for clinical practice and research (C. B. Fox et al., 2021; Xu et al., 2014).

While perceptual ratings are more commonly used in clinical settings because of their greater efficiency, they have been criticized for being more subjective and less reliable than transcription accuracy (Kreiman & Gerratt, 1998; Miller, 2013). Variability in listener attitudes, sensitivity to particular speech distortions, and difficulty assigning a single rating to fluctuating intelligibility contribute to this subjectivity, along with "listeners' individual internal yardsticks" (Miller, 2013, p. 603) for severity measurement. However, perceptual ratings have shown strong correlations with transcription accuracy, suggesting that ratings may serve as a reasonable alternative for assessing dysarthric speech (Abur et al., 2019; Stipancic et al., 2016; Tjaden et al., 2014).

Given the limitations of traditional human-based assessment, there is growing interest in leveraging communication technology to streamline the assessment process, particularly through the use of automatic speech recognition (ASR) systems as an alternative for evaluating intelligibility in clinical populations. Two main categories of ASR systems are (a) speaker-independent ASR systems and (b) personalized ASR systems, such as speaker-adaptive or speaker-dependent systems (Rowe et al., 2022). Most speaker-independent systems are commercially available to the public. Such systems include platforms such as Siri, Amazon Alexa, and Google Home. In contrast, personalized ASR systems such as Dragon or Cortana (e.g., Microsoft, 2020; Nuance, 2020) are designed for a specific user or population.

Currently, speaker-independent ASR systems demonstrate lower transcription accuracy for individuals with speech impairments than for neurotypical speakers (De Russis & Corno, 2019; Gutz et al., 2022; Schultz et al., 2021).[1] For instance, Schultz and colleagues suggested that while ASR transcription accuracy is lower for individuals with impaired speech, especially those with neurodegenerative conditions, such as multiple sclerosis (MS) and Friedreich's ataxia (FA), than for neurotypical adults, personalized ASR systems tailored to specific speech disorders could be trained to improve their transcriptions. Moreover, such systems show better performance than speaker-independent ASR systems, especially in recognizing severely dysarthric speech (Green et al., 2021). While personalized ASR systems offer better accuracy for individual users, their customization may limit generalizability in clinical settings (Yılmaz et al., 2019). Thus, despite their limitations, current speaker-independent ASR systems can provide immediate transcriptions that may support continuous monitoring of intelligibility and aid in assessing treatment efficacy (Gutz et al., 2023; Tobin et al., 2024). Given their ease of implementation and broader applicability, this study focuses on speaker-independent ASR systems, which are more feasible to apply across speakers, are without extensive customization, and hold greater potential for implementation in clinical speech assessment.

Various speaker-independent ASR systems have been evaluated for their transcription accuracy and potential clinical use for adult speech. Schultz et al. (2021) tested the performance of Amazon Web Services, Google

[1]Regarding terminology, word error rate (WER) is a standard measure of transcription performance reported in many ASR studies (e.g., C. B. Fox et al., 2021; Tetzloff et al., 2024; Tu et al., 2016). ASR transcription accuracy, which may be more intuitive than WER in clinical settings, is the complement of WER. For example, ASR transcription accuracy of 90% reflects a WER of 10%. For consistency and interpretability, we refer primarily to "ASR transcription accuracy" throughout this article. An additional term, word recognition rate (WRR; Gutz et al., 2022) denotes a metric for both ASR and human transcription accuracy, detailed in the Method section. We use the term *WRR* when describing transcription accuracy calculations.

Cloud, and IBM Watson ASR platforms for speech produced by adults with neurodegenerative diseases, including MS and FA. Amazon and Google outperformed IBM Watson in accuracy, suggesting that some ASR systems may be better suited than others for handling the complexities of impaired speech. Similarly, Jacks et al.'s (2019) comparison of IBM Watson and Google Cloud ASR for transcribing the speech of adults with aphasia secondary to stroke revealed strong correlations between ASR and human listeners' transcription accuracy, highlighting ASR's potential as a tool to approximate human transcriptions. Furthermore, Rodrigues et al. (2019) compared Google, Bing, and Nuance ASR systems in various noise conditions and distances from neurotypical speakers of various ages. Google consistently outperformed the other systems, demonstrating greater robustness to noise and age-related speech variations, suggesting that Google is the most reliable ASR for neurotypical speech in diverse environments.

The performance of ASR systems can be evaluated based on two key metrics: the system's success in transcribing what the speaker intended to say (i.e., ASR transcription accuracy) and its success in aligning with human listeners' perceptual assessment (i.e., comparability). The measure of ASR transcription accuracy itself provides information on the system's ability to recognize spoken words (Schultz et al., 2021; Tetzloff et al., 2024) and can be useful for determining its appropriateness for generating speech-to-text input for augmentative and alternative communication (AAC), for example. In contrast, comparability, examined through the correlation of ASR transcription accuracy with human listeners' transcription accuracy or ratings, provides insight into ASR's potential for approximating humans' experience in decoding speech. Thus, comparability is essential for determining ASR's potential as a tool for assessing a speaker's intelligibility in communication with other humans (Gutz et al., 2022; Jacks et al., 2019; Tu et al., 2016).

The comparability of ASR systems with human listeners has been examined for the speech of adult clinical populations. Very high correlations between ASR and human transcription accuracy (Spearman ρ = .96–.98) were observed when IBM Watson was used to transcribe speech from individuals with aphasia and/or apraxia of speech secondary to stroke (Jacks et al., 2019). Similarly, Google Cloud ASR performed comparably to human listeners in transcribing speech from adults with Parkinson's disease (PD) in noise, with a probability of .80 of matching or outperforming the average listener (Moya-Galé et al., 2022). Moreover, Tu et al. (2016) found strong correlations (Pearson r = .68–.84) between ASR transcription accuracy and human perceptual ratings of the speech of adults with various subtypes of dysarthria across domains, including overall speech severity, articulatory precision, nasality, vocal quality, and prosody. These findings suggest that ASR can approximate human perception of the speech of adults with communication disorders, highlighting its potential as an efficient and objective tool for assessing intelligibility of adults in clinical settings.

Building on advancements in ASR applications for evaluating dysarthric speech in adults, researches have shown growing interest in extending this technology to pediatric populations. However, children's speech presents distinct physiological, linguistic, and acoustic challenges that need to be addressed to achieve optimal transcription and intelligibility assessment. While state-of-the-art ASR systems for neurotypical adults achieve high transcription accuracy (approximately 99%; Shakhadri et al., 2025), ASR performance for typically developing children is lower, with transcription accuracy averaging approximately 91% (Attia et al., 2024). The discrepancies between ASR performance for the speech of adults and children are largely attributed to children's higher fundamental and formant frequencies (Wilpon & Jacobsen, 1996) and underdeveloped phonemic systems (Yeung & Alwan, 2018).

Much of the existing ASR research on pediatric clinical populations has evaluated how accurately ASR systems transcribe speech rather than how closely ASR output aligns with human listeners' perception (comparability). For example, C. B. Fox et al. (2021) evaluated Google Cloud Speech for transcribing narrative language samples from school-age children with developmental language disorder. Notably, transcriptions generated by ASR were more accurate and yielded more reliable language sample analyses, including measures of intelligibility and disfluency, compared to those provided by two groups of human listeners (speech-language pathologists [SLPs] and trained transcribers) who transcribed the samples in real time. Furthermore, Gale et al. (2019) fine-tuned ASR models using spontaneous language samples from children with autism spectrum disorder, increasing ASR transcription accuracy from 69% to 74%.[2] Such work has helped lay critical groundwork for more efficient, child-centered ASR applications, such as intelligibility assessment and automated language analysis. These studies indicate growing interest in applying ASR to pediatric clinical populations, an important first step toward evaluating its utility for intelligibility assessment.

However, as described above in studies of adults, evaluating ASR transcription accuracy alone may not represent speakers' intelligibility as experienced by human listeners. This is particularly the case when the speakers are children, whose phonological and articulatory patterns are

[2]Fine-tuning refers to adapting a pretrained ASR system to a smaller, domain-specific labeled data set so that it better captures the linguistic and acoustic features of the speech of that population, thereby improving transcription accuracy (Gale et al., 2019; Yang et al., 2022).

highly variable during development. Not surprisingly, different challenges arise for ASR versus for human listeners in decoding children's speech. For example, while ASR systems depend heavily on acoustic patterns, which are highly variable in children, human listeners integrate contextual, prosodic, and semantic cues, as well as their expectations of children's phonological processes, to make sense of less precise productions (Meylan et al., 2023). In contrast, human listeners' perception of intelligibility of the speech of children with speech sound disorders might be negatively affected by what they perceive as the "oddness" of the children's atypical speech (Strömbergsson et al., 2021), whereas ASR is not affected by such social constructs. Thus, evaluating ASR's comparability to human listeners' perception (i.e., comparing ASR transcription accuracy to listeners' transcription accuracy and to perceptual ratings) may offer a more clinically meaningful metric of communication success, representing human listeners' experience in communication than simply assessing ASR transcription accuracy. This is particularly important for pediatric populations, where speech is more variable and listeners' knowledge and judgment play key roles in intelligibility assessment.

Therefore, comparability studies that examine how closely ASR aligns with human perception are most clinically relevant for intelligibility assessment. To date, only a few studies have explored ASR's perceptual comparability in pediatric clinical populations. For instance, Maier et al. (2006) found strong negative correlations (Cohen's κ ranged from −.85 to −.90) between ASR transcription accuracy and expert listener ratings of speech severity in children with cleft lip and palate. This negative relationship indicates that as ASR accuracy decreased, expert-perceived severity of speech impairment increased. Similarly, Lilley et al. (2014) evaluated the relationship between ASR-derived scores and human listener scoring using a closed-set word identification task (i.e., identifying the correct words from a list of 12 phonetically similar choices) from the speech productions of children with normal hearing and hearing impairment. They reported strong agreement between ASR scoring and listener identification scoring. These findings indicate that ASR may have the potential to approximate human listeners' perceptual assessment in certain populations of children with speech disorders.

Despite promising research on ASR for the speech of children with various disorders, its potential for assessing intelligibility in children with dysarthria remains unknown. The speech features of dysarthria in children, including imprecise articulation, variable intensity, reduced prosodic control, and high variability across productions (Allison & Hustad, 2018), may pose particular challenges for ASR systems, indicating the need for studies evaluating ASR's alignment with human perception of their speech. Therefore, determining the potential of current ASR systems for intelligibility assessment of children with dysarthria may be an essential first step to improving the efficiency of assessment and intervention practices for this population. Nonetheless, current speaker-independent ASR systems often face challenges with the high acoustic variability of children's speech and may perform particularly poorly for children with severe speech impairments. These limitations highlight the need to interpret ASR outcomes cautiously and to evaluate not only ASR's accuracy but also its alignment with human listeners' perception.

Current Study

This study evaluated the potential of ASR technology for assessing the speech intelligibility of children with dysarthria due to CP by determining ASR's comparability with human listeners' perception. ASR transcription accuracy was compared to the gold standard measure of human listeners' orthographic transcription accuracy (Hustad, 2006) and to commonly implemented perceptual VAS ratings (Abur et al., 2019; Levy et al., 2017, 2021; Stipancic et al., 2016; Tjaden et al., 2014).

Our first aim was to determine the relationship between ASR transcription accuracy and human listeners' transcription accuracy of the speech of children with dysarthria. To investigate this, speech samples from children with dysarthria were transcribed by five ASR systems (i.e., Google Cloud, Wav2Vec2, WhisperX-small, WhisperX-medium, WhisperX-large) and 168 human listeners. ASR transcription accuracy was compared to human listeners' transcription accuracy to assess the degree of alignment. It was hypothesized that ASR transcription accuracy would show moderate-to-strong correlations with human listeners' transcription accuracy (Gutz et al., 2022; Jacks et al., 2019), supporting its potential for intelligibility assessment in children with dysarthria. However, given previous findings in adults with dysarthria (De Russis & Corno, 2019; Gutz et al., 2022; Jiao et al., 2019; Tu et al., 2016), ASR was expected to transcribe dysarthric speech with lower accuracy than human listeners (Gutz et al., 2022). It was hypothesized that the transcription accuracy yielded by Google Cloud would show a stronger correlation with human listeners' transcription accuracy than would the transcription accuracy yielded by the other four ASR systems (Gutz et al., 2022), suggesting the potential of Google Cloud as the most appropriate ASR system for approximating human transcription in this population.

Our second aim was to determine the relationship between ASR transcription accuracy and human perception of intelligibility, specifically listeners' EoU ratings (Fletcher et al., 2017; Landa et al., 2014; Levy et al., 2017, 2021). It was hypothesized that, in all five ASR systems, ASR transcription accuracy would strongly correlate with human EoU ratings, suggesting that ASR could serve

as an efficient, objective tool for estimating intelligibility as perceived by a listener (Tu et al., 2016). With regard to the five ASR systems investigated, it was hypothesized that Google Cloud would be the best proxy for human listeners for rating the speech of children with dysarthria in that its transcription accuracy would show the strongest correlation with human EoU ratings (Gutz et al., 2022; Tu et al., 2016).

Method

This study was approved by the institutional review board (IRB) at Teachers College, Columbia University, New York City (IRB Nos. 24-394 and 24-287). Prior to beginning the study, child participants provided assent, and their parents provided consent. Adult listeners also provided informed consent online.

Participants

Children With Dysarthria

A total of 20 children (seven girls, 13 boys) with dysarthria due to CP were included in this study. All children were participants in an ongoing study on intelligibility and intervention in children with dysarthria. All had been diagnosed with CP by a neurologist. The children ranged in age from 4;6 to 17;5 (years;months), with a mean age of 9;7. Dysarthria was diagnosed in all participants by three experienced SLPs, who independently reviewed audio and video recordings of each child. Dysarthria diagnosis was based on notable speech deficits across at least two speech subsystems (e.g., short breath groups for speech, strained vocal quality, imprecise articulation, hypernasality, reduced speech rate, and monotone speech; Allison & Hustad, 2018; Levy et al., 2021), supported by associated confirmatory signs, including abnormal orofacial and respiratory movements. Pure phonological disorders, childhood apraxia of speech, and other speech disorders were ruled out based on the children's speech characteristics (Levy et al., 2017, 2021). The SLPs reached 100% agreement on the presence of dysarthria. Additional inclusion criteria were the use of speech as the primary communication modality, passing a bilateral hearing screening at 20 dB HL (at 500, 1000, 2000, and 4000 Hz), the ability to follow simple directions, and English as the dominant language. Participant characteristics, including Gross Motor Function Classification System scores and language comprehension skills, are listed in Table 1. Receptive language skills were assessed using selected subtests from either the Clinical Evaluation of Language Fundamentals–Fifth Edition (Wiig et al., 2013) or the Test for Auditory Comprehension of Language–Third Edition (Carrow-Woolfolk, 1999). It is important to note that most standardized language tests are normed on children with typical development. Thus, such tests may not fully capture the receptive language abilities of children with CP, in part because many children with CP also experience visual deficits, motor difficulties with pointing, and other challenges (Hustad et al., 2012; Levy et al., 2017, 2021).[3]

Human Listeners

A total of 168 American-English–speaking adult listeners (86 women, 82 men), aged 18–40 years (M = 29.2 years, SD = 6.12), participated in a listening task via the crowdsourcing website Prolific Academic (Palan & Schitter, 2018). Exclusion criteria were a history of speech, language, or hearing disorders, as well as long-term health conditions or disabilities besides CP.

ASR Systems

Five speaker-independent ASR systems, including both open-source and commercial systems trained on diverse speech data sets, were evaluated for their comparability to human listeners in assessing speech intelligibility in children with dysarthria. Three types of speaker-independent ASR systems were evaluated in this study: open-source, self-supervised, and commercial. The open-source systems included three variants of WhisperX (i.e., WhisperX-small, WhisperX-medium, and WhisperX-large) transformer-based models, fine-tuned for automatic speech transcription (Bain et al., 2023). WhisperX has demonstrated strong performance in transcription accuracy for decoding disfluent speech (Tetzloff et al., 2024).[4] The three different WhisperX model sizes tested (small, medium, large) reflect a trade-off between computational efficiency and transcription accuracy. Larger models typically achieve higher accuracy but require greater computational resources (e.g., memory, hardware capacity) compared to smaller models (Graham & Roll, 2024; Ma et al., 2024; Teleki et al., 2024). The fourth ASR system tested was Wav2Vec2, a widely used, self-supervised ASR system, which learns to recognize speech patterns by analyzing large amounts of unlabeled audio rather than using explicit speech-to-text pairs. Because it is trained on unlabeled speech, Wav2Vec2 captures general acoustic–phonetic representations but may require fine-tuning on clinical speech to achieve strong performance (Baevski et al., 2020; Tetzloff et al., 2024). This system has been used

[3]Speech data were also collected from children with typical development but are not reported in the present study, which aimed to determine the potential of ASR for approximating human judgments of intelligibility of the speech of children with dysarthria.

[4]Disfluent speech here refers to interruptions in the flow of speech, including interjections (e.g., "um," "uh"), parentheticals (e.g., "You know," "I mean"), and edited speech (e.g., revisions or restarts). These interruptions can serve various communicative functions, such as signaling delays in speech processing or reflecting difficulties in speech production (Teleki et al., 2024).

Table 1. Demographic and language characteristics of children with dysarthria due to cerebral palsy (CP).

Child	Age (years; months)	Sex	CP type	GMFCS	Dysarthria severity	Receptive language test		
						Test	Percentile rank	Interpretation
CP01	4;6	F	Spastic hemiplegia	I	Moderate-to-severe	TACL	91st	Above average
CP02	5;7	M	Spastic quadriplegia	IV	Moderate	TACL	< 1st	Very poor
CP03	5;11	F	Spastic diplegia	IV	Moderate	TACL	73rd	Average
CP04	6;8	M	Ataxic diplegia	III	Mild-to-moderate	TACL	< 1st	Very poor
CP05	7;1	M	Spastic triplegia	IV	Moderate-to-severe	TACL	9th	Below average
CP06	7;7	F	Spastic hemiplegia	IV	Mild-to-moderate	n/a	n/a	Average[a]
CP07	7;8	M	Spastic diplegia	III	Moderate	TACL	9th	Below average
CP08	8;3	F	Dyskinetic–hypotonic	II	Moderate	TACL	37th	Average
CP09	8;8	M	Spastic diplegia	II	Mild	TACL	37th	Average
CP10	9;1	F	Spastic hemiplegia	II	Mild	TACL	25th	Average
CP11	9;6	M	Ataxic diplegia	I	Moderate	TACL	< 1st	Very poor
CP12	10;2	M	Spastic triplegia	III	Moderate	TACL	37th	Average
CP13	10;8	M	Spastic quadriplegia	V	Moderate	CELF	0.5th	Poor
CP14	11;0	M	Spastic quadriplegia	V	Mild-to-moderate	TACL	< 1st	Very poor
CP15	11;3	M	Spastic hemiplegia	I	Moderate	CELF	95th	Above average
CP16	11;8	M	Spastic diplegia	III	Severe	TACL	5th	Poor
CP17	12;1	M	Spastic quadriplegia	IV	Mild	CELF	25th	Below average
CP18	13;4	F	Spastic quadriplegia	IV	Mild-to-moderate	CELF	0.4th	Poor
CP19	14;11	M	Spastic quadriplegia, epilepsy, VP shunt	V	Severe	CELF	0.4th	Poor
CP20	17;5	F	Spastic hemiparesis	I	Severe	CELF	50th	Average

Note. GMFCS rating: I = no/mild impairment, V = severe impairment. GMFCS = Gross Motor Function Classification System (Palisano et al., 1997); F = female; TACL = Test for Auditory Comprehension of Language; M = male; CELF = Clinical Evaluation of Language Fundamentals; VP = ventriculoperitoneal.

[a]Language test recording was not available (n/a) for this child; however, based on responses to questions in language sample, receptive language was judged by three speech-language pathologists to be age appropriate.

to assess speech in individuals with primary progressive apraxia of speech, showing strong correlations between apraxia severity and ASR transcription accuracy (Tetzloff et al., 2024). Finally, the fifth ASR system tested was Google Cloud ASR, a proprietary system optimized for real-world transcription tasks (Google LLC, 2020). This system has been used to evaluate intelligibility and speech severity in individuals with amyotrophic lateral sclerosis (ALS), demonstrating strong correlations with human transcriptions (Gutz et al., 2022), and in individuals with PD (Moya-Galé et al., 2022). Unlike Wav2-Vec2, the WhisperX and Google Cloud ASR systems are trained on paired audio–text data sets, providing direct word mappings between acoustic input and linguistic output. Including this range of systems allowed us to evaluate both widely used research models and real-world applications, as well as to explore trade-offs between accuracy and usability.

Speech Samples

Speech Stimuli

A total of 14 phrase and sentence stimuli were included in this study. Eleven phrases and sentences ranging from four to seven words in length from the Test of Children's Speech (Hodge et al., 2007) were selected for the children of all dysarthria severity levels to produce. These stimuli are developmentally appropriate for young children in terms of lexical, phonetic, syntactic, and morphological complexity (Hodge et al., 2007). Additionally, three sentences ranging from four to seven words in length (i.e., "Buy Bobby a puppy," "The blue spot is on the key," and "The potato stew is in the pot") were selected for their range of vowels and consonants (Weismer, 1984) for the current study and for future phonemic-level analyses of ASR transcription performance. The sentences contain simple words with some early-developing speech sounds. A total of 280 recordings of phrases and sentences were collected (20 Children × 14 Utterances).

Speech Recording Procedure

The children with dysarthria participated in a single recording session held in a quiet room at Teachers College, Columbia University (see Levy et al., 2017, for detailed recording procedure). Because vocal intensity is a key variable in intelligibility (C. M. Fox & Boliek, 2012;

Levy et al., 2020; Švec & Granqvist, 2018), a calibration procedure was performed to maximize the accuracy of sound pressure level (SPL) measurement for later replication in the listening task. During calibration, the experimenter audio-recorded a pure tone produced by a tuner (OT-120, Korg Orchestral) positioned 8 cm away from a sound-level meter (Galaxy Check Mate CM-140) and noted the tone's SPL as indicated on the sound-level meter. For the recording, a Countryman EMW Lavalier microphone was attached to the child's forehead, positioned 8 cm from the child's lips by means of a headband. Speech recordings were made with Sound Forge 17.0 software (VEGAS Creative Software) on a Dell OptiPlex 7090 computer, connected via a Scarlett 2i2 audio interface (Focusrite 2 × 2 USB2). The input dial setting was kept constant throughout the study. The children's speech was recorded at a sampling rate of 22050 Hz with 16-bit resolution in mono.

Tasks

Children's Speech Task

During the speech recording session, the children with dysarthria were asked to repeat the target phrases produced by a model speaker. These phrases were presented through loudspeakers (Altec Lansing ADA 215) placed at a consistent distance of 30 cm from them (Hall et al., 1968). The model speaker for the recordings was a female adult native of American English from the New York City area.

Adults' Listening Task: Human Orthographic Transcription and EoU Ratings

In order for listeners to evaluate the children's speech recordings, a total of 586 naive listeners were recruited via the crowdsourcing website Prolific (Palan & Schitter, 2018). Crowdsourcing has been validated for listener testing in previous research, demonstrating consistency with in-lab testing results and offering a cost-effective and scalable approach to data collection (Jiao et al., 2019; McAllister Byun et al., 2015; Stipancic et al., 2025). Listeners accessed the testing via Prolific and were asked through written instructions to sit in a quiet environment, without wearing headphones (to be more representative of real-world environments), positioned 30 cm from the loudspeaker of their computer. They were further instructed to download SoundMeter X, a sound-level meter app found to be one of the most accurate and reliable (Kardous & Shaw, 2014), on their phone and place their phone 8 cm (3 in.) from the left loudspeaker of their computer, in order to maintain a consistent calibration position across listeners.

The following listener calibration process was designed to replicate the original SPL of the child's speech (Hwang et al., 2022; Švec & Granqvist, 2018). First, listeners were instructed to click on an arrow to play a pure tone. The SPL of this tone matched the original pure-tone SPL that had been measured (at 8 cm from the sound-level meter) during the speech recording calibration procedure. The listeners were then instructed to adjust their computer's volume until their phone's sound-level meter app displayed the corresponding target SPL. For example, if the original pure tone was recorded at 70 dB SPL, listeners were presented with the tone and asked to adjust their computer's volume until their phone app displayed 70 dB SPL. In subsequent calibration steps, listeners were asked to measure the SPL of three additional pure-tone sound files (40–80 dB SPL) using the same app. For each step, they typed the most stable dB SPL reading into the testing interface. Listeners were given unlimited attempts, but progression to the next step was restricted until the entered dB SPL values fell within a predetermined acceptable range of ±1 dB (Švec & Granqvist, 2018), optimizing accurate calibration before beginning the listening session. Because the calibration tones spanned a wide range, it was highly unlikely that listeners would guess the dB SPL values. Only listeners who successfully completed all four calibration steps accurately were able to proceed. They were subsequently instructed to not adjust their speaker volume during the perceptual task so that listening conditions would be consistent throughout the experiment.

The listening session began with a familiarization task, in which listeners completed a brief practice session to become accustomed to the task format. For this task, they transcribed (orthographically) four speech samples from a child with dysarthria whose recordings were not included in the main experiment. Following the familiarization task, listeners proceeded to the experimental listening tasks. The instructions for the experimental task were: "Type the words you think the child said" for orthographic transcription. Listeners first heard the recorded speech produced by a child with dysarthria and were then instructed to type the words. They were permitted to play the speech sample no more than twice. For the EoU rating task, the experiment's instructions were: "How easy was the phrase/sentence to understand?" The listeners were instructed to rate how easy the phrase/sentence was to understand by sliding a marker along a VAS, with anchor points from *difficult* to *easy* (Levy et al., 2017, 2021). For intrarater reliability assessment, 20% of the sentences were randomly selected and replayed to listeners during the task (Allison & Hustad, 2018; Levy et al., 2017).

Of the 586 individuals who initiated the study, 168 completed the full listening task, with many exiting during the calibration phase. As considered in the Discussion section, this attrition may have been due to the added step of downloading an external app and setting up the environment, rather than confusion or poor usability. However, each child's speech was ultimately rated by an average of nine listeners, a sample size consistent with prior

crowdsourcing research demonstrating that valid speech ratings can be obtained with approximately nine listeners per speaker (McAllister Byun et al., 2015). Post-task feedback from participants who completed the study indicated that the instructions were clear and easy to follow.

Transcription Generation Task: ASR Transcription

For ASR transcription generation, the five speaker-independent ASR systems (WhisperX-large, WhisperX-medium, WhisperX-small, Google Cloud ASR, Wav2Vec2) were employed. These systems converted the spoken utterances of the children into written transcriptions.

Analysis

Human Transcription and VAS Ratings

Human transcriptions were analyzed in terms of the percentage of words correct, that is, the number of correctly transcribed words divided by the total number of words in each sentence. Words were counted as correct if they matched the target exactly, were homonyms, or were clear misspellings of the target or of its homonym (Hustad, 2008; Levy et al., 2017). To complement orthographic transcription and ASR-based measures in the present study, VAS ratings were used to derive the EoU variable. VAS ratings provide a rapid and sensitive index of listeners' perceived EoU and are particularly useful for capturing gradations of intelligibility across clinical populations (Abur et al., 2019; Hustad, 2006; Levy et al., 2017, 2021; Stipancic et al., 2016; Tjaden et al., 2014).

ASR Transcription

ASR transcriptions were assessed using the WER, which is a gold standard metric for evaluating speech recognition systems (Gutz et al., 2022, 2023; Tetzloff et al., 2024). The WER is the sum of the number of substitutions (S), insertions (I), and deletions (D) divided by the total number of words in the target utterance (N), as shown in Equation 1 (Moore et al., 2018).

$$\mathrm{WER} = \frac{(S + I + D)}{N} \tag{1}$$

Comparison Between Two Modalities

For the purpose of generating a single metric indicating percentage of words correctly transcribed for comparing human transcriptions with ASR, WRR was determined. This was calculated by means of the formula WRR = 1 − WER. The WRR indicates the percentage of words correctly transcribed on a scale from 0% to 100% (Gutz et al., 2022). Because WER can exceed 1 (e.g., if ASR detects an extra word produced by a speaker), resulting in negative WRR values, any negative WRR values were replaced with 0% (Gutz et al., 2022). A higher WRR indicates greater transcription accuracy. Distortions or misarticulations of the target utterance were scored within the WER formula as word-level substitution, insertion, or deletion errors, depending on the deviation. For example, if the target utterance was "three little pink pigs" and the child produced "three little pink pig," which ASR also transcribed as "three little pink pig," this would be scored as one substitution error out of four target words. Thus, the WER would be 1/4 = 0.25, and the WRR would be 1 − 0.25 = 0.75. Human listeners' transcriptions were scored in the same manner, for consistency of analysis between ASR and human transcription.

Statistical Analysis

All statistical analyses were completed in R (R Posit Team, 2023). To address our first research aim, ASR transcription accuracy (i.e., ASR WRR) was compared with human listeners' transcription accuracy (i.e., human WRR) for the speech produced by the children with dysarthria. Due to violations of the normality assumption as assessed by the Kolmogorov–Smirnov test ($p < .05$), Spearman rank-order correlations were used to examine the relationship between ASR WRR and human WRR (Jacks et al., 2019; Tetzloff et al., 2024).[5] To evaluate each ASR system's performance, the WRR of the five speaker-independent ASR systems (WhisperX-small, WhisperX-medium, WhisperX-large, Google Cloud, Wav2Vec2) was compared with human WRR. Spearman correlation coefficients were calculated between each ASR system's WRR and human WRR, with stronger correlations suggesting greater alignment with human listeners' transcription accuracy.

For our second research aim—to assess the use of ASR in representing human perceptual ratings—the relationship between ASR transcription accuracy and perceptual ratings was examined. For each of the five ASR systems, Spearman correlation coefficients were calculated between ASR WRR and human EoU ratings. A strong, positive correlation would support the use of ASR output as an approximation of human perceptual ratings in assessing intelligibility in children with dysarthria. To address multiple testing across 10 correlations (5 ASR Systems ×

[5]Given the modest sample size (20 speakers), Spearman rank correlations were used to evaluate the monotonic relationships between ASR transcription accuracy and human listener measures. While generalized linear models can handle nonnormal distributions and allow for simultaneous comparisons, our data set did not provide sufficient power for such models without overfitting. We therefore report effect-size estimates (ρ) with 95% confidence intervals. Because multiple correlations were conducted (5 ASR Systems × 2 Human Measures), we treat p values as exploratory and guard against Type I error by focusing on effect sizes/CIs. Adjusted p values are provided in Supplemental Material S1 for transparency.

2 Human Measures), a Holm–Bonferroni correction was conducted (see Supplemental Material S1).

In an assessment of intrarater and interrater reliabilities, intraclass correlation coefficients (ICCs) were computed for human listeners' transcription accuracy and EoU ratings. All reliability analyses were completed with the *irr* package (Gamer et al., 2012) and *lme4* package (Bates et al., 2015). Reliability was interpreted following Koo and Li (2016): ICC > .90 = excellent, .75–.90 = good, .50–.75 = moderate, and < .50 = poor. A single-rater, two-way model ICC(2,1) was used to measure for intrarater reliability, and a multirater, two-way model ICC(2, *k*) was used to measure for interrater reliability, accounting for consistency in transcription accuracy and ratings. To further examine whether individual factors contributed to variability in ASR WRR and human WRR, exploratory linear regression analyses were conducted with dysarthria severity and age as predictors of WRR in ASR systems and human listeners.

Results

Table 2 presents descriptive data on WRR for the speech of children with dysarthria, transcribed by human listeners and five ASR systems. The ASR performance showed considerable variability, with two systems (WhisperX-medium and WhisperX-large) even exceeding human WRR for this population. WhisperX-large and WhisperX-medium achieved the highest ASR WRR (48.24% and 46.04%, respectively), surpassing human listeners (38.95%). Google Cloud yielded a WRR of 30.23%, while Wav2Vec2-960h-large demonstrated the lowest performance, with a WRR of 5.63%.

Addressing the first research aim—to determine the relationship between ASR transcription accuracy and human listeners' transcription accuracy—we found significant correlations between all ASR systems and human listeners, as shown in Figure 1. The WRR of WhisperX-medium demonstrated the strongest correlation with human WRR ($\rho = .83$, $p < .001$). The WRR of WhisperX-large and WhisperX-small also demonstrated strong correlations with human WRR ($\rho = .82$, $p < .001$ [for WhisperX-large]; $\rho = .79$, $p < .001$ [for WhisperX-small]). The WRR of Google Cloud also showed strong alignment with human WRR ($\rho = .77$, $p < .001$). In contrast, the WRR of Wav2Vec2 exhibited a weak correlation with human WRR ($\rho = .34$, $p < .001$).

Table 2. Word recognition rate (WRR) for the speech of children with dysarthria performed by 168 human listeners and five automatic speech recognition (ASR) systems.

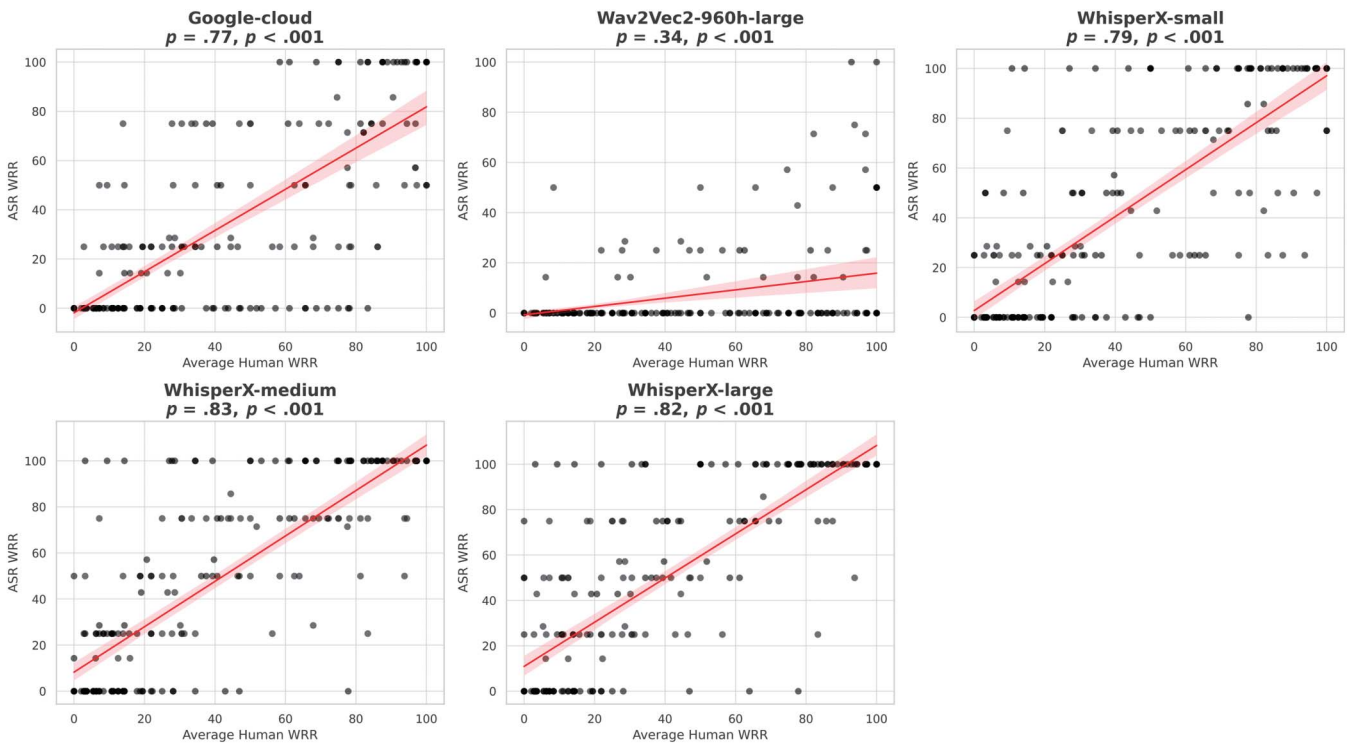
Modality	WRR (%), *M* (*SD*)
Human listeners	38.95 (40.43)
Google Cloud (ASR)	30.23 (37.43)
Wav2Vec2-960h-Large (ASR)	5.63 (16.17)
WhisperX-small (ASR)	38.90 (40.86)
WhisperX-medium (ASR)	46.04 (41.78)
WhisperX-large (ASR)	48.24 (41.82)

Addressing the second research aim, to determine the relationship between ASR transcription accuracy and human perceptual ratings, Spearman correlations were calculated between ASR WRR and average human EoU ratings in the five ASR systems. As shown in Figure 2, significant positive correlations were found for all ASR systems, suggesting that higher ASR transcription accuracy was associated with higher listener-rated intelligibility. The WRR of Google Cloud demonstrated the strongest correlation with EoU ratings ($\rho = .72$, $p < .001$), suggesting strong alignment with human perceptual ratings. The WRR of the WhisperX systems showed moderate correlations with human EoU ratings ($\rho = .67$–$.69$, $p < .001$). Wav2Vec2's WRR showed a weak correlation with human EoU ratings ($\rho = .28$, $p < .001$). Notably, the WRR of Wav2Vec2 demonstrated weak alignment with both human WRR and EoU ratings. This underperformance may reflect both the model's limited adaptation to atypical child speech and the absence of disorder-specific fine-tuning, as described in further detail in the Discussion section. All correlations remained significant after Holm–Bonferroni adjustment (adjusted *p*s < .001; see Supplemental Material S2), although ceiling and floor effects may have inflated some of the observed values.

The ICCs were computed to assess the intrarater and interrater reliabilities of human listeners' transcription accuracy and perceptual ratings across repeated trials during the listener testing. Intrarater reliability for human WRR was excellent, as indicated by ICC(2,1) = .945, $p < .001$. Interrater reliability for human WRR was good, with ICC(2, *k*) = .865, $p < .001$. Similarly, intrarater reliability for EoU ratings was good, as suggested by ICC(2,1) = .854, $p < .001$. Interrater reliability for EoU ratings was also good, as indicated by ICC(2, *k*) = .803, $p < .001$.

To illustrate individual variability, Supplemental Material S2 lists speaker-level WRRs for all ASR systems and human listeners. Exploratory regression analyses were performed to examine relationships between WRR and dysarthria severity and age. The analyses indicated that dysarthria severity was associated with WRRs for human listeners and all ASR systems, except Wav2Vec2 ($\beta = -21$ to -23, $p < .01$) such that children with milder dysarthria tended to yield higher WRR than those with more severe dysarthria. In contrast, age did not show a reliable relationship with WRR. Because these analyses were

Figure 1. Scatter plots showing the relationship between automatic speech recognition (ASR), word recognition rate (WRR), and average human WRR across five ASR systems for the speech of children with dysarthria.



exploratory and not central to the study aims, the findings should be interpreted with caution, and the patterns should be investigated more systematically in future work.

Discussion

This study examined the potential of ASR systems for assessing intelligibility in children with dysarthria secondary to CP. Overall, ASR WRR correlated with human WRR and EoU ratings, demonstrating ASR's capability as a tool for approximating human listeners' transcription accuracy and perceptual ratings, although the different ASR systems varied in their success in approximating listeners' perception. Implications for clinical and technological applications are discussed.

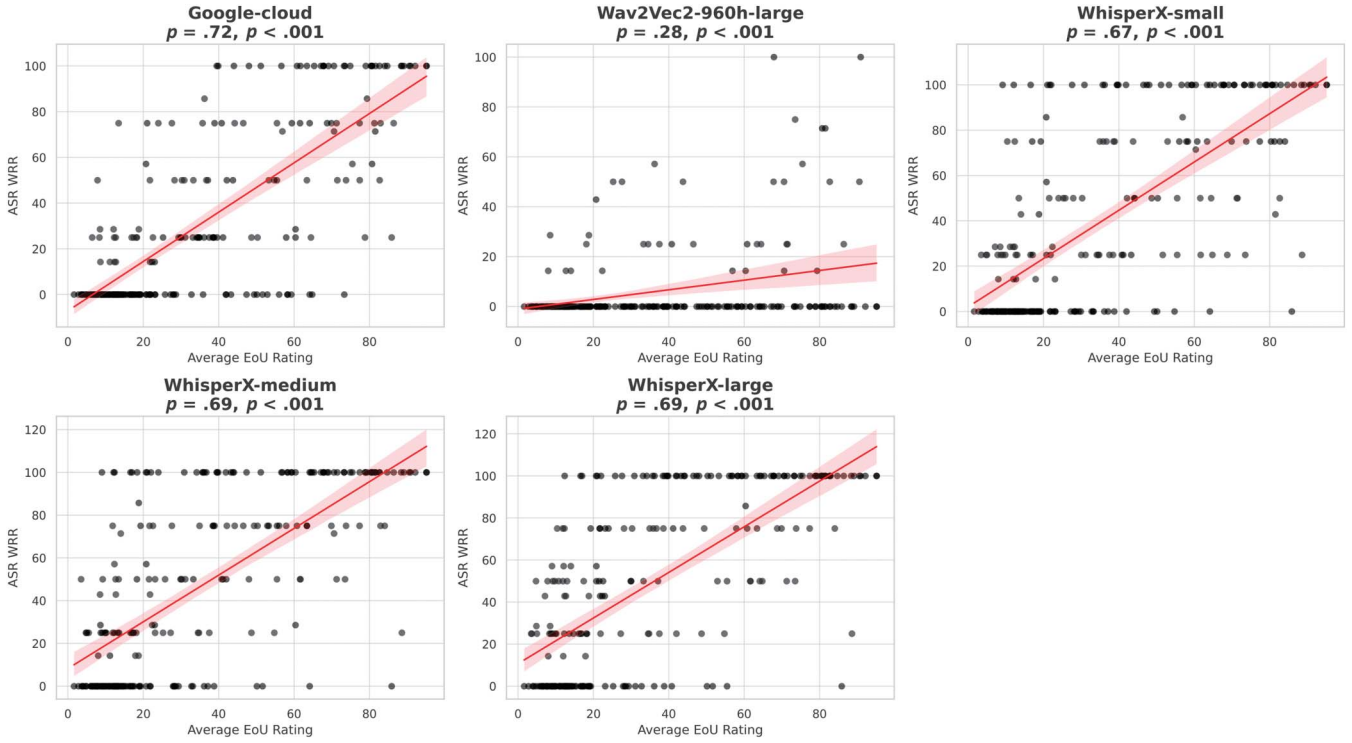
ASR Transcription and Human Transcription

In four of the five ASR systems tested, transcription accuracy (i.e., ASR WRR) yielded by WhisperX-small, WhisperX-medium, WhisperX-large, and Google Cloud demonstrated strong correlations with human listeners' transcription accuracy (i.e., human WRR), as had been hypothesized, supporting their potential clinical utility in children with dysarthria. In contrast, Wav2Vec2 exhibited a weaker correlation with human listeners' transcription accuracy, indicating that it needs further refinement for such purposes. The finding that transcription accuracy yielded by WhisperX-medium, rather than Google Cloud, demonstrated the strongest correlation with human listeners' transcription accuracy was unexpected (Attia et al., 2024). A possible explanation is that WhisperX-medium may have better captured the acoustic characteristics of children's speech in this data set, potentially due to differences in model architecture or training data. While this is the first examination of ASR for children with dysarthria, the findings are consistent with prior research evaluating ASR systems for transcribing the speech of adults with dysarthria (Gutz et al., 2022; Moya-Galé et al., 2022). The findings suggest that ASR could be useful if incorporated into intelligibility assessment, reducing the time and effort required.

ASR Transcription and Human EoU Ratings

ASR transcription accuracy (i.e., ASR WRR) also correlated with human EoU ratings, although the degree of alignment varied across ASR systems. Transcription accuracy yielded by Google Cloud exhibited a strong correlation with human EoU ratings, suggesting its potential as a tool for intelligibility assessment in children with dysarthria. The transcription accuracy of the three WhisperX

Figure 2. Scatter plots showing the relationship between automatic speech recognition (ASR) word recognition rate, and average human ease of understanding (EoU) ratings for five ASR systems for the speech of children with dysarthria.



systems demonstrated moderate correlations with human EoU ratings. However, these correlations were somewhat weaker than the correlation between Google Cloud's transcription accuracy and EoU ratings, suggesting that WhisperX systems may benefit from further refinement. Transcription accuracy of Wav2Vec2 showed a weak correlation with human EoU ratings, again revealing its limitations in assessing intelligibility in alignment with human listeners. These results support our hypothesis that, overall, ASR transcription accuracy would be associated with human perceptual ratings of intelligibility.

Clinical and Technological Implications

The findings suggest that WhisperX and Google Cloud ASR systems may be appropriate tools for use in the intelligibility assessment of children with dysarthria due to CP. Their strong correlations with human listeners' transcription accuracy and perceptual ratings support ASR as an efficient, objective tool that offers faster and more consistent assessments than human transcription alone. These systems can be implemented by clinicians and researchers with minimal technical expertise, potentially improving and expediting assessment and intervention processes.

From a clinical viewpoint, the findings highlight the importance of system selection and adaptation when using ASR for assessing intelligibility in children with dysarthria. For instance, while WhisperX-medium and WhisperX-large achieved even higher transcription accuracy than human listeners (see Table 2), this does not necessarily indicate stronger alignment with human perceptual ratings. As described in the first part, transcription accuracy may be particularly relevant for applications that require a precise record of speech output, such as those tracking articulation and phonological patterns over time or generating speech-to-text input for AAC devices (Chen et al., 2024; Hui et al., 2024). WhisperX-medium's transcription accuracy demonstrated the strongest correlation with human listeners' transcription accuracy, suggesting the system's utility as a proxy for human listeners when listeners' transcription is impractical, particularly for applications that require precise speech-to-text conversion, such as some AAC devices. In contrast, the transcription accuracy of Google Cloud, despite its lower accuracy (30.23%), demonstrated the strongest correlation with human perceptual ratings. This suggests that Google Cloud may be better suited for evaluating intelligibility and tracking speech progress in clinical contexts. Its alignment with human perception may stem from the system's ability to detect speech features that listeners use when judging dysarthric speech (Gutz et al., 2022).

Together, these findings indicate that while WhisperX-medium may be optimal for transcription-focused applications, both WhisperX-medium and Google Cloud show promise for clinical intelligibility assessment because each can approximate different aspects of human listener experience. WhisperX-medium's transcription accuracy aligns more closely with human listeners' transcription accuracy, whereas the transcription accuracy of Google Cloud better reflects subjective perceptual judgments. While the transcription accuracy of WhisperX systems and Google Cloud showed moderate-to-strong correlations with human perception of intelligibility, occasional "hallucinations" (i.e., instances where ASR systems generated words that were not present in the child's speech) were observed, particularly in WhisperX systems. Although these hallucinations were infrequent, they underscore the importance of continued refinement and validation to ensure consistent clinical reliability for the speech of children with dysarthria. From a clinical perspective, such hallucinations are not trivial, as they could misrepresent a child's speech output and lead to inaccurate estimates of intelligibility and/or inappropriate treatment planning. Future ASR research may mitigate these errors by incorporating pediatric disordered speech into training corpora, refining decoding strategies to reduce false insertions, and integrating clinician judgment to verify or contextualize ASR outputs for clinical decision making.

It is also important to note that overall WRR values for both ASR and human listeners were relatively low (< 50%; see Table 2). These low values likely reflect factors such as the severity of dysarthria in the speech sample, the stringent scoring method requiring exact word matches, and the absence of contextual cues that listeners would typically use in real-world communication. While two out of the five ASR systems produced higher WRR than human listeners, this should not be interpreted as evidence that children were more intelligible to ASR systems than to human listeners. As shown in Figure 1, the speech of a subset of the children yielded large discrepancies between ASR WRR and human WRR. These cases likely reflect methodological differences; as previously described, ASR depends strictly on acoustic–phonetic patterns, whereas human listeners can integrate contextual and prosodic cues but may also be influenced by perceptual biases. Such biases can sometimes lower WRR and raise concerns about the fairness and consistency of human-based intelligibility assessment, yet they may also provide valuable insights into how listeners perceive atypical speech in real-world contexts. In contrast, ASR does not incorporate social or affective judgments and therefore has the potential to reduce these listener-based biases. Nevertheless, current ASR systems may face challenges in capturing perceptual qualities (e.g., prosody, vocal quality, nasality), which could partly explain recognition errors (Goldwater et al., 2010; Hernandez et al., 2020; Maier et al., 2008; Moore, 2021; Tetzloff et al., 2024). Thus, clinical recommendations include that human-based measures be interpreted alongside ASR WRR to avoid mischaracterizing children's communicative abilities. In fact, if ASR systems were used in isolation, they could lead to underidentifying children's communicative difficulties, potentially affecting service eligibility decisions. Therefore, ASR output may be considered a useful complement to, rather than a replacement for, human listener judgments.

Furthermore, ASR systems and human listeners transcribed speech from children with milder dysarthria more accurately than from those with more severe dysarthria, suggesting ASR's sensitivity to clinically meaningful severity differences while also highlighting challenges for transcription of severely dysarthric speech. The WRR of ASR systems and human listeners was also higher for older children than for younger children, likely reflecting increases in speech motor control and phonological development over time. However, advancing ASR for clinical use will require models and training data sets that better capture the acoustic, phonetic, and developmental variability associated with both dysarthria severity and age-related speech changes.

Notably, Wav2Vec2's transcription accuracy did not align well with human perceptual judgments of the speech of children with dysarthria in this study. One possible explanation, as described in the Method section, is this system's self-supervised learning approach in transcriptions, which trains on large amounts of unlabeled audio without being paired with corresponding text or phonetic labels. Because it is not explicitly trained to map speech to specific phonemic or lexical targets, Wav2Vec2 may be less adaptable to the atypical and highly variable speech patterns characteristic of dysarthria. While self-supervised models such as Wav2Vec2 have performed well on neurotypical speech, they may not fare as well with dysarthric speech due to their lack of exposure to atypical phonetic variability during training. A recent study found that when Wav2Vec2 was fine-tuned (i.e., trained further using a smaller, labeled data set of dysarthric speech), its performance improved (Yang et al., 2022). Therefore, tailoring this model to disordered speech may be necessary for clinical applications. In contrast, most ASR systems (e.g., WhisperX, Google Cloud) are trained using paired speech–text data sets, allowing them to learn more direct mappings between speech input and expected linguistic outputs. Although these systems are usually trained on neurotypical speech, their structure may enable better generalization to intelligibility patterns, even in atypical speech. This performance gap likely reflects fundamental differences in training approaches: Wav2Vec2 is trained in a self-supervised manner on large amounts of unlabeled audio, without paired text supervision, whereas

WhisperX and similar models are trained on extensive paired speech–text data sets. As a result, WhisperX models may generalize more effectively to atypical pediatric speech, while Wav2Vec2 remains less robust in handling atypical pediatric speech without disorder-specific fine-tuning.

Limitations and Future Directions

Despite the promising findings on ASR for intelligibility assessment in this study, several limitations must be acknowledged. First, the data set was limited to dysarthric speech produced by children with CP, the majority of whom had spastic CP or a mix of spastic and other subtypes of CP (as indicated in Table 1). The children's speech characteristics were thus primarily, although variably, consistent with spastic dysarthria. While spastic CP is by far the most common subtype (Centers for Disease Control and Prevention, 2025), the results may not generalize to other populations or CP/dysarthria subtypes. Second, additional exploratory analyses suggested that severity was associated with WRR and that age did not show a reliable relationship with WRR in most ASR systems and human listeners. These exploratory findings may be followed up with more systematic study of larger, more representative samples, investigating the potential contributions of children's ages, speech features, and severity to ASR performance. Future work should also examine how these factors influence the comparability of ASR outcomes to human listener judgments.

A further limitation is the brevity of the speech stimuli used in this study, namely, sentences ranging from four to seven words. Although these short utterances were selected for their appropriateness for a range of children with dysarthria (Allison et al., 2019; Levy et al., 2017), studies on adult populations (e.g., those with ALS) suggest that speech severity and utterance length can interact to influence intelligibility (Allison et al., 2019; Gutz et al., 2022; Moya-Galé et al., 2022). Future work should determine any effects of utterance length on ASR performance and listener perception of children's speech. To enhance generalizability, future data sets could also incorporate longer utterances, varying levels of background noise, and larger samples spanning severity levels. Such additions would provide more representative input for ASR models and help identify conditions under which performance degrades.

An important limitation in interpreting the listener data is the relatively high attrition rate observed in the online task, with only approximately 30% of recruited participants completing the full protocol. While this completion rate is low, high attrition rates have also been reported in prior crowdsourced intelligibility studies that implemented strict quality control procedures (e.g., Hwang et al., 2022; Jiao et al., 2019). In our study, the added requirement of calibrating SPL using external apps likely contributed to the reduced completion rate. However, this rigor ensured that those who completed the protocol provided high-quality calibrated data that included the important dimension of vocal intensity (C. M. Fox & Boliek, 2012; Levy et al., 2020; Švec & Granqvist, 2018). Thus, what was lost in numbers was likely offset by gains in precision, as the final data set still included a large sample of listeners and robust transcription data for every child. However, it should be acknowledged that the inherent variability of online testing environments (e.g., differences in device type, speaker quality, and listener setup) may have introduced additional variability into the perceptual data. These factors should be taken into account when considering the generalizability of the findings.

Additionally, only five ASR systems were evaluated, and other systems or continued fine-tuning may yield different results. It will be important to explore a broader range of ASR systems (e.g., ElevenLabs Scribe, AssemblyAI) and assess their applicability across age groups, severity levels, types of dysarthria, and other speech disorders. Another limitation is the reliance on WRR as a primary metric for ASR performance. As noted in the Method section, distortions or misarticulations were scored as substitution, insertion, or deletion errors within the WER formula. Although WRR provides insight into word transcription accuracy, it does not specify phoneme-level errors or other distortions that may impact intelligibility (Miller, 2013). Further investigations may incorporate additional evaluation metrics, such as phoneme error rates and prosodic analyses, to gain a more nuanced understanding of ASR performance and the level of detail that ASR can provide on the speech production of children with dysarthria. Examining how different ASR architectures handle phoneme-level distortions and prosodic variations would support efforts in refining their clinical applicability in this population. Differences in ASR architecture raise important questions about the types of errors ASR systems make compared to human listeners. Human raters may be more sensitive to phoneme distortions and prosodic deviations, for example, while ASR errors often stem from misrecognition of specific phonemes or omitted segments (Southwell et al., 2022). Future research should further investigate these discrepancies to optimize ASR systems for dysarthric speech applications.

Certainly, translating these systems into practice presents challenges. Although, in theory, several ASR systems are available for consumer or research use, calculating WRR is not as straightforward as simply uploading an audio file to a website, for example. Current implementation requires additional steps such as audio preprocessing, transcription formatting, and metric computation, which may be beyond the scope of SLPs' routine clinical practice

or capabilities. Thus, further development of SLP-friendly tools is necessary before these systems can be directly integrated into practice.

To improve clinical applicability, future studies may also determine the real-time usability of ASR systems for providing feedback during speech treatment. Moreover, it would be informative to test SLPs as listeners and compare their judgments with ASR performance to further evaluate the clinical utility and reliability of ASR systems. Given that WhisperX systems may be particularly suited for integration into assistive communication devices, further research could explore their use in AAC contexts to enhance both ASR transcription accuracy and accessibility for individuals with severe speech impairments. A long-term goal will be the development of ASR systems specifically trained on the speech of children with dysarthria to optimize clinical utility.

Conclusions

This study provides evidence that ASR transcription accuracy yielded by various speaker-independent ASR systems aligns closely with both human listeners' transcription accuracy and ease-of-understanding ratings for the speech of children with dysarthria, although the degree of alignment varies across ASR systems. WhisperX-medium best approximated human listeners' transcription accuracy. Google Cloud emerged as the best proxy for human perceptual ratings. Certainly, ASR system selection should be guided by clinical goals, whether for intelligibility assessment, intervention planning, or assistive communication. In this first examination of ASR in children with dysarthria, the findings underscore the potential of ASR as a tool to be incorporated into intelligibility assessment protocols. The study also highlights the need for further ASR refinements to improve clinical applications and ultimately enhance communication outcomes for this population.

Data Availability Statement

In compliance with IRB guidelines, the speech data from this study are not publicly accessible. However, deidentified participant data in spreadsheet format may be available upon request from the first author.

Acknowledgments

The authors thank the children and their families for their participation, as well as the online listeners who contributed to the study. The authors also thank the research assistants, Jie Gao, Blake Vente, and Eunjung Yeo.

References

Abur, D., Enos, N. M., & Stepp, C. E. (2019). Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure. *American Journal of Speech-Language Pathology, 28*(3), 1222–1232. https://doi.org/10.1044/2019_AJSLP-18-0275

Allison, K. M. (2020). Measuring speech intelligibility in children with motor speech disorders. *Perspectives of the ASHA Special Interest Groups, 5*(4), 809–820. https://doi.org/10.1044/2020_PERSP-19-00110

Allison, K. M., & Hustad, K. C. (2018). Acoustic predictors of pediatric dysarthria in cerebral palsy. *Journal of Speech, Language, and Hearing Research, 61*(3), 462–478. https://doi.org/10.1044/2017_JSLHR-S-16-0414

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

Atkins, M. S., Boyce, S. E., MacAuslan, J., & Silbert, N. (2019). Computer-assisted syllable complexity analysis of continuous speech as a measure of child speech disorders. *Proceedings of the 19th International Congress of Phonetic Sciences, 2019,* 1054–1058. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1103.pdf [PDF]

Attia, A. A., Liu, J., Ai, W., Demszky, D., & Espy-Wilson, C. (2024). *Kid-Whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults*. arXiv. https://doi.org/10.48550/arXiv.2309.07927

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 33,* 12449–12460. https://dl.acm.org/doi/abs/10.5555/3495724.3496768

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. *Proceedings of Interspeech, 2023,* 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., & Bolker, M. B. (2015). Package 'lme4.' *Convergence, 12*(1), 1–139. https://cran.r-project.org/web/packages/lme4/lme4.pdf [PDF]

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual learning of dysarthric speech: A review of experimental studies. *Journal of Speech, Language, and Hearing Research, 55*(1), 290–305. https://doi.org/10.1044/1092-4388(2011/10-0349)

Carrow-Woolfolk, E. (1999). *Test for Auditory Comprehension of Language–Third Edition*. Pro-Ed.

Centers for Disease Control and Prevention. (2020, December). *Data and statistics for cerebral palsy*. https://archive.cdc.gov/#/details?url=https://www.cdc.gov/ncbddd/cp/data.html

Centers for Disease Control and Prevention. (2025, July). *About cerebral palsy*. https://www.cdc.gov/cerebral-palsy/about/index.html

Chen, S.-H. K., Saeli, C., & Hu, G. (2024). A proof-of-concept study for automatic speech recognition to transcribe AAC speakers' speech from high-technology AAC systems. *Assistive*

Technology, 36(4), 319–326. https://doi.org/10.1080/10400435.2023.2260860

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

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., Sinex, D. G., & Liss, J. M. (2017). Predicting intelligibility gains in individuals with dysarthria from baseline speech features. *Journal of Speech, Language, and Hearing Research, 60*(11), 3043–3057. https://doi.org/10.1044/2016_JSLHR-S-16-0218

Fox, C. B., Israelsen-Augenstein, M., Jones, S., & Gillam, S. L. (2021). An evaluation of expedited transcription methods for school-age children's narrative language: Automatic speech recognition and real-time transcription. *Journal of Speech, Language, and Hearing Research, 64*(9), 3533–3548. https://doi.org/10.1044/2021_JSLHR-21-00096

Fox, C. M., & Boliek, C. A. (2012). Intensive voice treatment (LSVT LOUD) for children with spastic cerebral palsy and dysarthria. *Journal of Speech, Language, and Hearing Research, 55*(3), 930–945. https://doi.org/10.1044/1092-4388(2011/10-0235)

Gale, R., Chen, L., Dolata, J., van Santen, J., & Asgari, M. (2019). Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques. *Proceedings of Interspeech, 2019,* 11–15. https://doi.org/10.21437/Interspeech.2019-3161

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr.' *Various Coefficients of Interrater Reliability and Agreement, 22,* 1–32. https://cran.r-project.org/web/packages/irr/irr.pdf [PDF]

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

Google LLC. (2020). *Speech-to-text: Turn speech into text using Google AI*. Google Cloud. https://cloud.google.com/speech-to-text

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters, 4*(2), Article 025206. https://doi.org/10.1121/10.0024876

Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M. A., Tobin, J., Brenner, M. P., Nelson, P. C., & Tomanek, K. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. *Proceedings of Interspeech, 2021,* 4778–4782. https://doi.org/10.21437/Interspeech.2021-1384

Gutz, S. E., Maffei, M. F., & Green, J. R. (2023). Feedback from automatic speech recognition to elicit clear speech in healthy speakers. *American Journal of Speech-Language Pathology, 32*(6), 2940–2959. https://doi.org/10.1044/2023_AJSLP-23-00030

Gutz, S. E., Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2022). Validity of off-the-shelf automatic speech recognition for assessing speech intelligibility and speech severity in speakers with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 65*(6), 2128–2143. https://doi.org/10.1044/2022_JSLHR-21-00589

Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold, A. R., Durbin, M., Edmonson, M. S., Fischer, J. L., Hymes, D., Kimball, S. T., La Barre, W., McClellan, J. E., Marshall, D. S., Milner, G. B., Sarles, H. B., Trager, G. L., & Vayda, A. P. (1968). Proxemics [and comments and replies]. *Current Anthropology, 9*(2/3), 83–108. https://doi.org/10.1086/200975

Hernandez, A., Kim, S., & Chung, M. (2020). Prosody-based measures for automatic severity assessment of dysarthric speech. *Applied Sciences, 10*(19), Article 6999. https://doi.org/10.3390/app10196999

Hodge, M., Daniels, J., & Gotzke, C. L. (2007). *TOCS+ intelligibility measures*. University of Alberta.

Hui, M., Zhang, J., & Mohan, A. (2024). *Enhancing AAC software for dysarthric speakers in e-health settings: An evaluation using TORGO*. arXiv. https://doi.org/10.48550/arXiv.2411.00980

Hustad, K. C. (2006). Estimating the intelligibility of speakers with dysarthria. *Folia Phoniatrica et Logopaedica, 58*(3), 217–228. https://doi.org/10.1159/000091735

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

Hustad, K. C., Schueler, B., Schultz, L., & DuHadway, C. (2012). Intelligibility of 4-year-old children with and without cerebral palsy. *Journal of Speech, Language, and Hearing Research, 55*(4), 1177–1189. https://doi.org/10.1044/1092-4388(2011/11-0083)

Hwang, K., Chang, Y., Berisha, V., McAuliffe, M., van Brenk, F., Choi, J., Brisman, R., Jeong, J., Hanselman, E., Hong, E., & Levy, E. S. (2022). *Validity of online recordings in children with dysarthria: Acoustic and perceptual measures* [Poster presentation]. Annual Convention of the American Speech-Language-Hearing Association, New Orleans, LA, United States (Hybrid).

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

Jiao, Y., LaCross, A., Berisha, V., & Liss, J. (2019). Objective intelligibility assessment by automated segmental and suprasegmental listening error analysis. *Journal of Speech, Language, and Hearing Research, 62*(9), 3359–3366. https://doi.org/10.1044/2019_JSLHR-S-19-0119

Kardous, C. A., & Shaw, P. B. (2014). Evaluation of smartphone sound measurement applications. *The Journal of the Acoustical Society of America, 135*(4), EL186–EL192. https://doi.org/10.1121/1.4865269

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104*(3), 1598–1608. https://doi.org/10.1121/1.424372

Landa, S., Pennington, L., Miller, N., Robson, S., Thompson, V., & Steen, N. (2014). Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International Journal of Speech-Language Pathology, 16*(4), 408–416. https://doi.org/10.3109/17549507.2014.927922

Levy, E. S., Chang, Y. M., Ancelle, J. A., & McAuliffe, M. J. (2017). Acoustic and perceptual consequences of speech cues for children with dysarthria. *Journal of Speech, Language, and Hearing Research, 60*(6S), 1766–1779. https://doi.org/10.1044/2017_JSLHR-S-16-0274

Levy, E. S., Chang, Y. M., Hwang, K., & McAuliffe, M. J. (2021). Perceptual and acoustic effects of dual-focus speech treatment in children with dysarthria. *Journal of Speech, Language, and Hearing Research, 64*(6S), 2301–2316. https://doi.org/10.1044/2020_JSLHR-20-00301

Levy, E. S., Moya-Galé, G., Chang, Y. M., Freeman, K., Forrest, K., Brin, M. F., & Ramig, L. A. (2020). The effects of intensive

speech treatment on intelligibility in Parkinson's disease: A randomised controlled trial. *EClinicalMedicine, 24,* Article 100429. https://doi.org/10.1016/j.eclinm.2020.100429

Lilley, J., Nittrouer, S., & Bunnell, H. T. (2014). Automating an objective measure of pediatric speech intelligibility. *Proceedings of Interspeech, 2014,* 1578–1582. https://doi.org/10.21437/Interspeech.2014-376

Ma, H., Peng, Z., Shao, M., Li, J., & Liu, J. (2024). *Extending Whisper with prompt tuning to target-speaker ASR*. arXiv. https://doi.org/10.48550/arXiv.2312.08079

Maier, A., Hacker, C., Noth, E., Nkenke, E., Haderlein, T., Rosanowski, F., & Schuster, M. (2006). Intelligibility of children with cleft lip and palate: Evaluation by speech recognition techniques. *Proceedings of 18th International Conference on Pattern Recognition (ICPR'06),* 274–277. https://doi.org/10.1109/ICPR.2006.718

Maier, A., Reuß, A., Hacker, C., Schuster, M., & Nöth, E. (2008). Analysis of hypernasal speech in children with cleft lip and palate. *Text, Speech and Dialogue: 11th International Conference, TSD 2008, Brno, Czech Republic, September 8-12, 2008, Proceedings,* 389–396. https://doi.org/10.1007/978-3-540-87391-4_50

McAllister Byun, T., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders, 53,* 70–83. https://doi.org/10.1016/j.jcomdis.2014.11.003

Mei, C., Reilly, S., Bickerton, M., Mensah, F., Turner, S., Kumaranayagam, D., Pennington, L., Reddihough, D., & Morgan, A. T. (2020). Speech in children with cerebral palsy. *Developmental Medicine & Child Neurology, 62*(12), 1374–1382. https://doi.org/10.1111/dmcn.14592

Meylan, S. C., Foushee, R., Wong, N. H., Bergelson, E., & Levy, R. P. (2023). How adults understand what young children say. *Nature Human Behaviour, 7*(12), 2111–2125. https://doi.org/10.1038/s41562-023-01698-3

Microsoft. (2020). *Cortana: Your personal productivity assistant in Microsoft 365*. https://www.microsoft.com/en-us/cortana

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

Moore, M. (2021). Speech recognition for individuals with voice disorders. In T. McDaniel & X. Liu (Eds.), *Multimedia for accessible human computer interfaces* (pp. 115–144). Springer. https://doi.org/10.1007/978-3-030-70716-3_5

Moore, M., Venkateswara, H., & Panchanathan, S. (2018). Whistle-blowing ASRs: Evaluating the need for more inclusive speech recognition systems. *Proceedings of Interspeech, 2018,* 466–470. https://doi.org/10.21437/Interspeech.2018-2391

Moya-Galé, G., Walsh, S. J., & Goudarzi, A. (2022). Automatic assessment of intelligibility in noise in Parkinson disease: Validation study. *Journal of Medical Internet Research, 24*(10), Article e40567. https://doi.org/10.2196/40567

Nuance. (2020). *Dragon Speech*. https://www.dragon-speech.us/

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17,* 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Palisano, R., Rosenbaum, P., Walter, S., Russell, D., Wood, E., & Galuppi, B. (1997). Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Developmental Medicine & Child Neurology, 39*(4), 214–223. https://doi.org/10.1111/j.1469-8749.1997.tb07414.x

R Posit Team. (2023). *RStudio: Integrated development environment for R*. Posit Software, PBC.

Rodrigues, A., Santos, R., Abreu, J., Beça, P., Almeida, P., & Fernandes, S. (2019). Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender. *Interacción '19: Proceedings of the XX International Conference on Human Computer Interaction,* 1–8. https://doi.org/10.1145/3335595.3335635

Rowe, H. P., Gutz, S. E., Maffei, M. F., Tomanek, K., & Green, J. R. (2022). Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective. *Frontiers in Computer Science, 4,* Article 770210. https://doi.org/10.3389/fcomp.2022.770210

Schultz, B. G., Tarigoppula, V. S. A., Noffs, G., Rojas, S., van der Walt, A., Grayden, D. B., & Vogel, A. P. (2021). Automatic speech recognition in neurodegenerative disease. *International Journal of Speech Technology, 24*(3), 771–779. https://doi.org/10.1007/s10772-021-09836-w

Shakhadri, S. A. G., Kruthika, K., & Angadi, K. B. (2025). *Samba-ASR: State-of-the-art speech recognition leveraging structured state–space models*. arXiv. https://doi.org/10.48550/arXiv.2501.02832

Southwell, R., Pugh, S., Perkoff, E. M., Clevenger, C., Bush, J., Lieber, R., Ward, W., Foltz, P., & D'Mello, S. (2022). Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. *Proceedings of the 15th International Educational Data Mining Society,* 302–315. https://doi.org/10.5281/ZENODO.6853109

Stipancic, K. L., Brenk, F., Qiu, M., & Tjaden, K. (2025). Progress toward estimating the minimal clinically important difference of intelligibility: A crowdsourced perceptual experiment. *Journal of Speech, Language, and Hearing Research, 68*(7S), 3480–3494. https://doi.org/10.1044/2024_JSLHR-24-00354

Stipancic, K. L., Tjaden, K., & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research, 59*(2), 230–238. https://doi.org/10.1044/2015_JSLHR-S-15-0271

Strömbergsson, S., Edlund, J., McAllister, A., & Lagerberg, T. (2021). Understanding acceptability of disordered speech through Audience Response Systems–based evaluation. *Speech Communication, 131,* 13–22. https://doi.org/10.1016/j.specom.2021.05.005

Švec, J. G., & Granqvist, S. (2018). Tutorial and guidelines on measurement of sound pressure level in voice and speech. *Journal of Speech, Language, and Hearing Research, 61*(3), 441–461. https://doi.org/10.1044/2017_JSLHR-S-17-0095

Teleki, M., Dong, X., Kim, S., & Caverlee, J. (2024). Comparing ASR systems in the context of speech disfluencies. *Proceedings of Interspeech, 2024,* 4548–4552. https://doi.org/10.21437/Interspeech.2024-1270

Tetzloff, K. A., Wiepert, D., Botha, H., Duffy, J. R., Clark, H. M., Whitwell, J. L., Josephs, K. A., & Utianski, R. L. (2024). Automatic speech recognition in primary progressive apraxia of speech. *Journal of Speech, Language, and Hearing Research, 67*(9), 2964–2976. https://doi.org/10.1044/2024_JSLHR-24-00049

Tjaden, K., Kain, A., & Lam, J. (2014). Hybridizing conversational and clear speech to investigate the source of increased intelligibility in speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 57*(4), 1191–1205. https://doi.org/10.1044/2014_JSLHR-S-13-0086

Tobin, J., Nelson, P., MacDonald, B., Heywood, R., Cave, R., Seaver, K., Desjardins, A., Jiang, P.-P., & Green, J. R. (2024). Automatic speech recognition of conversational speech in individuals with disordered speech. *Journal of Speech, Language, and Hearing Research, 67*(11), 4176–4185. https://doi.org/10.1044/2024_JSLHR-24-00045

Tu, M., Wisler, A., Berisha, V., & Liss, J. M. (2016). The relationship between perceptual disturbances in dysarthric speech

and automatic speech recognition performance. *The Journal of the Acoustical Society of America, 140*(5), EL416–EL422. https://doi.org/10.1121/1.4967208

Weismer, G. (1984). Articulatory characteristics of Parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal–supraglottal coordination. In M. McNeil, J. Rosenbek, & A. Aronson (Eds.), *The dysarthrias: Physiology, acoustics, perception and management* (pp. 101–129). College Hill Press.

Wiig, E. H., Semel, E. M., & Secord, W. (2013). *Clinical Evaluation of Language Functions–Fifth Edition Screening Test*. Pearson.

Wilpon, J. G., & Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 1,* 349–352. https://doi.org/10.1109/ICASSP.1996.541104

Xu, D., Richards, J. A., & Gilkerson, J. (2014). Automated analysis of child phonetic production using naturalistic recordings. *Journal of Speech, Language, and Hearing Research, 57*(5), 1638–1650. https://doi.org/10.1044/2014_JSLHR-S-13-0037

Yang, M., Hirschi, K., Looney, S. D., Kang, O., & Hansen, J. H. L. (2022). Improving mispronunciation detection with Wav2Vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment. *Proceedings of Interspeech, 2022,* 4481–4485. https://doi.org/10.21437/Interspeech.2022-11039

Yeung, G., & Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. *Proceedings of Interspeech, 2018,* 1661–1665. https://doi.org/10.21437/Interspeech.2018-2297

Yılmaz, E., Mitra, V., Sivaraman, G., & Franco, H. (2019). Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. *Computer Speech & Language, 58,* 319–334. https://doi.org/10.1016/j.csl.2019.05.002