Validity of Automatic Speech Recognition for Intelligibility Assessment in Children with Dysarthria

Jiyoung Choi

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

Abstract

Validity of Automatic Speech Recognition for Intelligibility Assessment

in Children with Dysarthria

Jiyoung Choi

Purpose: Accurate assessment of speech intelligibility is critical for children with dysarthria secondary to cerebral palsy (CP). Traditional human assessment, such as orthographic transcription and perceptual ratings (e.g., ease of understanding; EoU) can be highly time-consuming or subjective in clinical practice and research. Automatic speech recognition (ASR) may provide a more efficient, objective alternative, but its validity for intelligibility assessment in this population remains unexamined. This study evaluated the validity of ASR as a tool for intelligibility assessment in children with dysarthria. The most suitable ASR systems for approximating human intelligibility assessment were identified.

Methods: Five ASR systems transcribed speech samples produced by twenty children with dysarthria. Additionally, 168 adult listeners provided orthographic transcriptions and EoU ratings of the samples. Word recognition rate (WRR) was measured for both ASR and human listener transcriptions. Pearson correlations were used to assess the relationship between ASR-generated WRR and human WRR, as well as between ASR-generated WRR and human EoU ratings.

Results: 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. The four systems also exhibited strong correlations with EoU ratings, with Google Cloud ASR showing the strongest correlation. In contrast, Wav2Vec2 demonstrated a weak correlation with both human WRR and EoU ratings.

Conclusions: ASR shows promise as an adjunct tool for intelligibility assessment in children with dysarthria. If developed further, ASR could also be used for real-time feedback on intelligibility to help the children practice their speech skills independently. Of the ASR systems tested, WhisperX-medium appears most promising for approximating human transcription accuracy, whereas Google Cloud ASR is best suited for approximating perceptual ratings. However, differences in ASR performance highlight the need for careful system selection for appropriate clinical applications in this population.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to take this opportunity to express my deepest gratitude to the many individuals who have supported me throughout my doctoral journey.

First and foremost, I would like to express my deepest gratitude to my doctoral advisor, Dr. Erika S. Levy. She has been more than a mentor —she has been an academic mother, a role model, and an inspiring researcher. Her passion for research is truly remarkable, and it has shaped the way I approach my own work. When I decided on a doctoral program, my greatest consideration was the opportunity to work with Dr. Levy, and choosing Teachers College because of her was one of the hardest yet best decisions of my life. Over the years, I have learned so much from her about research, teaching, service, mentorship, and writing. Her unwavering support has extended far beyond academic guidance; she has written countless recommendation letters for me and encouraged me at every step of my doctoral journey. I will carry the lessons she has taught me throughout my professional career, striving to embody the dedication, generosity, and excellence she has demonstrated.

I would also like to express my sincere appreciation to Dr. Gemma Moya-Galé, an outstanding mentor, researcher, and friend. As a member of my dissertation committee, she has been incredibly generous with her time, always making herself available despite her busy schedule. Whenever I had questions, she not only listened but also provided thoughtful, insightful answers. She has supported me in numerous ways, including writing recommendation letters and helping me navigate my academic and professional journey. Her brilliance and kindness make her a role model I deeply admire.

I am profoundly grateful to Dr. Julia Hirschberg for serving as a member of my dissertation committee and for her invaluable contributions to my academic growth. Her Spoken Language Processing class was pivotal in shaping my research interests in automatic speech recognition, speech processing, and computational linguistics. She has always been available to answer my questions and has provided exceptional insights throughout my doctoral milestones—from my certification exam presentation to my dissertation proposal and defense. Her encouragement and support, including writing recommendation letters and sharing job opportunities, mean the world to me. It is truly an honor to have such an esteemed scholar on my committee.

I would also like to thank Dr. Andy Gordon and Dr. Jie Gao for serving as members of my dissertation committee and for their thoughtful feedback on my study. Their insights have been instrumental in refining my research.

A special thanks to Blake Vente and Dr. Eunjung Yeo for their expertise in computer science and computational linguistics. Their knowledge and insights helped me better understand automatic speech recognition (ASR) models and provided invaluable guidance with coding.

I would also like to extend my gratitude to Dr. KyungHae Hwang and Dr. Michelle YoungHwa Chang, former doctoral students in our lab, for their advice on participant recruitment, study design, and, most importantly, their moral support throughout this process. Their encouragement has meant so much to me. I am also grateful to Keren Tetteh, a doctoral student in our lab, for her warm hospitality in Ghana and for being an inspiration through her passion for the field as both a speech-language therapist and researcher.

A very special thank you to the research assistants who assisted with this study, as well as the children, their families, and the adult listeners who participated. This research would not

have been possible without their time and effort. Their contributions provided invaluable speech and listener data, which will help advance our understanding of speech intelligibility in children with cerebral palsy.

A special thank you to Dr. Cate Crowley for the incredible opportunity to participate in the international cleft lip and palate training in Ghana. This experience broadened my clinical perspective and deepened my appreciation for global outreach in speech-language pathology. I am truly grateful for the chance to learn from this training and from the professionals and families involved.

Additionally, I am grateful to the faculty and staff of the Communication Sciences and Disorders program and the Biobehavioral Sciences Department at Teachers College, including Dr. Michelle Troche and Dr. Lori Quinn, for their unwavering support of doctoral students. A special thank you to Eunji, Corey, Steven, and Maria for their administrative assistance throughout my time in the program.

I extend my heartfelt appreciation to my iTalk family for shaping me into a better speech-language pathologist. The memories of our laughter and camaraderie will always be cherished.

To my friends, thank you for your moral and emotional support. The time spent together has been a much-needed source of joy and a welcome break from academic life. I am especially grateful to those who participated in my pilot study, providing valuable data that helped refine the research design of my dissertation.

A very special thank you to my American mom, Dr. Lorraine Giordano, for her unwavering moral support, life lessons, laughter, and friendship over the past ten years in New York. Her kindness and wisdom have been a source of strength and joy throughout my journey.

Finally, I want to express my deepest love and gratitude to my family. Their unwavering support from South Korea, as well as my younger brother's presence here in New York, has been my anchor throughout this journey. Coming from a family of Ph.D. holders has provided me with invaluable guidance and encouragement. I truly could not have made it through this doctoral journey without them.

This dissertation is not just my achievement—it is a testament to the incredible people who have supported me along the way. I am deeply grateful to each and every person who has supported me throughout this journey.

Chapter 1: Introduction

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 more than 50% of children with CP and is associated with speech intelligibility impairment (Allison & Hustad, 2018; Hustad, 2008; Mei et al., 2020). Because reduced intelligibility 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). Given the pivotal role of intelligibility in determining successful communication, accurate measurement of the children's intelligibility as part of assessment and for determining any treatment-related gains is essential (Miller, 2013).

1.1 Intelligibility Assessment in Children with Dysarthria

Speech intelligibility in individuals with dysarthria is typically assessed through human perceptual evaluation (Allison, 2020; Hustad, 2008). Two primary human assessment measures often implemented are orthographic transcription accuracy and perceptual ratings. Human listeners' orthographic transcription accuracy, considered the gold standard and most objective measure of intelligibility (Hustad, 2006), requires listeners to indicate the words they believe they heard. Intelligibility is generally measured in terms of the percentage of words correctly transcribed (PWC; Hustad, 2008; Levy et al., 2017). Perceptual ratings 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). Human judgments of intelligibility, including ratings and percent estimates, are widely used in both clinical and research settings due to their ease of administration and minimal time requirements (Hustad, 2006; Miller, 2013). Such

judgments are also frequently employed in research to assess intelligibility in individuals with dysarthria across various age groups (Levy et al., 2017; 2021; Stipancic et al., 2016; Tjaden et al., 2014). 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, as well as analyzing their ratings and/or transcriptions (Allison, 2020; Jiao et al., 2015). Additionally, factors such as listeners' familiarity with the speaker, dysarthria, or test materials, as well as their age, hearing, and attitude can impact both transcription and ratings (Borrie et al., 2012; Miller, 2013). Furthermore, determining transcription accuracy can require testing of 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 research and clinical practice (Fox et al., 2021, Xu et al., 2014).

While perceptual ratings are more commonly used in clinical settings because of their efficiency, they have been criticized for being more subjective and less reliable than transcription accuracy (Kreiman & Gerratt, 1998, 2000; 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" (p. 603) for severity measurement. However, research has also revealed that perceptual ratings can show strong correlations with transcription accuracy, suggesting that ratings may serve as a

reasonable alternative in assessing dysarthric speech (Abur et al., 2019; Stipancic et al., 2016; Tjaden et al., 2014).

1.2 Automatic Speech Recognition in Dysarthric Speech

Given the limitations of traditional human-based assessment, there is growing interest in leveraging technology to streamline the assessment process, particularly through the use of automatic speech recognition (ASR) systems as an alternative for evaluating intelligibility in dysarthric speech. Two main categories of ASR systems are 1) speaker-independent ASR systems and 2) personalized ASR systems such as speaker-adaptive or speaker-dependent systems (Jaddoh et al., 2023; Rowe et al., 2022). Most SI systems are commercially available to the public. Such systems include platforms such as Siri, Amazon Alexa, and Google Home. In contrast, personalized ASR systems (e.g., Dragon or Cortana; Nuance, 2020; Microsoft, 2020) are designed for a specific user or population.

Various speaker-independent ASR systems have been evaluated for their transcription accuracy and potential clinical use, with the different systems' performance compared for speech from both adults with dysarthria and neurotypical speakers. Schultz and colleagues (2021) tested the performance of Amazon Web Services, Google Cloud, and IBM Watson ASR platforms, on speech produced by adults with neurodegenerative diseases including multiple sclerosis (MS) and Friedreich's ataxia (FA), along with speech produced by healthy controls. Among the ASR systems tested, 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. The study highlights the need for further improvements in ASR systems to recognize target words more accurately in dysarthric speech. Rodrigues et al. (2019) compared Google, Bing, and Nuance ASR systems in various noise conditions and distances from the neurotypical

speaker. Google consistently outperformed the other systems, demonstrating greater robustness to noise and age-related speech variations, suggesting Google as the most reliable ASR for neurotypical speech in diverse environments.

Currently, speaker-independent ASR systems demonstrate lower transcription accuracy for individuals with speech impairments than for neurotypical speakers (Gutz et al., 2022; Schultz et al., 2021). Nonetheless, ASR technology has shown potential for clinical application despite its limited accuracy in transcribing dysarthric speech. Schultz et al. (2021) highlighted that while ASR accuracy is lower for individuals with impaired speech, especially those with neurodegenerative conditions such as MS or FA, than for neurotypical adults, personalized ASR systems tailored to specific speech disorders could be trained to improve their transcriptions. Consistent with this, research has shown that personalized ASR systems show better performance than speaker-independent ASR systems, especially for recognizing severely dysarthric speech (Green et al., 2021). ASR provides immediate transcription of speech, aiding in assessment and in the continuous monitoring of intelligibility, which is critical for tracking the efficacy of dysarthria treatments (Gutz et al., 2023). If further developed, ASR-based tools could make speech intelligibility assessment more accessible and affordable. Additionally, ASR could facilitate personalized treatment plans by providing immediate and objective feedback on speech performance, helping clinicians tailor interventions to clients' individual needs (Tobin et al., 2024). However, Yılmaz et al. (2019) noted that while personalized ASR systems can enhance usability and transcription accuracy for individual users, their customization for a specific speaker may limit generalizability in clinical settings, where standardized assessments are required for diagnostic and therapeutic purposes. Thus, speaker-independent ASR systems remain essential for broader clinical applicability.

ASR performance can be evaluated based on two key metrics: transcription accuracy or its comparability to human perceptual measures. Transcription accuracy refers to the match between ASR output and the intended speech, providing information on the system's ability to recognize and correctly transcribe spoken words. Comparability, examined through correlation, assesses how closely ASR transcriptions align with human transcriptions or ratings, providing insight into ASR's potential for approximating human perceptual understanding. Both metrics are important when evaluating ASR performance. However, for clinical intelligibility assessment, alignment with human perceptual judgments is particularly important, as ASR systems are evaluated based on their ability to produce results similar to those of clinicians (e.g., SLPs) in assessing intelligibility. Comparability is typically assessed through correlation, measuring how well ASR outputs reflect human transcriptions or perceptual ratings. This metric is key to determining ASR's clinical validity as an intelligibility assessment tool (Goldsack et al., 2020; Gutz et al., 2022). A clinically valid ASR system would closely approximate human intelligibility judgments, ensuring it can detect intelligibility to a similar degree as clinicians would.

ASR systems have been examined in the assessment of speech intelligibility in adults with dysarthria, including adults with conditions such as Parkinson's disease (PD), cerebral palsy, aphasia, apraxia of speech, and amyotrophic lateral sclerosis (ALS; Gutz et al., 2022; Jacks et al., 2019; Moya-Galé et al., 2022; Tu et al., 2016). For example, Gutz and colleagues evaluated the clinical validity of Google Cloud ASR (Google LLC, 2020), for assessing speech severity in adults with ALS. Their findings showed strong correlations between ASR transcription accuracy and human listener transcription accuracy, supporting ASR's validity as a tool for assessing speech intelligibility. Similarly, prior research has demonstrated strong

correlations between ASR and human listeners' transcription accuracy in adults with dysarthria, including those with cerebral palsy or ALS (Doyle et al., 1997; Ferrier et al., 1995).

As described above, perceptual ratings may convey subjective listener experiences of intelligibility. Recent research has examined whether ASR can serve as a proxy for human ratings by generating intelligibility estimates that align with human perceptual judgments. Tu et al. (2016) examined ASR's potential for approximating human intelligibility assessment in adults with dysarthria, analyzing correlations between ASR transcription scores and human perceptual ratings across multiple speech domains, including overall speech severity, nasality, vocal quality, articulatory precision, and prosody. Results showed strong correlations between WER and human perceptual ratings (r = 0.68–0.84), suggesting that ASR can approximate human judgments of intelligibility across various dimensions. Although ASR and human listeners rely on vastly different mechanisms to interpret speech, these findings suggest that ASR systems can yield outcomes aligned with human perceptual judgments. This highlights ASR's potential as an efficient, objective tool for assessing intelligibility in a way that reflects listener experiences.

1.3 Automatic Speech Recognition in Children

Building on the advancements in ASR applications for evaluating dysarthric speech in adults, there is a growing interest in extending this technology to pediatric populations. However, children's speech presents distinct physiological, linguistic, and acoustic challenges that must be addressed to achieve accurate transcription and intelligibility assessment.

Most state-of-the-art ASR systems for healthy adults achieve high transcription accuracy, with low word error rate (WER)[1] of approximately 1.9% (i.e., greater than 98% accuracy; Gulati

[1] Word error rate (WER) is a gold standard measurement of performance in ASR systems. It is the sum of the number of word substitutions, insertions, and deletions divided by the total number of words in the target utterance (Moore et al., 2018; Tetzloff et al., 2024). The lower the WER is, the higher the ASR transcription accuracy is.

et al., 2020; Han et al., 2020). However, ASR performance for typically developing children is notably lower, with WER of approximately 10% (Liao et al., 2015). This discrepancy underscores the challenges of applying ASR systems trained on adult data to children's speech.

One key challenge in applying ASR to children's speech is the acoustic mismatch between children's speech and ASR training data. Wilpon and Jacobsen (1996) found that ASR systems performed less accurately for younger children, particularly those under 15 years old, possibly, in part, due to their higher fundamental and formant frequencies, which fall outside the ASR's optimal acoustic range, resulting in higher WER and lower transcription accuracy. Additionally, Yeung and Alwan (2018) highlighted that younger children's underdeveloped phonemic systems contribute to higher WER in ASR transcription, especially for words with greater syllable complexity. In particular, younger children often produce less accurate phonemic contrasts and less complex syllable structures, which challenge ASR systems designed primarily for adult speech patterns.

Of relevance, Rodrigues and colleagues (2019) found that, compared to two other speaker-independent ASR systems (i.e., Microsoft Bing, Nuance ASR), Google Cloud ASR performance was relatively unaffected by age, showing similar accuracy for the speech of neurotypical individuals across an age range of 8 to 51 years. Therefore, this system may be a useful tool for transcribing children's speech. While ASR transcription is less accurate for children's speech than for adults', it is important to consider that human listeners also face challenges transcribing younger children's speech, in part because of young children's phonemic imprecision. Thus, WER alone may not be the best benchmark for ASR performance in pediatric populations. Instead, comparability with human perceptual ratings may provide more clinically

meaningful insights into ASR's utility for assessing intelligibility, especially for children's speech.

Studies investigating ASR have extended beyond children with typical development (TD) to clinical populations, including children with speech sound impairment (Atkins et al., 2019; Benway & Preston, 2023, 2024), cleft lip and palate (Maier et al., 2006, 2007), hearing impairment (Lilley et al., 2014), language delays (Fox et al., 2021), and autism spectrum disorders (ASD; Gale et al., 2019). The adoption of ASR has been suggested as an adjunct to determine children's speech and language development and is motivated by several factors. First, as previously described, traditional transcription methods are labor-intensive and require significant time and effort (Atkins et al., 2019; Flipsen Jr., 2006; Fox et al., 2021; Xu et al., 2014). Such approaches have typically involved careful transcription of children's speech by individuals with specialized training (Atkins et al., 2019). Moreover, special challenges for the speech of children with language disorders include their greater number of lexical and grammatical errors than for the speech of children with TD, leading to additional transcription time. Miller and colleagues (2016) reported that transcribing language samples from children with TD required approximately four to six minutes for each minute of audio. In contrast, transcribing audio samples from children with language disorders required approximately seven to eight minutes per minute of audio. These challenges further complicate and delay the transcription and analysis process (Fox et al., 2021; Heilmann et al., 2010; Miller et al., 2016) and limit the scope and comprehensiveness of clinical assessment and research studies in children with speech and language disorders (Atkins et al., 2019). Therefore, if further developed for children's speech, ASR technology may offer a more efficient solution, enabling rapid transcription of speech data without the need for extensive human assessment. Additionally,

ASR technology, if further advanced, could enable comprehensive analysis of children's speech patterns and language characteristics (Atkins et al., 2019; Gale et al., 2019).

Given the need for objective and efficient intelligibility assessment tools, it is important to evaluate the validity of ASR in children with dysarthria to determine its potential clinical applications. Specifically, assessing ASR's alignment with human perceptual judgments can provide insight into its suitability for speech intelligibility assessment. While the previous research exploring ASR transcription accuracy for pediatric populations with various speech and language disorders has shed light on ASR's potential applications and limitations (Atkins et al., 2019; Fox et al., 2021; Gale et al., 2019; Xu et al., 2014), appropriateness of ASR for evaluating intelligibility in children with dysarthria has not been determined. ASR performance for children with dysarthria may differ from its performance for other populations because of the unique speech characteristics associated with dysarthria in this young population (Duffy, 2019), which may pose challenges for ASR systems.

Thus, despite the high prevalence of dysarthria among children with CP and the urgent need for timely and accurate intelligibility assessment, research on ASR's applicability for this purpose remains limited. A key question in addressing this gap is whether ASR systems can approximate traditional intelligibility measures assessed by humans.

1.4 Current Study

This study evaluated the validity of ASR technology for assessing the speech intelligibility of children with dysarthria due to CP. ASR transcription was compared to the (objective) measure of orthographic transcription accuracy (Hustad, 2006) and to commonly implemented perceptual ratings (Abur et al., 2019; Levy et al., 2017; 2021; Stipancic et al., 2016; Tjaden et al., 2014).

Our first aim was to determine whether ASR ouptut can be a valid representation of human listeners' transcription of the speech of children with dysarthria. To investigate this, speech samples from children with dysarthria were transcribed by five ASR systems and 168 human listeners. ASR transcription accuracy was compared to human transcription accuracy to assess the degree of alignment. It was hypothesized that ASR transcription accuracy would show moderate to strong correlations with human transcription accuracy (Gutz et al., 2022; Jacks et al., 2019), supporting its validity for intelligibility assessment. 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 Google Cloud would show the strongest correlation with human listeners, as previous research has demonstrated its high alignment with human transcription accuracy (Gutz et al., 2022), suggesting its potential as the most appropriate ASR system for approximating human transcription in this population.

Our second aim was to determine whether ASR can provide a valid representation of human perceptual ratings of intelligibility, specifically, EoU ratings (Fletcher et al., 2017; Landa et al., 2014; Levy et al., 2017; 2021). It was hypothesized that ASR transcription accuracy would strongly correlate with EoU ratings**,** suggesting that ASR could serve as a valid proxy for estimating intelligibility from a listener's perspective (Tu et al., 2016). Additionally, this study examined the performance of five ASR systems and identified those that may be best proxies for human EoU ratings for the speech of children with dysarthria.

While perceptual ratings currently offer a far more time-efficient means of assessing intelligibility than the more objective measure of human transcription accuracy, ASR would provide additional advantages, if found to approximate human judgments. If developed further,

ASR advantages could include provision of immediate and objective feedback, potentially without the consistent need for human transcribers or raters. This capability would ultimately be particularly beneficial for children with dysarthria, who may have mobility constraints, as it would allow them to practice their speech skills independently, receive immediate feedback on their intelligibility, and track their speech progress in real-time.

Chapter 2: Methods

This study was approved by the Institutional Review Board at Teachers College, Columbia University, New York City.

2.1 Participants

2.1.1 Speakers – Children with Dysarthria

A total of twenty children (7 girls, 13 boys) with dysarthria due to CP were included in this study. (Data from seventeen of the children were reported in Levy et al.'s [2017] speech cueing study. Data from children with TD are not included in the present study examining the validity of ASR for children with dysarthria[2].) All of the children 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 observable characteristics, including abnormal orofacial and respiratory movements, and 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 (Allison & Hustad, 2018; Levy et al., 2021). 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

[2] Speech data were also collected from children with TD, but are not reported in the present study, which aimed to determine the validity of ASR in approximating human judgments of intelligibility of the speech of children with dysarthria. A comparison of ASR performance for speech produced by children with dysarthria and their age- and sex-matched TD peers was reported in Choi et al. (2025). Findings revealed that twelve out of the thirteen ASR systems tested showed significant differences in WER between the two groups of children ($p < .001$).

to follow simple directions, and English as the dominant language. Participant characteristics, including Gross Motor Function Classification System (GMFCS) 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). (For one child, CP06, no language test recording was available.) However, due to the limitations of these tests for children with CP, the scores likely underestimated the children's abilities (Hustad et al., 2012; Levy et al., 2017; 2021).

Table 1: Demographic and language characteristics of children with dysarthria due to cerebral palsy.

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-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-moderate	TACL	<1st	very poor
CP05	7;1	M	Spastic triplegia	IV	Moderate-severe	TACL	9th	below average
CP06	7;7	F	Spastic hemiplegia	IV	Mild-moderate	n/a	n/a	average*
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	I	Moderate	TACL	<1st	very poor

			diplegia					
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-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-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 = Gross Motor Function Classification System (Palisano et al., 1997); GMFCS rating: I = no/mild impairment, V=severe impairment; F = female; M = male. * 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 SLPs to be age appropriate.

2.1.2 Listeners – Human Listeners

A total of 168 American-English-speaking adult listeners (86 female, 82 male), ages 18 to 40 years (M = 29.2, SD = 6.12) participated in a listening task via crowdsourcing. Exclusion criteria were a history of speech, language, or hearing disorders, long-term health conditions or disabilities.

2.2 ASR Systems

Five speaker-independent ASR systems, including both open-source and commercial systems trained on diverse speech datasets, were evaluated for assessing speech intelligibility in

children with dysarthria. The open-source systems were variants of WhisperX, a transformer-based model fine-tuned for automatic speech transcription (Bain et al., 2023). Wav2Vec2 is a self-supervised ASR system designed to learn speech representations without direct phonetic supervision (Baevski et al., 2020). The commercial system was Google Cloud ASR, a proprietary system optimized for real-world transcription tasks (Google LLC, 2020). WhisperX has shown strong performance in WER and disfluency handling, suggesting that it could outperform other systems for disfluent speech[3] (Graham & Roll, 2024; Teleki et al., 2024). Three different versions of WhisperX were assessed (i.e., WhisperX-small, WhisperX-medium, Whisper X-large). The systems have been evaluated for accuracy performance and computational cost across different model sizes (Ma et al., 2024). Google Cloud ASR has been used to assess speech intelligibility and severity in individuals with ALS, demonstrating strong correlations with human transcriptions (Gutz et al., 2022). Similarly, Wav2Vec2 has been applied to assess speech in individuals with primary progressive apraxia of speech (PPAOS), also showing strong correlations between AOS severity and WER (Tetzloff et al., 2024).

2.3 Speech Samples

2.3.1 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 (e.g., *three little pink pigs*; TOCS+; Hodge et al., 2007) were selected. These stimuli are developmentally appropriate for young children in terms of lexical, phonetic, syntactic, and morphological complexity. Additionally, three sentences ranging from four to seven words in

[3]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).

length (i.e., *Buy Bobby a puppy, The blue spot is on the key, The potato stew is in the pot*; Weismer, 1984) were selected for their range of vowels and consonants. 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).

2.3.2 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 (Fox & Boliek, 2012; Levy et al., 2020; Miller, 2020; Švec & Granqvist, 2018), a calibration procedure was performed to maximize accuracy of sound pressure level (SPL) measurement. During calibration, the experimenter audio-recorded a pure tone produced by a tuner (OT120-Korg Orchestral) positioned 8 cm away from a sound level meter (Galaxy CheckMate CM140) and noted the tone's SPL. During 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 SoundForge 17.0 software on a Dell OptiPlex 7090 computer, connected via a Scarlett 2i2 audio interface (Scarlett2i2, Focusrite 2x2 USB2). The input dial setting was kept constant throughout the study. The children's speech was recorded at a sampling rate of 22,050 Hz with 16-bit resolution in mono.

2.4 Tasks

2.4.1 Children's Speech Task

During the speech recording session, the children were asked to repeat target phrases or sentences produced by a model speaker. For each utterance, they were presented with the target phrase or sentence through loudspeakers (Altec Lansing ADA 215) placed at a consistent

distance of 30cm from them (Hall et al., 1968). The model speaker for the recordings was a female adult native speaker of American English from the New York City area.

2.4.2 Adults' Listening Task – Human Orthographic Transcription and ease of understanding (EoU) ratings

In order for listeners to evaluate the children's speech recordings, a total of 586 naïve listeners were recruited via crowdsourcing (Prolific Academic; 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., 2024). A total of 168 listeners completed the task. 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 (Hwang, 2023), 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 inches) from the left loudspeaker of their computer.

The following listener calibration process was designed to replicate the original SPL of the child's speech (Hwang, 2023; 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 would be presented with the tone and be 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 additional pure-tone sound files using the same app and to type the most stable dB SPL reading into the testing interface. Progression to the next step was restricted until the entered dB SPL values fell within a pre-determined acceptable range of +/-1dB (Švec & Granqvist, 2018), optimizing accurate calibration before beginning the listening session.

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 what they heard. 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 visual analog scale, with anchor points from "difficult" to "easy" (Levy et al., 2017; 2021). For interrater reliability assessment, 20% of the sentences were randomly selected and replayed to listeners during the task (Allison & Hustad, 2018; Levy et al., 2017). Each child's speech was transcribed and rated by eight to nine listeners (McAllister Byun et al., 2015).

2.4.3 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.

2.5 Analysis

2.5.1 Human Transcription

Human transcriptions were analyzed in terms of the percentage of words correct (PWC). The PWC was calculated by the number of correctly transcribed words divided by the total number of words in each target sentence. Response words were counted as correct if they match the target exactly, were homonyms, or were clear misspellings of the target or of its homonym (Hustad, 2008; Levy et al., 2017).

2.5.2 ASR Transcription

ASR transcriptions were assessed using the word error rate (WER), which is a gold standard metric for evaluating speech recognition systems (Gutz et al., 2022, 2023; Tetzloff et al., 2024). As described above, 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).

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

2.5.3 Comparison between Two Modalities

For the purpose of generating a single metric indicating percentage of words correctly transcribed for comparing human transcriptions with ASR, WER was converted using 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.

2.5.4 Statistical Analysis

All statistical analyses were completed in R (R Posit team, 2023). To address our first research aim, the clinical validity of ASR was assessed by comparing ASR transcription accuracy to human listeners' transcription accuracy for the speech produced by the children with dysarthria. Pearson correlation coefficients were calculated to determine the agreement between ASR-generated WRR and human WRR (Gutz et al., 2022, Tu et al., 2016).

To evaluate each ASR system's performance, the transcription accuracy of the five speaker-independent ASR systems (WhisperX-small, WhisperX-medium, WhisperX-large, Google Cloud, Wav2Vec2) was compared with human WRR. Pearson correlation coefficients were calculated between each ASR system's WRR and human WRR, with stronger correlations suggesting greater alignment with human transcription accuracy.

For our second research aim, of assessing ASR for its capacity to represent human perceptual ratings, the relationship between ASR transcription accuracy and perceptual ratings was examined by means of Pearson correlation coefficients. Correlations between ASR-generated WRR and human EoU ratings were analyzed to determine whether ASR accuracy would be a valid representation of human perceptual ratings. A strong, positive correlation would indicate a valid proxy for EoU ratings in assessing intelligibility in children with dysarthria.

ASR-generated WRR in the five ASR systems was compared with human EoU ratings. Pearson correlation coefficients were calculated between each ASR system's WRR and human EoU ratings. A strong, positive correlation (i.e., $r \geq 0.70$) would indicate that ASR can

approximate perceptual ratings, supporting its validity for intelligibility assessment in children with dysarthria.

In an assessment of intrarater and interrater reliability, intraclass correlation coefficients (ICCs) were computed for human orthographic transcriptions 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 > 0.90 = excellent, 0.75–0.90 = good, 0.50–0.75 = moderate, and < 0.50 = poor. A single-rater, two-way model ICC (2,1) was used to measure for intrarater reliability, and a multi-rater, two-way model ICC (2,k) was used for interrater reliability, accounting for consistency in transcriptions and ratings.

Chapter 3: Results

Table 2 presents descriptive data on word recognition rate (WRR) for the speech of children with dysarthria, transcribed by human listeners and five ASR systems. WhisperX-large and WhisperX-medium achieved the highest 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%. The ASR performance showed considerable variability, with two systems (WhisperX-medium and WisperX-large) even exceeding human transcription accuracy for this population.

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		DYS Mean (SD)
Human	Human listeners	38.95 (40.43)
ASR System	Google Cloud	30.23 (37.43)
	Wav2Vec2-960h-Large	5.63 (16.17)
	WhisperX-small	38.90 (40.86)
	WhisperX-medium	46.04 (41.78)
	WhisperX-large	48.24 (41.82)

Note. DYS = children with dysarthria

Addressing the first research aim, to determine whether ASR can provide a valid representation of human listeners' transcription, Pearson correlations indicated that ASR-generated WRR was associated with average human WRR in the five speaker-independent ASR systems: WhisperX-small, WhisperX-medium, WhisperX-large, Google Cloud, and Wav2Vec2-960h-large. Significant linear correlations were found between all ASR systems and human listeners, as shown in Figure 1. WhisperX-medium demonstrated the strongest correlation with human transcription ($r = 0.83$, $p < .001$). Similarly, WhisperX-large and WhisperX-small demonstrated strong correlations, for WhisperX-large ($r = 0.82$, $p < .001$) and for WhisperX-

small ($r = 0.81$, $p < .001$). Google Cloud also showed strong alignment with human WRR ($r = 0.79$, $p < .001$). In contrast, Wav2Vec2 exhibited a weak correlation ($r = 0.36$, $p < .001$).

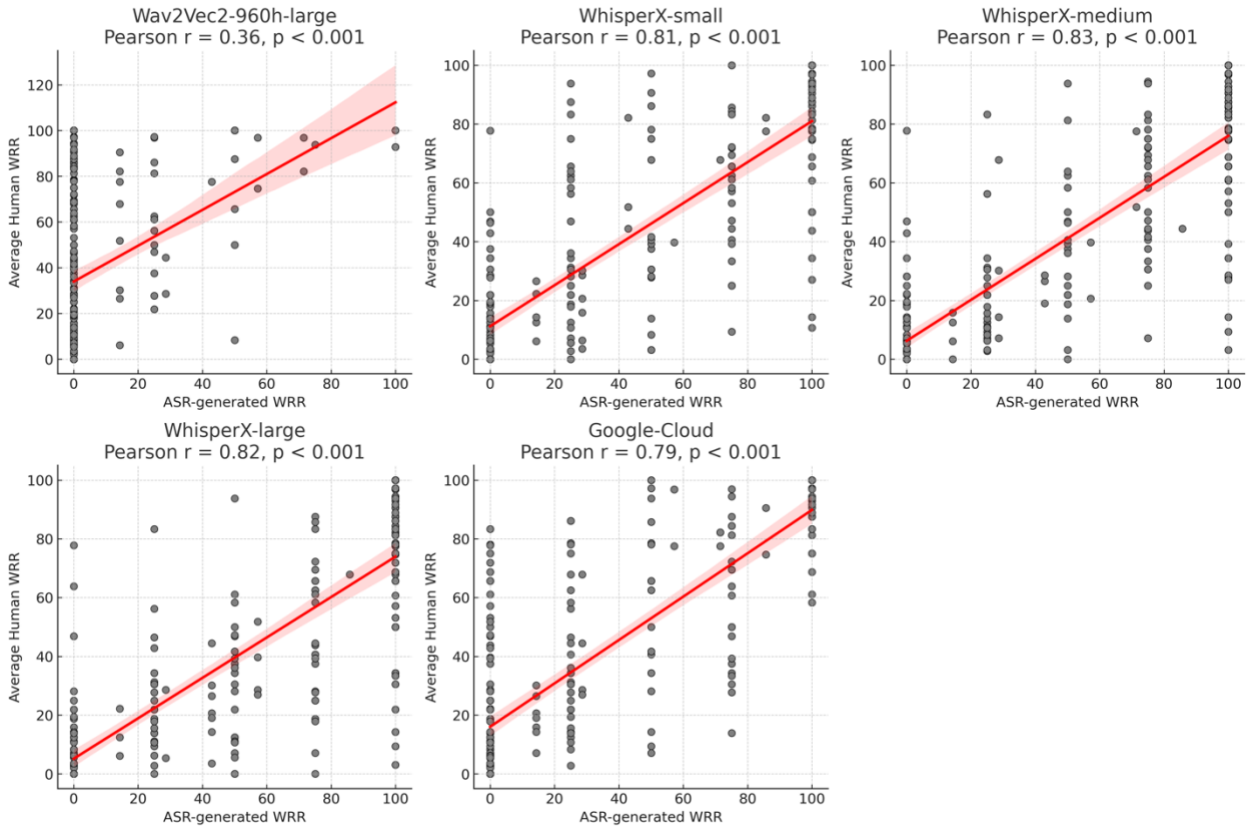

Figure 1. Scatterplots showing the relationship between ASR-generated word recognition rate (WRR) and average human WRR across five automatic speech recognition (ASR) systems.

To further evaluate the validity of ASR, addressing the second research aim, to determine whether ASR output is a valid proxy for human perceptual ratings, Pearson correlations were calculated between ASR transcription accuracy and average human EoU ratings in the five ASR systems. As shown in Figure **2**, ASR-generated WRR was correlated with average human EoU ratings across five ASR systems. Significant positive correlations were found for all ASR systems, suggesting that higher ASR transcription accuracy was associated with higher listener-rated intelligibility. Google Cloud demonstrated the strongest correlation with EoU ratings ($r =$

0.76, p < .001), indicating a valid proxy for human perceptual ratings. The WhisperX systems also showed strong correlations (r = 0.69, p < .001). Wav2Vec2 showed a weak correlation with human EoU ratings (r = 0.32, p < .001).

An intraclass correlation coefficient (ICC) was computed to assess the intrarater and interrater reliability of human transcription accuracy and perceptual ratings across repeated trials during the listener testing. Intrarater reliability for WRR was excellent, as indicated by ICC(2,1) = 0.945, p < .001. Interrater reliability for WRR was good, with ICC(2,k) = 0.865, p < .001. Similarly, intrarater reliability for EoU ratings was good, as suggested by ICC(2,1) = 0.854, p < .001. Interrater reliability for EoU ratings was also good, as indicated by ICC(2,k) = 0.803, p < .001.

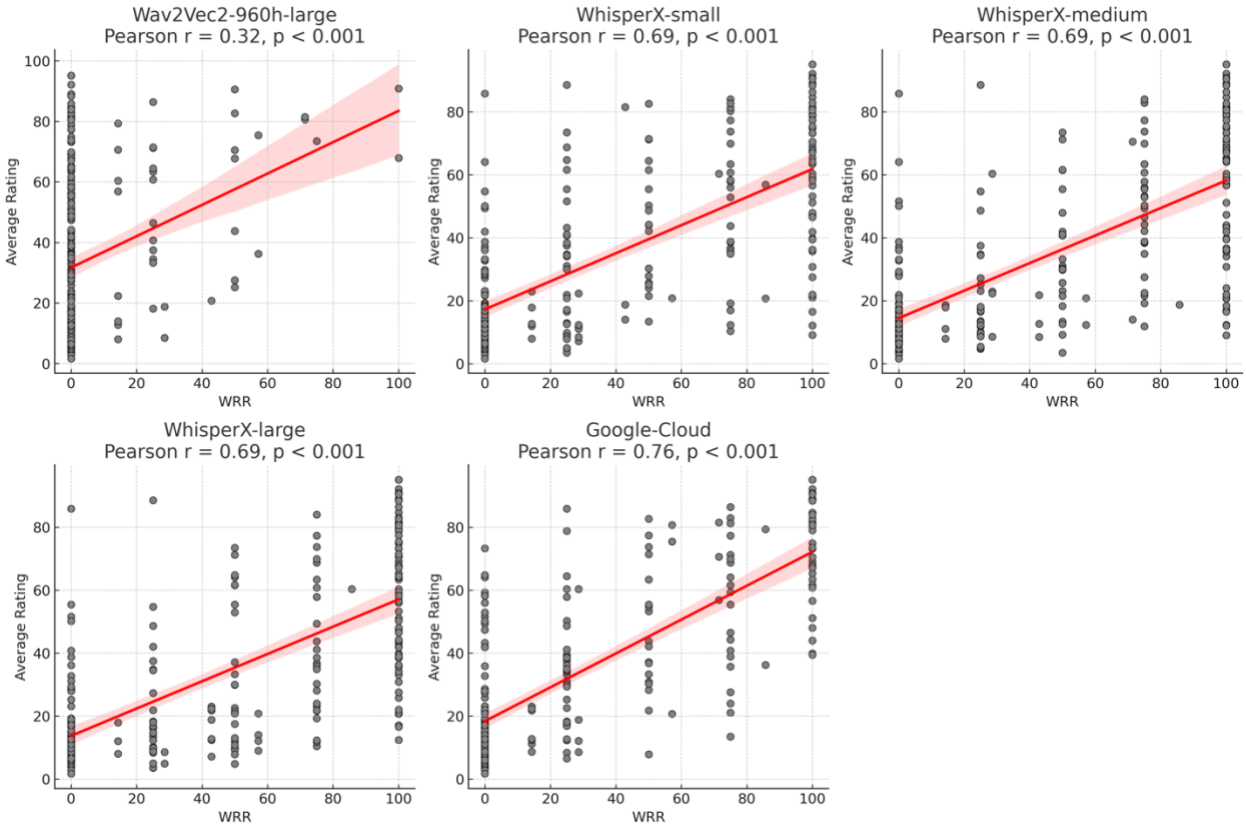

Figure 2. Scatterplots demonstrating the relationship between automatic speech recognition (ASR)-generated word recognition rate (WRR) and average human ease of understanding (EoU) ratings for five ASR systems.

Chapter 4: Discussion

This study examined the validity of automatic speech recognition (ASR) systems for assessing intelligibility of speech produced by children with dysarthria secondary to CP. Overall, ASR-generated WRR correlated with human WRR and EoU ratings. The findings provide insight into the extent to which ASR systems can approximate human listeners' transcriptions and perceptual ratings, with implications for clinical and technological applications.

4.1 ASR Transcription and Human Transcription

Four of the five ASR systems tested demonstrated strong correlations with human WRR, suggesting their potential clinical utility for assessing speech intelligibility in children with dysarthria. In contrast, Wav2Vec2 exhibited a weaker correlation, indicating challenges in transcribing dysarthric speech and the need for further refinement.

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 strong correlations found for most of these ASR systems' WRR with human WRR highlight ASR's promise for supplementing traditional human-based measures, potentially reducing the time and resources needed for assessing the children's speech.

4.2 ASR Transcription and Human EoU Ratings

ASR-generated WRR showed varying degrees of correlation with human EoU ratings, with some systems aligning more closely with perceptual judgments than others. Google Cloud exhibited the strongest correlation, suggesting its potential as a valid tool for intelligibility assessment in children with dysarthria. The three WhisperX systems also demonstrated strong correlations, albeit with slightly weaker alignment than Google Cloud, suggesting these systems

may benefit from further refinement. Wav2Vec2 showed the weakest correlation, again revealing its limitations in assessing intelligibility in alignment with human listeners. These results indicate that overall, ASR transcription accuracy is associated with human perceptual ratings of intelligibility, although the degree of alignment varies among ASR systems.

4.3 Clinical and Technological Implications

Findings suggest that WhisperX systems and Google Cloud ASR may be suitable for assessing speech intelligibility in children with dysarthria in clinical and research settings. The strong correlation between ASR transcription accuracy and human perceptual ratings supports ASR-generated WRR as a valid indicator of intelligibility. This technology is at least as objective as listener transcription, more objective than listener ratings, and significantly more time- and resource-efficient than perceptual assessments, especially those involving human listeners' transcription and its analysis. Fortunately, these speaker-independent ASR systems can be implemented with minimal technical expertise, making them accessible to SLPs and researchers.

Miller (2013) suggests that focus on economy of time has led to the implementation of unreliable, but rapid, human perceptual measures. Contrary to the goal of efficiency, use of such measures may result in prolonged therapy and reduced client and clinician morale due to poor understanding of salient treatment targets and inadequate documentation of intervention progress. Our findings indicate that ASR systems show promise as assistive tools for objective and cost-efficient intelligibility assessment. ASR transcription can streamline clinical assessment by reducing the burden of manual transcription, allowing clinicians to allocate more time to interpretation and intervention. Beyond its speed and objectivity, ASR could offer a unique advantage by providing immediate intelligibility feedback without consistently requiring human

raters. Recent studies have demonstrated the feasibility of ASR for real-time intelligibility feedback in adults with dysarthria (Gutz et al., 2023; Tobin et al., 2024). The present study extends this research by evaluating ASR performance in children with dysarthria, highlighting its potential not only for assessment but also, if developed further, for real-time feedback. Such ASR feedback could empower children with dysarthria to engage in independent practice and self-monitoring, reducing reliance on in-person clinical visits and enabling more consistent speech practice at home.

From an ASR evaluation standpoint, the findings highlight the importance of system selection and adaptation when using ASR for assessing intelligibility in children with dysarthria. Different ASR systems vary in their ability to transcribe speech accurately and to align with human perceptions of intelligibility. While some systems, such as WhisperX-medium and WhisperX-large, actually achieved higher transcription accuracy than naïve human listeners (see Table 2), this does not necessarily indicate stronger alignment with human perceptual ratings. These systems may be particularly useful for applications requiring precise documentation of speech production, such as tracking articulation and phonological patterns or generating speech-to-text output for augmentative and alternative communication (AAC) devices (Chen et al., 2024; Hui et al., 2024). Notably, WhisperX-medium demonstrated the strongest correlation with human transcription accuracy among the ASR systems tested, suggesting it may serve as a promising proxy for clinician transcriptions in contexts where human annotations are impractical or unavailable. However, despite its lower transcription accuracy (30.23%), Google Cloud demonstrated the strongest correlation with human perceptual ratings, suggesting that it may be better suited for evaluating intelligibility and tracking speech progress. This alignment with

human perception may stem from ASR's ability to recognize phonetic and linguistic patterns similar to those used by human listeners when decoding dysarthric speech (Gutz et al., 2022).

Google Cloud's closest alignment with perceptual ratings may be attributed to its proprietary deep learning architecture, which is trained on extensive, diverse datasets, including noisy and disordered speech. In contrast to open-source systems, Google Cloud benefits from continuous optimization and adaptation to real-world linguistic variability, which may allow it to recognize more phonetic and prosodic cues relevant to intelligibility in dysarthric speech (De Russis & Corno, 2019). Prior research has shown that Google Cloud ASR is robust in handling speaker variability and environmental noise, compared to other ASR systems. Rodrigues et al. (2019) demonstrated that Google ASR outperformed Bing and Nuance in recognizing speech across different speaker characteristics of neurotypical individuals and noise conditions, which may explain its stronger association with human EoU ratings in the current study, despite its lower WRR. Moreover, Google Cloud ASR has demonstrated superior performance in recognizing dysarthric speech compared to other ASR systems. Schultz et al. (2021) found that Google Cloud and Amazon Web Services outperformed IBM Watson in transcribing speech from individuals with neurodegenerative disorders, although accuracy declined with increased disease severity. Additionally, Gutz et al. (2022) found that Google Cloud ASR aligned closely with human judgments, further supporting its appropriateness for assessing intelligibility.

Variation in ASR performance underscores the importance of assessing ASR systems beyond transcription accuracy alone, particularly when considering clinical applications. Teleki et al. (2024) found that WhisperX outperformed Google ASR in transcribing interjection-type disfluencies (e.g., "uh," "um"), particularly in non-scripted speech, whereas Google ASR better captured edited disfluencies, such as sentence repairs and revisions, suggesting differential

sensitivity to disfluency types across systems. WhisperX, whose WRR aligned most closely with human transcription accuracy, may excel at recognizing individual words accurately, but may be less capable of interpreting atypical speech in a way that aligns with human decoding of speech.

Conversely, Wav2Vec2 did not align well with human perceptual judgments of the speech of children with dysarthria in this study, possibly due to its self-supervised learning approach. Unlike supervised ASR systems that are trained explicitly with labeled speech-to-text data, Wav2Vec2 learns latent speech representations without direct exposure to phonetic or linguistic rules. While this approach has been applied for recognizing neurotypical speech, it faces challenges with dysarthric speech, which contains significant deviations from typical phonetic patterns. The weakness of Wav2Vec2 in this context suggests that unlike other open-source ASR systems, self-supervised ASR systems such as Wav2Vec2 require targeted fine-tuning on dysarthric speech data to improve alignment with human intelligibility judgments. A recent study found a strong negative correlation between phoneme error rate from fine-tuned Wav2Vec2-based systems and human intelligibility scores, indicating that with appropriate fine-tuning, Wav2Vec2 has the potential to better assess intelligibility in dysarthric speech (Yang et al., 2022).

Improving ASR systems would increase accessibility and customization for clinical applications, including for children with dysarthria. Future research should focus on domain-specific training strategies to enhance ASR performance for this population. Furthermore, investigating how different ASR architectures handle phoneme-level distortions and prosodic variations supports efforts in refining their clinical applicability. These variations 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.

4.4 Limitations and Future Work

Despite the promising findings on ASR for intelligibility assessment of children with dysarthria in this study, several limitations must be acknowledged. First, the dataset was limited to the speech produced by children with this motor speech disorder; thus, the results may not generalize to other populations or speech disorders. Additionally, only five ASR systems were evaluated, and other systems or additional fine-tuning may yield different results. Future research should 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.

An additional 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 (Hustad et al., 2021; Levy et al., 2017), studies on adult populations (e.g., those with amyotrophic lateral sclerosis) suggest that speech severity and utterance length can interact to influence intelligibility (Allison et al., 2019, Gutz et al., 2019). Future work should determine whether utterance length affects ASR performance and listener judgments of children's speech.

Another limitation is the reliance on WRR as a primary metric for ASR performance. Although, as previously noted, WRR provides insight into transcription accuracy, it does not capture phoneme-level errors or other distortions that may impact intelligibility (Miller, 2013). Future work should 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. To improve clinical applicability, future studies should also investigate the real-time usability of ASR systems for providing feedback during speech treatment. Moreover, future work should include testing speech-language pathologists as listeners and comparing their judgments with ASR performance to further evaluate the clinical utility and reliability of ASR systems. Additionally, exploring the integration of ASR into assistive communication devices may enhance accessibility for individuals with severe speech impairments. A long-term goal should be the development of ASR systems specifically trained on the speech of children with dysarthria to optimize ASR accuracy and clinical utility.

Chapter 5: Conclusions

This study provides evidence that ASR transcription accuracy yielded by various speaker-independent ASR systems aligns with both human transcription accuracy and ease-of-understanding ratings for the speech of children with dysarthria. WhisperX systems and Google Cloud exhibited validity as proxies for human transcription. Additionally, these systems may provide valid representations of human perceptual ratings, with Google Cloud most closely aligning with human judgments of intelligibility in children with dysarthria. Certainly, ASR system selection should align with clinical goals, whether for intelligibility assessment, intervention planning, or assistive communication. As the first examination of ASR for intelligibility assessment in children with dysarthria, this study highlights the need for further refinements in ASR. Advancing ASR technology could lead to more effective tools for assessment and intervention, ultimately improving communication outcomes for this population.

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., & 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. (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., 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.

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 (ICPhS 2019),* Melbourne, Australia, 4–10.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Bain, M., Huh, J., Han, T., Zisserman, A. (2023) WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *Interspeech 2023*, 4489-4493, doi: 10.21437/Interspeech.2023-78

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Bolker, M. B. (2015). Package 'lme4'. *convergence*, *12*(1), 2.

Benway, N. R., & Preston, J. L. (2023). Prospective validation of motor-based intervention with automated mispronunciation detection of rhotics in residual speech sound disorders. *Interspeech 2023*, 4558–4562. https://doi.org/10.21437/Interspeech.2023-1882

Benway, N. R., & Preston, J. L. (2024). Artificial Intelligence–Assisted Speech Therapy for /ɹ/: A Single-Case Experimental Study. *American Journal of Speech-Language Pathology*, *33*(5), 2461–2486. https://doi.org/10.1044/2024_AJSLP-23-00448

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*. Austin, TX: Pro-Ed.

Centers for Disease Control and Prevention. (2020, December). *Data and Statistics for Cerebral Palsy | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/cp/data.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

Choi, J., Vente, B., Yeo, E., Hwang, K., Hirschberg, J., & Levy, E. S. (2025). Comparing automatic speech recognition models for assessing intelligibility in children with dysarthria. Manuscript submitted for publication.

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

Doyle, P. C., Leeper, H. A., Kotler, A. L., Thomas-Stonell, N., O'Neill, C., Dylke, M.C., & Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development*, 34(3), 309-316.

Duffy, J. R. (2019). *Motor speech disorders e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences.

Ferrier, L., Shane, H., Ballard, H., Carpenter, T., & Benoit, A. (1995) Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11, 165-175.

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 : JSLHR*, *60*(11), 3043–3057. https://doi.org/10.1044/2016_JSLHR-S-16-0218

Flipsen Jr., P. (2006). Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics*, *20*(4), 303–312. https://doi.org/10.1080/02699200400024863

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)

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

Fudala, J. B. (2001). *Arizona Articulation Proficiency Scale - Third Edition*. Los Angeles, CA: Western Psychological Services.

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.

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

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

Google LLC. (2020). Speech-to-text: Automatic speech recognition, 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), 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. *Interspeech 2021*, 4778–4782. https://doi.org/10.21437/Interspeech.2021-1384

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition (arXiv:2005.08100). arXiv. http://arxiv.org/abs/2005.08100

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

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

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

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Gulati, A., Pang, R., & Wu, Y. (2020). ContextNet: Improving convolutional neural networks for automatic speech recognition with global context (arXiv:2005.03191). arXiv. http://arxiv.org/abs/2005.03191

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech, and Hearing Services in Schools*, *41*(4), 393–404. https://doi.org/10.1044/0161-1461(2009/09-0023)

Hodge, M., Daniels, J., & Gotzke, C. L. (2007). TOCS+ intelligibility measures. *Edmonton, AB: 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:2411.00980). 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., Mahr, T. J., Natzke, P., & Rathouz, P. J. (2021). Speech development between 30 and 119 months in typical children I: Intelligibility growth curves for single-word and multiword productions. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3707–3719. https://doi.org/10.1044/2021_JSLHR-21-00142

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, Kyung Hae. (2023). Validity of speech measures collected online from children with dysarthria. [Doctoral dissertation, Columbia University]. Columbia University Libraries.

Jaddoh, A., Loizides, F., & Rana, O. (2023). Interaction between people with dysarthria and speech recognition systems: A review. *Assistive Technology*, *35*(4), 330–338. https://doi.org/10.1080/10400435.2022.2061085

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 Acoustical Society of America, 135,* 186 – 192. https://doi: 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.

Kreiman, J., & Gerratt, B. R. (2000). Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America*, *108*(4), 1867-1876.

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. *The Lancet's EClinicalMedicine, 24,* 100429. https://doi.org/10.1016/j.eclinm.2020.100429

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q.-M., Sainath, T. N., Senior, A., Beaufays, F., & Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. *Interspeech 2015*, 1611–1615. https://doi.org/10.21437/Interspeech.2015-373

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

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. *18th International Conference on Pattern Recognition (ICPR'06)*, 274–277. https://doi.org/10.1109/ICPR.2006.718

Maier, A., Haderlein, T., Schuster, M., Nkenke, E., & Nöth, E. (2007). Intelligibility Is more than a single word: Quantification of speech intelligibility by ASR and prosody. In V. Matoušek & P. Mautner (Eds.), *Text, Speech and Dialogue* (Vol. 4629, pp. 278–285). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74628-7_37

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 and child neurology*, *62*(12), 1374–1382. https://doi.org/10.1111/dmcn.14592

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

Miller, J. F., Andriacchi, K., & Nockerts, A. (2016). Using language sample analysis to assess spoken language production in adolescents. *Language, Speech, and Hearing Services in Schools*, *47*(2), 99–112. https://doi.org/10.1044/2015_LSHSS-15-0051

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., Venkateswara, H., and Panchanathan, S. (2018). Whistle blowing ASRs: evaluating the need for more inclusive speech recognition systems. *Interspeech 2018*, 466–470. doi: 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), e40567. https://doi.org/10.2196/40567

Murphy, K. R., & Davidshofer, C. O. (1998). Psychological testing: Principles and applications (2nd ed.). Prentice Hall.

Nuance. (2020). Dragon Naturally Speaking software. https://www. nuance.com/en-au/dragon/support/dragon-naturallyspeaking.html.

Palan, S., & Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22-27.

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,* 214–223.

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

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. *Proceedings of the 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*, 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

Southwell, R., Pugh, S., E. Margaret Perkoff, 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. *International Educational Data Mining Society*. https://doi.org/10.5281/ZENODO.6853109

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

Stipancic, K. L., Brenk, F., Qiu, M., & Tjaden, K. (2024). Progress toward estimating the minimal clinically important difference of intelligibility: A crowdsourced perceptual experiment. *Journal of Speech, Language, and Hearing Research*, 1–15. https://doi.org/10.1044/2024_JSLHR-24-00354

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. *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). San Diego, CA: 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. *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. *Interspeech 2018, 1661-1665,* doi: 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.

Zimmerman, I., Steiner, V., & Pond, R. (2012). *Preschool Language Scales–Fifth Edition Screening Test.* The Psychological Corporation.