Interaction between people with dysarthria and speech recognition systems: A review

Aisha Jaddoh, Fernando Loizides, Omer Rana

jaddoha@cardiff.ac.uk, loizidesf@cardiff.ac.uk, ranaof@cardiff.ac.uk

In recent years, rapid advancements have taken place for automatic speech recognition (ASR) systems and devices. Though ASR technologies have increased, the accessibility of these novel interaction systems is underreported and may present difficulties for people with speech impediments. In this article, we attempt to identify gaps in current research on the interaction between people with dysarthria and ASR systems and devices. We cover the period from 2011, when Siri (the first and the leading commercial voice assistant) was launched, to 2020. The review employs an interaction framework in which each element (user, input, system, and output) contributes to the interaction process. To select the articles for review, we conducted a search of scientific databases and academic journals. A total of 36 studies met the inclusion criteria, which included use of the word error rate (WER) as a measurement for evaluating ASR systems. This review determines that challenges in interacting with ASR systems persist even in light of the most recent commercial technologies. Further, understanding of the entire interaction process remains limited; thus, to improve this interaction, the recent progress of ASR systems must be elucidated.

Keywords: automatic speech recognition, dysarthria, interaction, virtual home assistants

Introduction

Automatic speech recognition (ASR) is now a part of daily life. ASR technologies may be built into phones (e.g., Apple's Siri), computers (e.g., Microsoft Cortana), vehicles (e.g., Apple CarPlay), or standalone devices such as virtual home assistants (e.g., Amazon Echo and Google Home). ASR is a hands-free system with the potential to assist people with disabilities who may be unable to use other input methods, such as computer mice, keyboards.

However, people with speech impairments—specifically, those with moderate to severe dysarthria, a neurological motor speech disorder that results in speech difficulties, poor speech intelligibility, difficulty coordinating breaths and speech and inconsistency in the production of the same word over multiple trials —may have difficulty using ASR technologies. Dysarthria could be caused by different diseases, for example, multiple sclerosis or motor neuron disease, or occur after strokes and brain injuries. Research on the interactions between people with dysarthria and speech technologies has been recognized since 1985 (Fried-Oken, 1985), with several studies examining speech intelligibility's contribution to ASR performance. Ferrier et al. (1995) examined the correlation between speaker intelligibility and recognition success, finding that intelligibility has a strong correlation with ASR performance. Thomas-Stonell et al. (1998), Rosen and Yampolsky (2000), and Young and Mihailidis (2010)

Other factors have also been shown to affect ASR performance. Literature reviews (Rosen & Yampolsky, 2000; Young & Mihailidis, 2010) presented studies investigating the relationship between ASR functioning and dysarthric speech characteristics, such as disfluency, and inconsistencies or variations in speech articulation. Young and Mihailidis (2010) found that ASR performance decreases with increased dysarthria severity and speech variability, which may be caused by different characteristics of dysarthria, for example, poor articulation and pauses within words.

Over the years, progress has emerged in ASR technologies. ASR systems started by recognizing single words or digits. After that, the development of ASR led to speaker specific, isolated words and small vocabulary systems, wherein the user needs to train the system before using it. In the 1980s, context correlation with continues speech was introduced using hidden Markov models (HMM). When the performance reached its peak, new technologies arose using deep learning, which are currently the basis of many commercial ASR systems. End-to-end deep learning uses a single model to perform speech recognition tasks. It takes an input (audio), learns it, and produces a result. ASR systems not only recognize continues speech and spoken dialog, but also understand them, as those systems are trained on huge amounts of data from huge datasets. With the help of the cloud, these systems are currently deployed on devices utilizing cloud-processed data, such as mobile phones and TV.

Various systems using different approaches have been proposed to improve ASR systems. One example is STARDUST (Parker et al., 2006) which is a system that uses speaker training by involving speakers articulating words several times, to improve ASR for people with severe dysarthria; it achieved an average 5% increase in its speech recognition rate. Another is Hawley et al. (2006), who focused on augmentative and alternative communication devices (AAC). AAC device is a communication device that receives input through either a substitute mouse, keyboard, or speech and transfers the input to spoken speech. He used the voice-input voice-output communication aid (VIVOCA), which involves (i) a confusability manager anticipating words that could be incorrectly recognized; (ii) users practicing a set of words to aid their consistency in word production (similar to the STARDUST approach); and (iii) training records used to retrain the system. CanSpeak does not use training but is based on a list of keywords tailored to each user, resulting in a range of improved accuracy rates for the four participants (Hamidi et al., 2010). Another example is the Supplemented Speech Recognition (SSR) system (Fager et al., 2010), which does not solely rely only on speech for recognition but also includes first letter identification. The user starts by typing the first letter and then uttering a word. The system then interprets the input word to the closest matching word with the same first letter. This approach resulted in comparable keystroke savings in addition to having the system match the input with fewer words in the word dataset regardless of the severity of the dysarthria case. Using ASR has also been explored as a decoding technology for impaired speech to access AAC devices as an alternative input approach.

Several literature reviews address different areas of ASR and speech impairment (Balaji & Sadashivappa, 2015; Mustafa et al., 2015; Rosen & Yampolsky, 2000; Young & Mihailidis, 2010). However, to the best of our knowledge, no recent reviews have explored progress in commercial systems and ASR in relation to people with dysarthria. This systematic review thus aims to identify studies of ASR systems, analyze the interactions between people with dysarthria and ASR systems/devices, and understand the progress achieved over the last nine years.

Materials and Methods

To achieve the aim of this review, the following research questions were defined:

RQ1: How do characteristics of users' speech affect their interactions with ASR systems?

RQ2: What acoustic input is required for effectively interacting with ASR systems?

RQ3: What ASR systems have been evaluated in dysarthria research?

RQ4: Where future ASR research could be directed?

An exhaustive search was conducted on July 2020, using the following search terms: (dysarthri* OR dysarthri* speech) AND (automatic speech recognition system* OR ASR OR virtual home assistant* OR voice-controlled digital assistant* OR conversational agent* OR commercial voice assistant* OR voice interface* OR Personal Assistant*). The search utilized these databases: ACM Digital Library, Google Scholar, IEEE Xplore digital library, Springer Link, and Science Direct. A backward and forward reference search was also performed: the backward search identified the references from certain research, while the forward search identified the research that cited a certain study.

To identify studies, we used the following inclusion criteria: (i) studies published 2011– 2022 (Siri launched in 2011, making ASR mainstream and enabling people with dysarthria to use ASR ubiquitously); (ii) studies evaluating interactions between people with dysarthria and ASR systems/devices; (iii) studies using word error rate (WER) as the measurement criteria, the most commonly used metric for accuracy. Selecting studies that use the same metrics will allow the authors to compare one thing that is common across studies. Our focus is on human–computer interactions rather than on the disorder and therapeutic intervention; therefore, we excluded clinical and therapeutic research and research that examined dysarthria in individuals with language and cognitive impairment (e.g., aphasia and dementia) to eliminate factors that affect the interaction process. The Preferred Reporting Items for Systematic Review and Meta Analysis (PRISMA) was followed (Moher et al., 2015). One author conducted the screening process, while another undertook the final review of the screening results. Out of 101 studies that were selected and assessed for eligibility, 40 were chosen to answer the research questions (see flowchart in Appendix A). To systematically review the literature, the authors used the interaction framework (Abowd & Beale 1991), which proposes that interactions between users and systems follow a four-component cycle: user, input, system, and output. We classify the ASR literature and people with dysarthria according to these four components. A table that maps the framework component with the research questions is attached in Appendix B.

Results

RQ1: *How the characteristics of users' speech affect their interactions with ASR systems?*

The first element in the interaction cycle is the *user*, who triggers the system with a voice command. To effectively use ASR systems, users must possess understandable speech characteristics and behaviors; thus, systems should be designed according to users' needs and abilities.

Research studying the interactions between people with dysarthria and ASR systems classifies dysarthria by either severity, as mild, moderate, or severe (Ballati et al., 2018b; De

Russis & Corno, 2019; M. Kim et al., 2013; Park et al. 2011; Sriranjani et al., 2015; Xiong et al., 2018)—or intelligibility, as high, medium, and low (Moore et al., 2019; Yue et al., 2020). The present article does not use several clinically defined and separate dysarthria types (e.g., flaccid, spastic, ataxic, and hypokinetic), as research about interactions does not follow this classification. Intelligibility is the degree to which a listener can understand the utterance produced by an individual with dysarthria (Yorkston et al., 1996). Because dysarthria is caused by "disturbances in muscular control over the speech mechanism" (Darley et al., 1969), it leads to reduced speech intelligibility—and the greater the severity of dysarthria, the poorer the unintelligibility.

Dysarthric speech is characterized by poor articulation, abnormal speech rate, and speech muscle fatigue. These characteristics are not necessarily exhibited by all individuals with dysarthria and may vary from one individual to another. A slow speech rate in dysarthria could prolong some words or syllables and vary vowel spacing. Additionally, speech rate variations among individuals with dysarthria may make interactions with ASR systems challenging (M. Kim et al., 2017; Rosen & Yampolsky, 2000).

Poor articulation or pronunciation is another characteristic of dysarthria. Some studies have examined the correlations between consonants, vowels, and dysarthric speech. Findings on consonants show that pronouncing them is difficult for individuals with dysarthria. For instance, unvoiced consonants were pronounced improperly (Nordberg et al., 2014; Rudzicz, 2013). Similarly, problems with pronouncing vowels were mainly associated with mid-vowels—in which the tongue is positioned midway to the roof of the mouth—and front rounded vowels (Dhanalakshmi et al., 2018).

Muscle fatigue also affects interactions with ASR, making speech more susceptible to variability and reduced quality and lowering recognition accuracy (Mustafa et al., 2015). Moreover, fatigue can make data collection challenging (Xiong et al., 2019), constraining the collection process (Joy & Umesh, 2018), which may need to occur in multiple sessions or in a session with breaks to avoid participant fatigue (Ferrier et al., 1995). Consequently, the dataset of recordings for people with dysarthria is smaller than the dataset for people with typical speech patterns (Xiong et al., 2019).

The reviewed studies proposed various approaches to overcoming these challenges to achieve better ASR performance. To improve the recognition of variable dysarthric speech, some researchers have proposed systems that can model phonetic variations, including a Kullback–Leibler divergence-based hidden Markov model (M. Kim et al., 2017; M. Kim et al., 2016; Seong et al., 2016) and a deep learning model (Xiong et al., 2018). Overcoming the limitations of dysarthric speech as training data, researchers (1) used models that require less training data (Gemmeke et al., 2014), (2) augment data by artificially generating dysarthric speech (Green et al., 2021; Jin et al., 2021; Ko et al., 2017; Liu et al., 2021; Mariya Celin et al., 2020; Vachhani et al., 2018; Xiong et al., 2019), and (3) adapt data to a given speaker (Geng et al., 2021; Takashima et al., 2020). Further, Sriranjani et al. (2015) used "data pooling," in which normal speech recordings were pooled from databases and combined with dysarthric speech data to train systems.

Other improvements to ASR systems include creating a database of recordings to help researchers train and test their systems (Marini et al., 2021; Nicolao et al., 2016; Rudzicz et al., 2012; Turrisi et al., 2021), adding features such as word prediction to systems (S. Kim et al., 2013), and adding an interface that assists with command formulation—which the authors found to be an alternative interaction modality as the structure of their system could not understand incomplete commands (Derboven et al., 2014)—to reduce fatigue and frustration. However, users may experience further difficulties because their dysarthria is often accompanied by physical disabilities (Joy & Umesh, 2018), which can limit their ability to use other forms of input, such as keyboards or touchscreens. Another approach to reducing fatigue is to use a smaller vocabulary, which has also been found to result in better accuracy (Gemmeke et al., 2014; Hamidi et al., 2015).

RQ2: *What acoustic input is required for effectively interacting with ASR systems?*

The second element in the interaction cycle is *input*, which may involve various modalities: sounds, eye gazes, and gestures. Our review focuses on acoustic inputs, such as a single word, a sentence, or continuous speech.

Fourteen studies in our review used single-word inputs, six used sentences, nine used both words and sentences, and one used words, sentences, and continuous speech. Table 1 summarizes the results, showing that isolated words were the most common input mode.

Table 1. Summary of speech modalities used in the literature

Although not explicitly stated in all papers, the choice to use isolated words or sentences depended on the system. For example, some have argued that words may be problematic for systems because limited data will be insufficient for neural networks. Some researchers have tested voice assistants, which usually accept continuous speech only.

Isolated words are an easier input for individuals with dysarthria, who have difficulty controlling their breath and speech. Thus, uttering single words is easier than articulating sentences (Allison et al., 2019; Young & Mihailidis, 2010).

Recently, Jaddoh (2021) and Lea (2022) have studied the use of nonverbal sound as a method of instruction to extend the ability of interacting with ASR systems or devices. Lea used recordings with different accents to develop a model that detects different mouth sounds, such as “pop” and “click,” as inputs, while Jaddoh suggested using nonverbal sound as a technique to control virtual home assistance.

All inputs involved recordings, either collected by researchers from participants with dysarthria and then used to test and train the system (Ballati et al., 2018a; M. Kim et al., 2017; Malavasi et al., 2016; Park et al. 201; Takashima et al., 2020) or obtained from publicly available databases containing speech data from people with dysarthria, such as TORGO (Rudzicz et al., 2012), UA-Speech (H. Kim et al., 2008), Nemours (Menendez-Pidal et al., 1996) , IDEA (Marini et al., 2021), or (Turrisi et al., 2021) . Such databases may have been used because of difficulties with recruiting people with dysarthria—for example, M. Kim et al. (2017) were unable to recruit people with severe dysarthria, and thus focused their research on mild and moderate cases only.

RQ3: *What ASR systems have been evaluated in dysarthria research?*

Different systems explored in the reviewed literature can be classified. One classification is based on input modality. Another is based on the type of speaker. For example, speaker-independent systems, such as commercial ones, are designed for all public users. However, these systems have been trained using typical speech, making them somewhat inaccessible to people with dysarthria. By contrast, speaker-dependent systems, designed and trained for a specific user, will perform best with that particular user. However, obtaining training data from users with dysarthria may be challenged by the fatigue and frustration they experience (Joy & Umesh, 2018). The third classification is speaker-adaptive, in which systems are trained, first with normal speech and then with the user's speech. These systems require little training; however, their accuracy will initially be low until they have adapted to the user's speech and can then improve.

Two more classifications derived from the reviewed literature are commercial systems (Ballati et al., 2018a, 2018b; De Russis & Corno, 2019; Moore et al., 2018) and non-commercial systems. Commercial systems are mostly speaker-independent systems, sold in the market and used by any user, that may be either built into computers (e.g., Cortana) and mobile phones (e.g., Siri) or used as standalone devices (e.g., Amazon Alexa or Google Home). Overall, researchers have agreed that these devices require improvements to their recognition and robustness to fully understand dysarthric speech, although Ballati et al. (2018a) reported that these systems can understand moderate dysarthria to some extent. Commercial systems can time out before a command is completed (Ballati et al., 2018a; Derboven et al., 2014), which may be problematic for people with dysarthria, whose speech is typically slower than normal. While there are features in some devices – such as follow-up mode in Alexa – where the device can listen for more few seconds after the user completes his commands – to the best of our knowledge, this feature has not been tested by individuals with dysarthria. Moreover, dysarthric speech may feature a breath between word syllables; many systems cannot handle this type of noise (Moore et al., 2018).

Some researchers have attempted to enhance the performance of non-commercial systems by improving the model or involving users in building applications. Improving the model may involve boosting understanding of dysarthric speech characteristics or including ASR applications. CloudCast (Malavasi et al., 2016) is an example of an application that implements a low-cost voice-controlled system to automate the home environment using the open-source software Open Home Automation Bus. Another example of an application is ALADIN (Derboven et al., 2014; Gemmeke et al., 2013), a voice user interface system for home environment automation. In this application, the users themselves train the system. Moreover, the system was designed according to user requirements and needs. Another system, CanSpeak (Hamidi et al., 2015), uses selected keywords, customized according to the user's preference, to command computer systems. Users speak selected words to carry out particular commands. Similarly, S. Kim et al. (2013) designed a mobile voice interface system that is customized for the user, with a word/sentence prediction function employed to reduce user speaking trials. When the user spell the first letter/s of the word, the prediction would appear on the screen. The user selects the prediction either verbally or by clicking on a confirmation button. This would save the user from repeated attempts to utter the word/sentence that the system might not understand.

RQ4: *Where future ASR research could be directed?*

In this section, we examine the literature on how different ASR systems perform for people with dysarthria and the effectiveness of their interactions. Performance is based on the accuracy of recognition and is commonly measured using the WER metric (McCowan et al., 2005): **WER= (S+D+I)/N**, where S is number of words substituted, D is the number of words deleted, I is the number of insertions, and N is (S + D + I + number of correct words). The studies used different models, training sets, platforms, and training times, so their results could not be directly compared.

The reviewed research is categorized into (a) non-commercial ASR systems and (b) open-source and commercial ASR systems. The research is analyzed according to the two main factors affecting ASR accuracy: the severity of the dysarthria and the speaker's intelligibility level (Mustafa et al., 2015).

Non-commercial ASR systems

Seven studies calculated the WER according to dysarthria severity. Since many individual studies tested multiple systems and we sought those with the best performance, the data included here reflect the best tools that achieved the highest accuracy of those mentioned in each study. Table 2 summarizes the results.

Some studies provided the WER for each individual speaker rather than the average performance across all participants. Therefore, to examine ASR progress, we calculated the average results presented in the studies. We found that the lowest average WER (highest accuracy) for severe cases was 43.61%, while the highest average WER was 69.3%. Based on these data,the ability of ASR systems to understand severe dysarthric speech is not acceptable.

Table 2. Accuracy (word error rate) of ASR systems according to dysarthria severity

Only five studies tested ASR accuracy according to intelligibility level. The highest WER percentage for "very low" intelligibility was 83.36%. Table 3 summarizes the results. The numbers indicate that to date, ASR systems poorly understand dysarthric speech characterized by reduced intelligibility.

Table 3. Accuracy (word error rate) of ASR systems according to dysarthria intelligibility

Researchers have also provided general testing results of ASR systems without grouping the data set into severity or intelligibility. Table 4 summarizes the results, showing that the WER was 14.9–61.9. This wide range arose from the different types of evaluation in each study. Interestingly, using sentences rather than isolated words resulted in better performance (Hermann & Magimai-Doss, 2020). This was because the model was suitable for a slow speech rate and resulted in fewer spurious words. This contrasted with other studies in which smaller words led to better performance (Rosen & Yampolsky, 2000).

Table 4. General ASR system accuracy.

Commercial systems

In commercial systems, WER is 9% for normal speech (Këpuska & Bohouta, 2017) and much poorer for people with dysarthria. Several of the reviewed studies evaluated the performance of commercial systems when used by individuals with dysarthria (Ballati et al., 2018a, 2018b; De Russis & Corno, 2019). Table 5 shows the evaluation results in terms of WER. Siri, which constantly works to transcribe all commands it receives, performed worse than other systems. Other systems sometimes indicate if they are unable to understand a command so the user can restate the command more clearly (Ballati et al., 2018b).

When responding, commercial devices rely not only on transcribed text, but also on context (Ballati et al., 2018a). Thus, in addition to WER, researchers have used various measures to evaluate the performance of commercial systems. For example, they have qualitatively studied the level of comprehension of commands as well as device responses. Thus, in voice-controlled systems, many aspects must be evaluated to understand system performance. Table 5 shows that the performance of commercial systems is poor for people with dysarthria. A possible explanation is that the data used for testing are not designed for this kind of ASR system.

Discussion

This review discusses gaps in the current research on the interaction between people with dysarthria and ASR systems and devices. Scholars have noted the importance of understanding the characteristics of users' speech and how they affect interactions. The accuracy results improved somewhat; however, the performance of these systems when used by people without dysarthria was notably different. This finding is consistent with Young and Mihailidis (2010), who indicated the similarity in ASR performance trends or patterns, despite advancements in ASR technology and variances in research methodologies, study subjects, and the types of ASR systems employed.

Following previous research by Rosen and Yampolsky (2000), papers that focused on improving systems to enhance interactions for people with dysarthria showed that individualized systems are more suitable. However, this remains to be studied in commercial systems. Additionally, the literature on commercial systems reveals that attempts to address the challenges faced by people with dysarthria when using commercial systems do not provide approaches to overcoming these challenges.

A major difficulty in comparing the performance of different systems arises from the various methods used. For example, the training data inputs were normal speech, dysarthric speech, or both. The number of data sets, the type of input modality, and dysarthria severity also varied. Mustafa et al. (2015) similarly indicated that different techniques influenced the understanding of ASR performance.

Another challenge in comparing the studies was the varying calculation and presentation of the WER. Some studies calculated WER for the entire data set, potentially including dysarthric speakers with varying severity levels. This led to less informative results and difficulties with comparing the results. Others did not mention the level of dysarthria severity. Most studies evaluated systems using publicly available databases rather than real users, and most recordings are produced in controlled rather than natural environments, this may have affected the WER results. Consequently, generalizing these findings may prove difficult (Young & Mihailidis, 2010).

The generalizability of these findings is subject to certain limitations. For instance, due to the lack of a standardized measurement for ASR, other literature may be rejected. An additional limitation is the exclusion of research that examined dysarthria in individuals with language and cognitive impairment. Future research focusing on all individuals who experience dysarthria would strengthen the present findings. Additionally, a meta-analysis could be undertaken to fully examine the results of the studies in this review.

Conclusion

Despite improvements enumerated by various studies over the past nine years, challenges in interacting with ASR systems continue. According to the literature, using dysarthric speech for training data improves system performance. However, the difficulty in generating large databases of dysarthric speech remains a continual challenge. Moreover, the current available databases are limited because of repetition of utterances by the same users and the lack of many unique utterances. Additionally, these databases are not suitable for conversational commercial systems.

Ensuring that people with dysarthria have the same opportunities to use commercial ASR devices as those without dysarthria requires involving real users in design and testing processes. Finally, it is necessary to continue improving ASR system performance and to identify a methodology for interaction that allows more accessibility.

References

Abowd, G. D., & Beale, R. (1991). Users, systems and interfaces: A unifying framework for interaction. In D. Diaper & N. Hammond, *Proceedings of the Sixth Conference of the British Computer Society Human Computer Interaction Specialist Group: People and computers IV* (pp. 73–87). Cambridge University Press.

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

Balaji, V., & Sadashivappa, G. (2015). Speech disabilities in adults and the suitable speech recognition software tools—a review. In J. Wu & C. B. Westphall (Eds.), *2015 International Conference on Computing and Network Communications (CoCoNet)* (pp. 559–564). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/coconet.2015.7411243

Ballati, F., Corno, F., & De Russis, L. (2018a). Assessing virtual assistant capabilities with Italian dysarthric speech. In F. Hwang (Ed.), *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 93–101). Association for Computing Machinery. https://doi.org/10.1145/3234695.3236354

Ballati, F., Corno, F., & De Russis, L. (2018b). “Hey Siri, do you understand me?”: Virtual assistants and dysarthria. In I. Chatzigiannakis, Y. Tobe, P. Novais, & O. Amft (Eds.), *Intelligent Environments 2018: Workshop Proceedings of the 14th International Conference on Intelligent Environments* (Vol. 23, pp. 557–566). IOS Press.

Cai, S., Lillianfeld, L., Seaver, K., Green, J., Brenner, M., Nelson, P., & Sculley, D. (n.d.). A Voice-Activated Switch for Persons with Motor and Speech Impairments: Isolated-Vowel Spotting Using Neural Networks. Retrieved March 8, 2022, from *https://storage.googleapis.com/pub-tools-public-publication-data/pdf/70c549e87cbf7e8a3d9c7a3c8e6f862362a9dbee.pdf*

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential Diagnostic Patterns of Dysarthria. *Journal of Speech and Hearing Research*, 12(2), 246–269. https://doi.org/10.1044/jshr.1202.246

Derboven, J., Huyghe, J., & De Grooff, D. (2014). Designing voice interaction for people with physical and speech impairments. In V. Roto (Ed.), *Proceedings of the 8th Nordic Conference on Human–Computer Interaction: Fun, fast, foundational* (pp. 217–226). Association for Computing Machinery. https://doi.org/10.1145/2639189.2639252

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

Dhanalakshmi, M., Mariya Celin, T. A., Nagarajan, T., & Vijayalakshmi, P. (2018). Speech-input speech-output communication for dysarthric speakers using HMM-based speech recognition and adaptive synthesis system. *Circuits, Systems, and Signal Processing*, *37*(2), 674–703. https://doi.org/10.1007/s00034-017-0567-9

Españа-Bonet, C., & Fonollosa, J. A. R. (2016). Automatic speech recognition with deep neural networks for impaired speech. In A. Abad (Ed.), *Advances in speech and language technologies for Iberian languages*. *IberSPEECH 2016. Lecture notes in computer science* (Vol. 10077, pp. 97–107). Springer. https://doi.org/10.1007/978-3-319-49169-1_10.

Fager, S. K., Beukelman, D. R., Jakobs, T., & Hosom, J.-P. (2010). Evaluation of a Speech Recognition Prototype for Speakers with Moderate and Severe Dysarthria: A Preliminary Report. *Augmentative and Alternative Communication*, 26(4), 267–277. https://doi.org/10.3109/07434618.2010.532508

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*(3), 165–175. https://doi.org/10.1080/07434619512331277289

Fried-Oken, M. (1985). Voice recognition device as a computer interface for motor and speech impaired people. *Archives of Physical Medicine and Rehabilitation*, *66*(10), 678–68.

Gemmeke, J. F., Sehgal, S., Cunningham, S., & Van Hamme, H. (2014). Dysarthric vocal interfaces with minimal training data. In Institute of Electrical and Electronics Engineers (Ed.), *2014 IEEE Spoken Language Technology Workshop (SLT 2014)* (pp. 248–253). IEEE. https://doi.org/10.1109/SLT.2014.7078582

Gemmeke, J., Ons, B., Tessema, N. M., Van hamme, H., Van de Loo, J., De Pauw, G., Daelemans, W., Huyghe, J., Derboven, J., Vuegen, L., Van Den Broeck, B., Karsmakers, P., & Vanrumste, B. (2013, August 1). Self-taught assistive vocal interfaces: An overview of the ALADIN project. *Lirias.kuleuven.be;* ISCA; Baixas. https://lirias.kuleuven.be/301023?limo=0

Geng, M., Liu, S., Yu, J., Xie, X., Hu, S., Ye, Z., Jin, Z., Liu, X., & Meng, H. (2021). Spectro-Temporal Deep Features for Disordered Speech Assessment and Recognition. Interspeech 2021. https://doi.org/10.21437/interspeech.2021-60

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*. https://doi.org/10.21437/interspeech.2021-1384

Hamidi, F., Baljko, M., Ecomomopoulos, C., Livingston, N. J., & Spalteholz, L. G. (2015). Co-designing a speech interface for people with dysarthria. *Journal of Assistive Technologies*, *9*(3), 159–173. https://doi.org/10.1108/jat-10-2014-0026

Hamidi, F., Baljko, M., Livingston, N., & Spalteholz, L. (2010). CanSpeak: A customizable speech interface for people with dysarthric speech. In K. Miesenberger, J. Klaus, W. Zagler, & A. Karshmer (Eds.), *Computers helping people with special needs: 12th International Conference, ICCHP 2010: Proceedings* (pp. 605–612). Springer. https://doi.org/10.1007/978-3-642-14097-6_97

Hawley, M. S., Enderby, P., Green, P., Cunningham, S., & Palmer, R. (2006). Development of a voice-input voice-output communication aid (VIVOCA) for people with severe dysarthria. In K. Miesenberger, J. Klaus, W. L. Zagler, & A. I. Karshmer (Eds.), *Computers helping people with special needs: 10th International Conference, ICCHP 2006: Proceedings* (pp. 882–885). Springer. https://doi.org/10.1007/11788713_128

Hermann, E., & Magimai-Doss, M. (2020). Dysarthric speech recognition with lattice-free MMI. In Institute of Electrical and Electronics Engineers (Ed.), *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6109–6113). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053549

Holyfield, C., & Drager, K. (2021). Integrating Familiar Listeners and Speech Recognition Technologies into Augmentative and Alternative Communication Intervention for Adults with Down syndrome: Descriptive Exploration. Assistive Technology. https://doi.org/10.1080/10400435.2021.1934610

Jaddoh, A., Loizides, F., & Rana, O. (n.d.). Non-Verbal Interaction with Virtual Home Assistants for People with Dysarthria. *In Journal on Technology and Persons with Disabilities Santiago*, J. Retrieved March 8, 2022, from https://scholarworks.csun.edu/bitstream/handle/10211.3/219936/JTPD-2021-p071.pdf?sequence=1

Jin, Z., Geng, M., Xie, X., Yu, J., Liu, S., Liu, X., & Meng, H. (2021). Adversarial Data Augmentation for Disordered Speech Recognition. *ArXiv:2108.00899* [Cs, Eess]. https://arxiv.org/abs/2108.0089

Joy, N. M., & Umesh, S. (2018). Improving acoustic models in TORGO dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(3), 637–645. https://doi.org/10.1109/tnsre.2018.2802914

Këpuska, V. Z., & Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *International Journal of Engineering Research and Application*, *7*(3), 20–24.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. In International Speech Communication Association (Ed.), *9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)* (pp. 1741–1744). Curran Associates.

Kim, M., Kim, Y., Yoo, J., Wang, J., & Kim, H. (2017). Regularized speaker adaptation of KL-HMM for dysarthric speech recognition. *IEEE Transactions on Neural Systems* *and Rehabilitation Engineering*, *25*(9), 1581–1591. https://doi.org/10.1109/tnsre.2017.2681691

Kim, M., Wang, J., & Kim, H. (2016). Dysarthric speech recognition using Kullback–Leibler divergence-based hidden Markov model. In International Speech Communication Association (Ed.), *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016): Understanding speech processing in humans and machines* (pp. 2671–2675). Curran Associates. https://doi.org/10.21437/Interspeech.2016-776

Kim, M., Yoo, J., & Kim, H. (2013). Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models. In F. Bimbot (Ed.), *14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013): Speech in life sciences and human societies* (pp. 3622–3626). International Speech Communication Association.

Kim, S., Hwang, Y., Shin, D., Yang, C.-Y., Lee, S.-Y., Kim, J., Kong, B., Chung, J., Cho, N., Kim, J.-H., & Chung, M. (2013). VUI development for Korean people with dysarthria. *Journal of Assistive Technologies*,*7*(3). https://doi.org/10.1108/jat-10-2012-0031

Kim, Y., Kent, R. D., & Weismer, G. (2011). An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *Journal of Speech, Language, and Hearing Research*, *54*(2), 417–429. https://doi.org/10.1044/1092-4388(2010/10-0020)

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In Institute of Electrical and Electronics Engineers (Ed.), *ICASSP 2017—2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5220–5224). IEEE. https://doi.org/10.1109/icassp.2017.7953152

Lea, C., Huang, Z., Jain, D., Tooley, L., Liaghat, Z., Thelapurath, S., Findlater, L., & Bigham, J. P. (2022). Nonverbal Sound Detection for Disordered Speech. *ArXiv:2202.07750 [Cs, Eess]. https://arxiv.org/abs/2202.07750*

Liu, S., Geng, M., Hu, S., Xie, X., Cui, M., Yu, J., Liu, X., & Meng, H. (2021). Recent Progress in the CUHK Dysarthric Speech Recognition System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2267–2281. https://doi.org/10.1109/TASLP.2021.3091805

Liu, S., Hu, S., Liu, X., & Meng, H. (2019). On the use of pitch features for disordered speech recognition. In International Speech Communication Association (Ed.), *20th Annual*

Conference of the International Speech Communication Association (INTERSPEECH 2019): Crossroads of speech and language (pp. 4130–4134). Curran Associates.

Malavasi, M., Turri, E., Motolese, M. R., Marxer, R., Farwer, J., Christensen, H., Desideri, L., Tamburini, F., & Green, P. (2016). An innovative speech-based interface to control AAL and IoT solutions to help people with speech and motor disability. In N. Cassidu, C. Porfirione, A. Monteriù, & F. Cavallo (Eds.), *Ambient assisted living: Italian Forum 2017* (pp. 269–278). Springer.

Marini, M., Vigano, M., Corbo, M., Zettin, M., Simoncini, G., Fattori, B., D'Anna, C., Donati, M., & Fanucci, L. (2021). IDEA: An Italian Dysarthric Speech Database. 2021 *IEEE Spoken Language Technology Workshop (SLT).* https://doi.org/10.1109/slt48900.2021.9383467

Mariya Celin, T. A., Nagarajan, T., & Vijayalakshmi, P. (2020). Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 346–354. https://doi.org/10.1109/jstsp.2020.2972161

McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., & Bourlard, H. (2005). *On the use of information retrieval measures for speech recognition evaluation* (Research Report 04-73). Idiap Research Institute.

Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996). The Nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing: ICSLP'96* (Vol. 3, pp. 1962–1965). Institute of Electrical and Electronics Engineers.

Mengistu, K. T., & Rudzicz, F. (2011). Adapting acoustic and lexical models to dysarthric speech. In Institute of Electrical and Electronics Engineers (Ed.), *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4924–4927). IEEE. https://10.1109/ICASSP.2011.5947460

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews, 4(1). https://doi.org/10.1186/2046-4053-4-1

Moore, M., Saxon, M., Venkateswara, H., Berisha, V., & Panchanathan, S. (2019). Say what? A dataset for exploring the error patterns that two ASR engines make. In International

Speech Communication Association (Ed.), *20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019): Crossroads of speech and language* (pp. 2528–2532). Curran Associates.

Moore, M., Venkateswara, H., & Panchanathan, S. (2018). Whistle-blowing ASRs: Evaluating the need for more inclusive automatic speech recognition systems. In International Speech Communication Association (Ed.), *19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018): Speech research for emerging markets in multilingual societies*. Curran Associates.

Mustafa, M. B., Rosdi, F., Salim, S. S., & Mughal, M. U. (2015). Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Systems with Applications*, *42*(8), 3924–3932. https://doi.org/10.1016/j.eswa.2015.01.033

Na, M., & Chung, M. (2016). Optimizing vocabulary modeling for dysarthric speech recognition. In K. Miesenberger, C. Bühler, & P. Penaz (Eds.), *Computers helping people with special needs. ICCHP 2016. Lecture notes in computer science* (Vol. 9759, pp. 507–510). Springer. https://doi.org/10.1007/978-3-319-41267-2_71

Nicolao, M., Christensen, H., Cunningham, S., Green, P., & Hain, T. (2016, May 27). A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus. *Proceedings of LREC 2016*. https://eprints.whiterose.ac.uk/109262/

Nordberg, A., Miniscalco, C., & Lohmander, A. (2014). Consonant production and overall speech characteristics in school-aged children with cerebral palsy and speech impairment. *International Journal of Speech-Language Pathology*, *16*(4), 386–395. https://doi.org/10.3109/17549507.2014.917440

Park, J. H., Seong, W. K., & Kim, H. K. (2011). Preprocessing of Dysarthric Speech in Noise Based on CV–Dependent Wiener Filtering. *Proceedings of the Paralinguistic Information and Its Integration in Spoken Dialogue Systems Workshop*, 41–47. https://doi.org/10.1007/978-1-4614-1335-6_6

Parker, M., Cunningham, S., Enderby, P., Hawley, M., & Green, P. (2006). Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project. *Clinical Linguistics & Phonetics*, *20*(2–3), 149–156. https://doi.org/10.1080/02699200400026884

Rosen, K., & Yampolsky, S. (2000). Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication*, *16*(1), 48–60. https://doi.org/10.1080/07434610012331278904

Rudzicz, F. (2012). Using articulatory likelihoods in the recognition of dysarthric speech. *Speech Communication*, *54*(3), 430–444. https://doi.org/10.1016/j.specom.2011.10.006

Rudzicz, F. (2013). Adjusting dysarthric speech signals to be more intelligible. *Computer Speech & Language*, *27*(6), 1163–1177. https://doi.org/10.1016/j.csl.2012.11.001

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, *46*(4), 523–541. https://doi.org/10.1007/s10579-011-9145-0

Seong, W. K., Kim, N., Ha, H. K., & Kim, H. K. (2016). A discriminative training method incorporating pronunciation variations for dysarthric automatic speech recognition. In Institute of Electrical and Electronics Engineers (Ed.), *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 96–100). IEEE. https://doi.org/10.1109/APSIPA.2016.7820840

Sriranjani, R., Ramasubba Reddy, M., & Umesh, S. (2015). Improved acoustic modeling for automatic dysarthric speech recognition. In Indian Institute of Technology (Ed.), *2015 Twenty First National Conference on Communications (NCC 2015)* (pp. 242–247). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/NCC.2015.7084856

Takashima, R., Takiguchi, T., & Ariki, Y. (2020). Two-step acoustic model adaptation for dysarthric speech recognition. In Institute of Electrical and Electronics Engineers (Ed.), *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6104–6108). IEEE.

Thomas-Stonell, N., Kotler, A.-L., Leeper, H., & Doyle, P. (1998). Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication*, *14*(1), 51–56. https://doi.org/10.1080/07434619812331278196

Turrisi, R., Braccia, A., Emanuele, M., Giulietti, S., Pugliatti, M., Sensi, M., Fadiga, L., & Badino, L. (2021). EasyCall corpus: a dysarthric speech dataset. *ArXiv:2104.02542 [Cs]*. https://arxiv.org/abs/2104.02542

Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. In International Speech Communication Association

(Ed.), *19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018): Speech research for emerging markets in multilingual societies* (pp. 471–475). Curran Associates.

Wang, D., Wang, X., & Lv, S. (2019). An Overview of End-to-End Automatic Speech Recognition. Symmetry, 11(8), 1018. https://doi.org/10.3390/sym11081018

Xiong, F., Barker, J., & Christensen, H. (2018). Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In S. Doclo & P. Jax (Eds.), *Proceedings of the 13th ITG Conference on Speech Communication* (pp. 1–5). VDE Verlag.

Xiong, F., Barker, J., & Christensen, H. (2019). Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In Institute of Electrical and Electronics Engineers (Ed.), *ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5836–5840). IEEE. https://doi.org/10.1109/icassp.2019.8683091

Yorkston, K. M., Beukelman, D. R., & Bell, K. R. (1988). *Clinical management of dysarthric speakers*. Little, Brown and Company.

Yorkston, K., Strand, E., & Kennedy, M. (1996). Comprehensibility of Dysarthric Speech. American Journal Of Speech-Language Pathology, 5(1), 55-66. doi: 10.1044/1058-0360.0501.55

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, *22*(2), 99–112. https://doi.org/10.1080/10400435.2010.483646

Yue, Z., Xiong, F., Christensen, H., & Barker, J. (2020). Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In Institute of Electrical and Electronics Engineers (Ed.), *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*IEEE.

Table 1. Summary of speech modalities used in the literature

Input modality	Studies
Isolated words	Geng et al. (2021); Jin et al. (2021); M. Kim et al. (2017); M. Kim et al. (2016); S. Kim et al. (2013); Liu et al. (2019); Malavasi et al. (2016); Marini et al. (2021); Mariya Celin et al. (2020); Mengistu & Rudzicz (2011); Na & Chung (2016); Vachhani et al. (2018); Xiong et al. (2018, 2019); Park et al. (2011)
Sentences	Ballati et al. (2018b); De Russis & Corno (2019); Seong et al. (2016); Takashima et al. (2020); Rudzicz (2012); Green et al. (2021)

Words and sentences	España-Bonet & Fonollosa (2016); Hamidi et al. (2015); Hermann & Magimai-Doss (2020); Liu et al. (2021); Moore et al. (2019); Moore et al. (2018); Sriranjani et al. (2015); Turrisi et al., (2021); Yue et al. (2020)
Words, sentences, and continuous speech	Joy & Umesh (2018)

Table 2. Accuracy (word error rate) of ASR systems according to dysarthria severity (some papers are not included. They do not provide precise numbers. For example, they present WER in charts)

Study	Input modality	Dysarthria severity			
		Severe	Severe–moderate	Moderate	Mild
Xiong et al. (2019)	Isolated words	67.83	27.55	26.41	9.71
Joy & Umesh (2018)	All modalities	46.52	46.97	56.26	25.94
Yue et al. (2020)	Isolated words and sentences	57.6	–	33.0	14.3
Sriranjani et al. (2015)	Isolated words and sentences	43.61	–	32.63	21.14
Vachhani et al. (2018)	Isolated words	69.3	36.45	21.32	1.35
M. Kim et al. (2017)	Isolated words	–	–	28	28
Seong et al. (2016)	Sentences	–	–	46.00	29.11

Table 3. Accuracy (word error rate) of ASR systems according to dysarthria intelligibility

	Intelligibility			
Study	High	Medium	Low	Very low
Geng et al., 2021	7.91	16.80	27.16	59.83
Jin et al. (2021)	7.75	16.50	27.37	61.42
Liu et al., (2021)	7.55	16.47	26.84	62.37
Mariya Celin et al. (2020)	5.81	16.38	4.47	50.36
Xiong et al. (2018)	16.17	31.41	45.79	83.36

Table 4. General accuracy of ASR systems (some papers are not included. They do not provide precise numbers. For example, they present WER in charts)

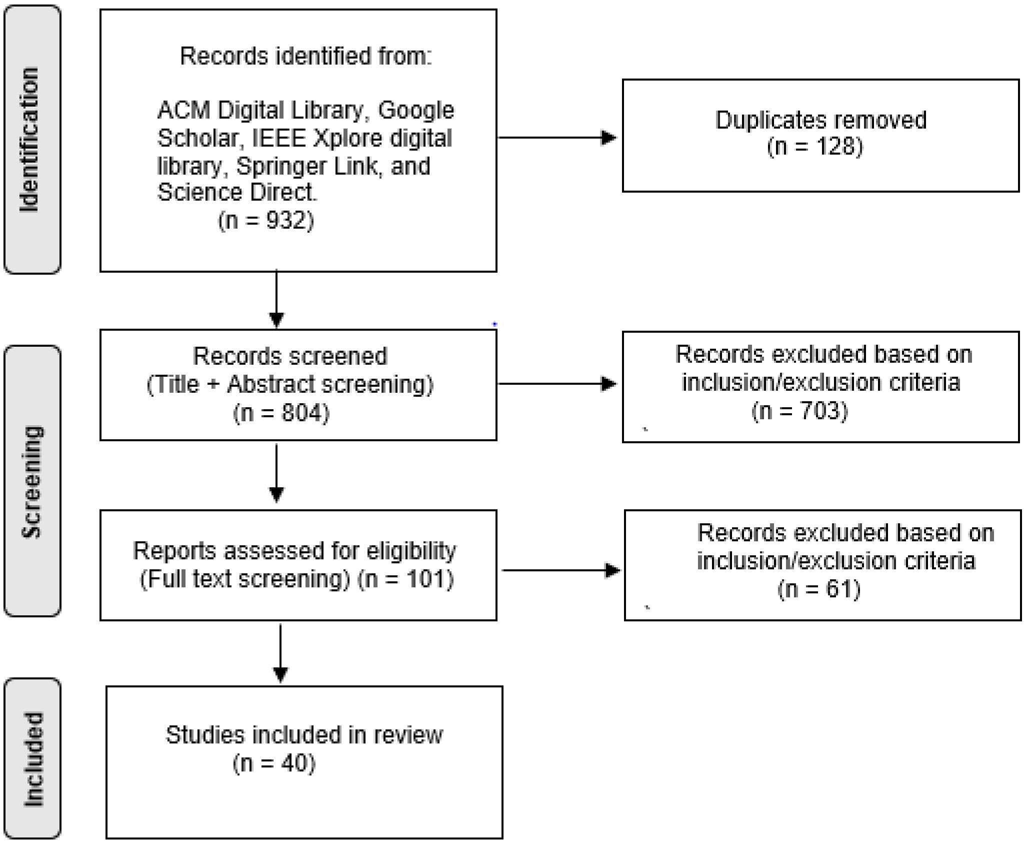
Study	Database	Input modality	Word error rate (%)
Geng et al. (2021)	UA-Speech	Isolated words	25.6
Jin et al. (2021)	UA-Speech	Isolated words	25.8
M. Kim et al. (2017)	Recordings	Isolated words	28.0
Liu et al. (2019)	CUDYS	Isolated words	28.2
	UA-Speech	Isolated words	31.0
M. Kim et al. (2016)	Korean	Isolated words	33.4
Marini et al. (2021)	IDEA	Isolated words	14.99
Hermann & Magimai-Doss (2020)	Recordings	Sentences	25.9
		Sentences	53.7
		Isolated words	42.9
Rudzicz (2012)	TORGO and MOCHA	Sentences	34.7
Turrisi et al.(2021)	Easycall	Sentences	61.9

Table 5. Accuracy (word error rate) of commercial ASR systems *(*some *p*apers are not included. They do not provide precise numbers. For example, *they present* WER in charts*)*

Study	Dysarthria severity	Automatic speech recognition system				
		Google	Siri	IBM	Microsoft	Sphinx
De Russis & Corno (2019)	Severe	78.21	–	89.08	78.59	–
Ballati et al. (2018a)	Moderate	24.88	70.89	–	39.39	–
Ballati et al. (2018b)	Various	15.38	69.41	–	–	–
Moore et al. (2018)	Various	43	–	–	–	126

Appendix A

Figure 1 PRISMA-P 2015 flowchart.
Figure 1 Alt Text: PRISMA-P flowchart that shows the studies identification, screening, and inclusion process.



Appendix B

Table 6. Mapping between codes and research question.

Interaction framework component	Explanation	Research Question
Users	The first element in the interaction cycle is the user, who triggers the system with a voice command. To effectively use ASR systems, users' speech characteristics and behaviours must be understood; thus, systems should be designed according to users' needs and abilities.	How do characteristics of users' speech affect their interactions with ASR systems?
Acoustic input	System input may involve various modalities, including sounds, eye gazes, and gestures. However, in this review, our focus was on acoustic input only. Acoustic input may involve a single word, a sentence, continuous speech.	What kind of acoustic input is required to effectively interact with ASR systems?

System	ASR systems categories	What ASR systems have been evaluated in dysarthria research?
output	the performance and accuracy of different ASR systems for people with dysarthria and the effectiveness of their interactions.	Where future ASR research could be directed?