International Journal of Speech Technology (2021) 24:771–779
https://doi.org/10.1007/s10772-021-09836-w

Automatic speech recognition in neurodegenerative disease

Benjamin G. Schultz[1,6] · Venkata S. Aditya Tarigoppula[2,3] · Gustavo Noffs[1] · Sandra Rojas[1] · Anneke van der Walt[4] · David B. Grayden[2,3] · Adam P. Vogel[1,5]

Received: 19 August 2020 / Accepted: 31 March 2021 / Published online: 4 May 2021

Abstract
Automatic speech recognition (ASR) could potentially improve communication by providing transcriptions of speech in real time. ASR is particularly useful for people with progressive disorders that lead to reduced speech intelligibility or difficulties performing motor tasks. ASR services are usually trained on healthy speech and may not be optimized for impaired speech, creating a barrier for accessing augmented assistance devices. We tested the performance of three state-of-the-art ASR platforms on two groups of people with neurodegenerative disease and healthy controls. We further examined individual differences that may explain errors in ASR services within groups, such as age and sex. Speakers were recorded while reading a standard text. Speech was elicited from individuals with multiple sclerosis, Friedreich's ataxia, and healthy controls. Recordings were manually transcribed and compared to ASR transcriptions using Amazon Web Services, Google Cloud, and IBM Watson. Accuracy was measured as the proportion of words that were correctly classified. ASR accuracy was higher for controls than clinical groups, and higher for multiple sclerosis compared to Friedreich's ataxia for all ASR services. Amazon Web Services and Google Cloud yielded higher accuracy than IBM Watson. ASR accuracy decreased with increased disease duration. Age and sex did not significantly affect ASR accuracy. ASR faces challenges for people with neuromuscular disorders. Until improvements are made in recognizing less intelligible speech, the true value of ASR for people requiring augmented assistance devices and alternative communication remains unrealized. We suggest potential methods to improve ASR for those with impaired speech.

Keywords Automatic Speech Recognition · Dysarthria · Neurodegenerative disease · Augmented assistive communication technology · Communication

Automatic speech recognition (ASR) systems help digital machines interpret spoken speech and automate human tasks, such as typing text and web searches. ASR-based technologies are ubiquitous with an ever-increasing inventory of applications including smart phones, in-car systems, healthcare, intelligent assistive devices, security, banking, retail, telephony, computers, education, and smart homes (Owens, 2006). At present, however, this technology is not optimized for all users. People with neurological disease often present with impaired speech (i.e., dysarthria). Dysarthria is a speech disorder caused by impaired neurological function, motor control, and/or speech articulators. These speech impairments can result from acquired brain or spinal cord injuries (e.g., stroke) as well as congenital and neurodegenerative diseases, and age-related neurological decline (Liégeois et al., 2010; Magee, Copland, & Vogel, 2019; Noffs et al., 2018; Poole et al., 2017; Rojas, Kefalianos, & Vogel, 2020). Commercial ASR services are developed

✉ Benjamin G. Schultz
ben.schultz@unimelb.edu.au

1 Centre for Neuroscience of Speech, Department of Audiology & Speech Pathology, The University of Melbourne, 550 Swanston Street, Carlton, Melbourne, VIC 3053, Australia

2 Department of Biomedical Engineering, The University of Melbourne, Melbourne, Australia

3 ARC Training Centre in Cognitive Computing for Medical Technologies, The University of Melbourne, Melbourne, Australia

4 Central Clinical School, Department of Neuroscience, Monash University, Melbourne, Australia

5 Redenlab, Melbourne, Australia

6 Department of Neuropsychology and Psychopharmacology, Faculty of Psychology & Neuroscience, Maastricht University, Maastricht, The Netherlands

and maintained using large datasets typically acquired from people with unimpaired speech. ASR services may, consequently, lose their proficiency with impaired speech due to their underrepresentation in training datasets. This leads to increased errors in ASR for dysarthric speech (De Russis & Corno, 2019; Mengistu & Rudzicz, 2011; Rosen & Yampolsky, 2000; Young & Mihailidis, 2010). Therefore, it is important to assess the functionality of ASR for populations that may exhibit abnormal speech to ensure these groups can access voice-assistant technology (e.g., emails, playing music, operating a television, using appliances).

Although previous studies inform us of shortcomings specific to ASR systems for dysarthric speech, they do not provide direct comparisons between commercial offerings of ASR nor do they compare performance between dysarthria types or causes. For example, some studies have examined ASR accuracy using databases of dysarthric speech (e.g., the TORGO database) but do not differentiate between causes of dysarthria (e.g., cerebral palsy and amyotrophic lateral sclerosis) which may produce different speech recognition errors (De Russis & Corno, 2019; Mengistu & Rudzicz, 2011). Previous studies have generally contained small sample sizes (e.g., < 10 per group) and, although qualitative comparisons were discussed, do not report quantitative statistical comparisons between controls and dysarthric speakers (Blaney & Wilson, 2000; De Russis & Corno, 2019; Mengistu & Rudzicz, 2011; Raghavendra, Rosengren, & Hunnicutt, 2001; Thomas-Stonell et al., 1998). As such, studies did not have sufficient statistical power to examine potential effects of individual differences (e.g., sex and age). The work presented here focuses on the performance of state-of-the-art commercial ASR services (Amazon Web Services Transcribe, Google Cloud Speech, and IBM Watson Speech-to-Text) for healthy speech and impaired speech from two groups with compromised speech systems: multiple sclerosis (MS) and Friedreich's ataxia (FA). We further examine the effects of age and sex on ASR accuracy.

An estimated 2.3 million people worldwide have MS (Wallin et al., 2019) and many present with dysarthria (Noffs et al., 2018). MS is a chronic autoimmune disease that disrupts the communication between the brain and body by attacking the myelin sheath of nerve fibers. These dysfunctions can result in weakness and reduced coordination of the muscles associated with speech production (tongue, lips, mandible, vocal cords, diaphragm), resulting in impaired speech. People with MS and dysarthria can present with reduced voice quality, imprecise articulation, impaired stress patterns, slower speech rate, reduced breath support and poor pitch and loudness control (Noffs et al., 2018).

Friedreich's ataxia (FA) is a neurodegenerative disease that is less common than MS; FA affects 1.7–4.7 out of 100,000 people or 132.6–336.6 thousand people worldwide (Delatycki, Williamson, & Forrest, 2000; Klockgether, 2007) but often has a more debilitating impact on the speech (Gibilisco & Vogel, 2013). FA is a multisystem degenerative disease and the most common hereditary ataxia. Symptoms typically present in the second decade of life but can occur earlier or later depending on genetic profiles. Deficits include progressive limb and gait ataxia, optic and auditory neuropathy, scoliosis, cardiomyopathy and swallowing and speech deficits (Delatycki & Bidichandani, 2019). Speech in FA is characterized by reduced pitch variation, reduced loudness control, impaired timing, strained voice quality, reduced breath support, hypernasality, and imprecise production of consonants (Folker et al., 2010; Poole et al., 2015; Vogel et al., 2017).

In both MS and FA, speech typically declines as disease severity increases (Noffs et al., 2020; Rosen et al., 2012). The nature of these diseases and how users interact with technology are sometimes different to those of "healthy speakers". In the case of people with dysarthria, there is often a disconnect between what ASR aspires to do and what it can offer in practice. The typical age of onset for MS is between 20 and 50 years with an average of 34 years (Stoppler, 2019) and before 25 years for FA (Harding, 1983). Therefore, MS and FA primarily affect those who could potentially be the most active users of ASR technology (Shih, 2020). It is imperative that people with impaired speech are not left behind in this age of technology. Our objective was to quantify the problem of poor performance of commercial ASR services in people with impaired speech with the aim of guiding future design solutions. Here, we examined ASR performance on impaired speech using accuracy measures for transcribing multiple consecutive words. Potential mediating variables of age, disease duration, and sex were also explored. The present research has the potential to increase the quality of life for people with dysarthria by improving their access to technologies that facilitate communication and automate manual tasks through voice commands.

1 Methods

To ensure that we could detect differences between groups and examine the effects of sex and age, we required a larger sample size than those in publicly available databases. Moreover, our objective was to examine ASR performance for healthy controls and populations with FA and MS performing the same task under comparable conditions; to the knowledge of the authors, no such speech database is available. Therefore, we developed our own speech database using the protocols discussed below.

Table 1 Demographic information for healthy controls (HC), Friedreich's ataxia (FA), and multiple sclerosis (MS)

Variable	Statistic	Group		
		HC	FA	MS
Age (years)	Mean	49.22	39.25	44.38
	SD	11.85	16.74	11.69
	Min–Max	26–67	13–68	24–66
Sex	Female (Male)	14 (18)	16 (16)	16 (16)
Disease duration (years)	Mean	NA	24.16	13.44
	SD	NA	14.85	8.00
	Min–Max	NA	3–49	2–32
Disease Severity (FA = FARS, MS = EDSS)	Mean	NA	1.41	3.94
	SD	NA	0.95	1.66
	Min–Max	NA	0–3	1–7

Disease severity was measured using the Friedreich's ataxia rating scale (FARS; 0 = less severe, 3 = most severe) for FA and the expanded disability status scale (EDSS; 0 = no disability, 2.5 = mild disability, 5 = impairment to daily activities, 7.5 = unable to take more than a few steps and restricted to wheelchair, 9.5 = confined to bed and completely dependent) for MS

1.1 Participants

Three groups of participants were recruited: Individuals with multiple sclerosis (N = 32), individuals with Friedreich ataxia (N = 32), and 32 healthy controls (see demographic information in Table 1). All speakers were Australians with Australian accents. Participants in the clinical groups had confirmed diagnosis of disease according to published diagnosis criteria and confirmed by a neurologist. The recruitment of these participants was agnostic to their speech quality. Groups contained a subsample of 21 participants matched for sex and age, with a maximum age difference of 2 years.

1.2 Materials

Speech was recorded via an AKG C520 cardioid head-mounted condenser microphone (frequency range = 20–20,000 Hz; sensitivity = −43 dB) connected to a Roland Quad-Capture recorder sound card at a sample rate of 44.1 kHz, quantized at 16 bits. Speech recordings were screened prior to automatic speech recognition to manually remove speech artefacts and background noise using Audacity (Mazzoni & Dannenberg, 2012).

1.3 Procedure

Participants were seated in a quiet room without acoustic isolation to reflect real-world settings. The head-mounted microphone was placed approximately 8 cm from their mouth at a 45° angle. They performed a reading task where a phonetically balanced written passage was read aloud (The Grandfather Passage, Van Riper, 1963, see "Appendix") and speech was recorded. No time restrictions were placed on the reading and durations ranged from 36 to 183 s. Manual transcriptions of the recordings were performed prior to ASR implementation using the written passage as a template.

1.4 Automatic speech recognition implementation

Automatic speech recognition algorithms were implemented using custom-made Python scripts (Python3.6; Rossum, 2019) that used the *SpeechRecognition* library (Zhang, 2017). These scripts were merely wrappers to load and submit the audio recordings to ASR services and save the output transcripts. ASR was implemented by each of the three services: AWS, Google Cloud, and IBM Watson. The Australian English language model was used for AWS and Google Cloud. IBM Watson does not yet provide an Australian English language model, so the US English model was used instead.[1] For AWS and IBM Watson, the full passage could be transcribed in one instance. Google Cloud could only transcribe 60 s of audio per file. Therefore, audio passages were analyzed in 50-s frames with 5 s of overlap. Google Cloud transcriptions were then aligned by matching the last 5 s of the previous frame with the first 5 s of the next frame, then visually inspected to ensure alignment occurred correctly. Text was analyzed using custom-made MATLAB (version R2019a; MathWorks, 2019) scripts that measured the percentage of consecutive words (nGrams; one word, two words, three words) that were transcribed correctly by the ASR services based on the manual transcripts. All capitalization and punctuation were removed prior to analyzing the text and all numbers were converted to text. The percentage of accurately transcribed phrases for each recording served as our measure for evaluating ASR performance. Confidence estimates of the ASR transcriptions were also extracted to assess the relationship between ASR accuracy and self-monitoring.

[1] We also implemented the UK English language model. This produced similar patterns of results but with lower accuracy on average than the US English model [$F(1, 92) = 28.5$, $p < .001$, $\eta^2_G = .003$]. We proceeded to report the results of IBM Watson using the US English model to provide the best representation of IBM Watson's capabilities. Similarly, we confirmed that the AU English model performed with higher accuracy for the UK and US English models for Amazon Web Services and Google Cloud ($ps < .001$).

1.5 Statistical analysis

Non-linear mixed effects models (nLMEM) were used due to unequal variance between healthy controls, and people with FA and MS. nLMEMs were fit to accuracy data with fixed factors Group (3; HC, FA, MS; *between-subjects*), ASR service (3; Amazon Web Service, Google Cloud, IBM Watson; *within-subjects*), and nGram (3; one word, two words, three words; *within-subjects*) with variance allowed to vary between levels of Group. We used the maximal random effects structure justified by the experimental design (Barr et al., 2013), that is, nGram nested within ASR service, and ASR service nested within Participant, with Age as a random predictor. To assess evidence for the null hypothesis, we used Bayes Factor where evidence for the null hypothesis (BF_{01}) was interpreted as 1 = no evidence, 3–10 = moderate evidence, and 10–30 = strong evidence (Jeffreys, 1998).

Data were analyzed using R software (R Core Team, 2013). nLMEMs were performed using the *lme* function of the *nlme* library (Pinheiro et al., 2015) using Satterthwaite's method of approximation for degrees of freedom. *F*-statistics, significance values, and effect sizes (generalized eta squared; η^2_G where .02 = small, .13 = medium, and over .26 = large; Bakeman, 2005) are reported. Two-tailed pairwise comparisons were computed using generalized linear hypothesis testing for Tukey's Honestly Significant Difference contrasts, using the *glht* function in the *multcomp* library (Hothorn et al., 2008). Bayes factor was calculated using the *anovaBF* function in the *BayesFactor* library (Morey, Rouder, & Jamil, 2018). Partial correlations were calculated using the *pcorr.test* function in the *ppcor* library (Kim & Kim, 2015) and associated Bayes factor partial correlations were calculated using the *jzs_partcor* function in the *BayesMed* library (Nuijten et al., 2014).

2 Results

To evaluate the performance of ASR services between groups, accuracy scores were subjected to the nLMEM. There were significant main effects of Group [$F(2, 92) = 44.69$, $p < 0.001$, $\eta^2_G = 0.45$], ASR service [$F(2, 184) = 66.72$, $p < 0.001$, $\eta^2_G = 0.08$], and nGram [$F(2, 552) = 3888.11$, $p < 0.001$, $\eta^2_G = 0.39$]. Two-way interactions were significant between Group and ASR service [$F(4, 184) = 3.91$, $p = 0.005$, $\eta^2_G = 0.01$], Group and nGram [$F(4, 552) = 37.74$, $p < 0.001$, $\eta^2_G = 0.01$], and ASR service and nGram [$F(4, 552) = 2.73$, $p = 0.03$, $\eta^2_G = 0.001$]. The three-way interaction between Group, ASR service, and nGram was also significant [$F(8, 552) = 2.07$, $p = 0.04$, $\eta^2_G = 0.002$]. Planned comparisons for hypothesis testing were performed considering all pairwise comparisons of interest.

For all Groups and ASR services, accuracy significantly decreased as nGram increased ($ps < 0.001$; see Fig. 1a). These results suggest that the ASR services have greater difficulty transcribing consecutive words regardless of whether speech is impaired. For all nGrams and ASR services, accuracy was significantly higher for controls relative to MS ($ps < 0.03$) and FA ($ps < 0.001$), and for MS relative to FA ($ps < 0.001$). These results support the hypothesis that speech recognition accuracy decreased for groups with neurodegenerative disease. As speech recognition accuracy was lower for FA compared to MS, results also suggest that the severity of the disease type influenced accuracy. For all nGrams and Groups, Amazon Web Services and Google Cloud outperformed IBM Watson ($ps < 0.003$). Amazon Web Services and Google Cloud did not significantly differ except for FA one-word accuracy where Google Cloud showed higher accuracy than Amazon Web Services ($p = 0.01$). A Bayes factor *t*-test between accuracy from Amazon Web Services and Google Cloud revealed strong evidence for the null hypothesis ($BF_{01} = 10.49 \pm 0.001\%$) suggesting that these two ASR services performed comparably.

To explore differences in ASR between females and male speakers, we performed the same analyses as above with Sex as factor. Results revealed a significant two-way interaction between ASR service and Sex [$F(2, 178) = 3.36$, $p = 0.04$, $\eta^2_G = 0.004$], and a significant three-way interaction between Group, ASR service, and Sex [$F(4, 178) = 3.25$, $p = 0.01$, $\eta^2_G = 0.008$]. Post hoc comparisons examining the two-way interaction between ASR service and Sex revealed that sex did not moderate differences between ASR services with Amazon Web Services and Google Cloud demonstrating higher accuracy than IBM Watson for males ($ps < 0.001$) and females ($ps < 0.001$; see Fig. 1b). Males and females did not significantly differ for any of the ASR services ($ps > 0.95$). Post hoc comparisons examining the three-way interaction between Group, ASR service, and Sex revealed that accuracy for males was significantly higher than females in the HC group for Amazon Web Services ($p = 0.03$) and trended towards significance for the HC group for IBM Watson ($p = 0.07$). No other sex differences approached significance ($ps > 0.1$). Finally, the same significant differences between groups (HC > MS > FA) were observed for both males ($ps < 0.003$) and females ($ps < 0.003$). Overall, these results show that sex did not moderate differences between groups or ASR services.

To explore the influences of age and disease duration, we examined correlations between accuracy and age across groups and within each group. As shown in Fig. 2, there were moderate significant negative correlations between accuracy and age for FA using all ASR services. No other correlations between age and accuracy were significant. For disease duration, the FA group demonstrated moderate significant negative correlations with

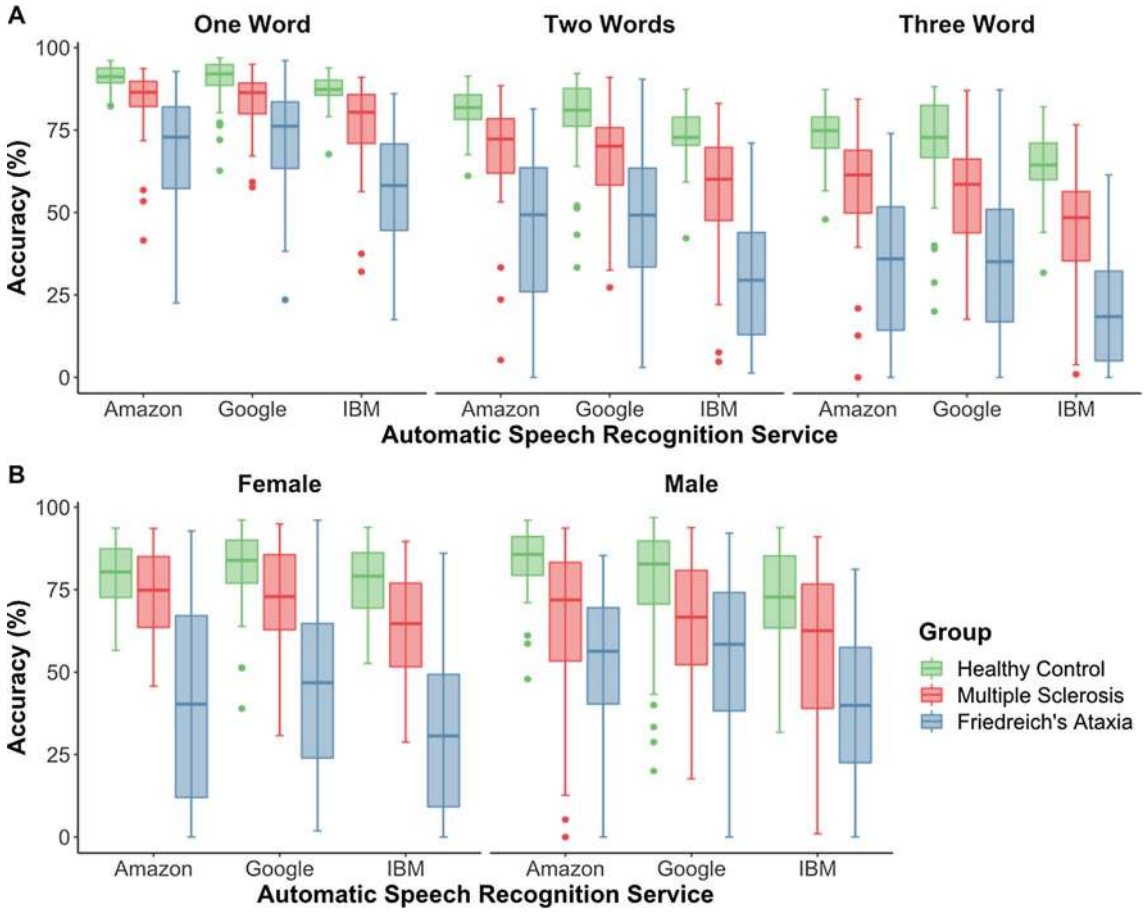

Fig. 1 Mean accuracy for automatic speech recognition methods for groups for **a** nGrams consisting of one, two, and three consecutive words, and **b** between females and males

accuracy for all ASR services. No significant correlations were observed between ASR accuracy and disease duration for any ASR service for the MS group ($ps > 0.2$; see Fig. 2c). These results indicate that speech recognition accuracy decreases as FA progresses over time, but MS did not demonstrate this relationship. Finally, we performed partial correlations between accuracy and age controlling for disease duration. No partial correlations were significant ($ps > 0.3$). To assess evidence for the null hypothesis we performed Bayes factor partial correlations, which revealed strong evidence for the null hypothesis for FA ($BF_{01} = 22.64 \pm 0.001\%$) and moderate evidence for MS ($BF_{01} = 8.98 \pm 0.001\%$). Bayes factor correlations between accuracy and age for healthy controls also demonstrated strong evidence for the null hypothesis ($BF_{01} = 16.93 \pm 0.4\%$). These results suggest that ASR accuracy is not influenced by healthy ageing outside of the effects of neurodegenerative disease.

To examine the relationship between ASR accuracy and confidence, we performed Pearson correlations between accuracy at the one-word level and confidence. As shown in Table 2, there were significant moderate-to-large positive correlations for all ASR services and all groups ($ps < 0.001$). These results suggest that word-level confidence is an indicator of ASR accuracy regardless of ASR service or neurodegenerative disease.

3 Discussion

We examined the accuracy of three publicly available state-of-the-art automatic speech recognition platforms for individuals with impaired speech. Accuracy for was higher for healthy controls compared to people with neurodegenerative disease, and higher accuracy for people with MS compared to FA. Accuracy declined with longer disease duration, suggesting that dysarthria severity negatively

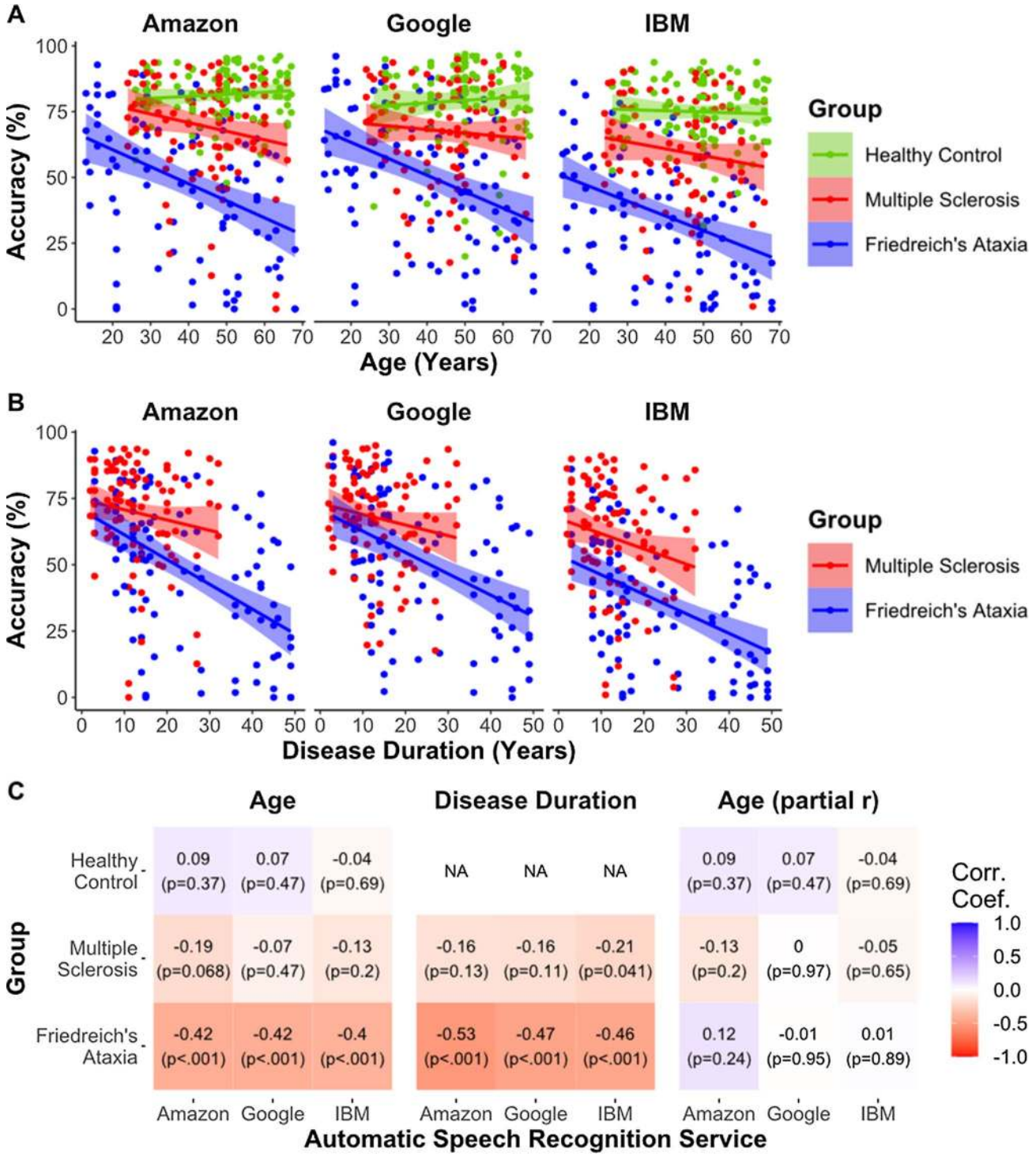

Fig. 2 Scatterplots **a** between accuracy and age, **b** between accuracy and disease duration, and **c** correlation coefficients between ASR accuracy and age, ASR accuracy and disease duration, and partial correlations between accuracy and age controlling for disease duration

impacts the capability of ASR services to interpret speech in neurodegenerative disease. Healthy ageing did not appear to affect accuracy. Performance varied across the three ASR services with Amazon Web Services and Google Cloud yielding higher accuracy than IBM Watson. Amazon Web Services provided higher accuracy for males than females in the healthy control group. Overall, findings suggest that these ASR services are not yet optimized

Table 2 Pearson correlation coefficients between one-word accuracy and ASR confidence

	Amazon	Google	Watson
Healthy Controls	0.82	0.88	0.86
Multiple Sclerosis	0.58	0.52	0.62
Friedreich's Ataxia	0.94	0.64	0.81

All $ps < 0.001$

for populations with mild (e.g., MS), moderate, or severe (e.g., FA) speech impairments associated with neurodegenerative disease.

People with neurodegenerative disease stand to benefit significantly from ASR. Tools designed to automate tasks and enhance communication may improve quality of life and provide access to activities that are inaccessible. ASR has been proposed as an alternative mode of communication for people with difficulty typing due to fine motor impairment, as they could use ASR to write emails or use instant messaging services. ASR would be particularly useful in telehealth applications where communication may be degraded further because of unreliable internet connections or poor bandwidth. There are several techniques that could be used to optimize ASR for people with neurodegenerative disease and other speech impairments. One approach would be to train the ASR models on speech data from people with neurodegenerative disease. These models could then either be selected as a language variant (e.g., English, Neurodegenerative disease) or be implemented automatically if the ASR service is unable to determine words with high confidence. Our results demonstrating moderate correlations between ASR accuracy and ASR confidence suggest that automatically switching the language variant for low-confidence transcripts could be a viable option to increase the accessibility of ASR services. Several projects are underway that aim to improve the accuracy and capacity of ASR in speakers with compromised speech (e.g., *Project Euphonia by Google AI*, n.d.) with the outcomes pending. Another approach would be to continue adapting existing ASR models idiosyncratically on individuals' speech by allowing some form of error correction either manually or through automatic error estimation. Personalization is more arduous for the user but may lead to better ASR accuracy with time and consistent usage than static non-adaptive models trained at the population level. Several commercially available ASR services train models for individuals, such as Dragon (Nuance, 2020), but do not appear to have been systematically tested on populations with impaired speech. A comparison of population- and individual-level ASR services would be useful to ascertain which produces better ASR accuracy for healthy controls and people with impaired speech. It is possible that hybrid ASR implementation would produce better results; a base model trained on population data which is then optimized using individual-level adaptations should, theoretically, outperform population- or individual-level ASR approaches.

Some evidence suggests a negative relationship between age and ASR accuracy (cf. Young & Mihailidis, 2010). In contrast, our results did not demonstrate a relationship between age and ASR accuracy for healthy controls. ASR accuracy for people with neurodegenerative disease was primarily associated with disease duration, a proxy of disease severity. It is possible that previous research did not consider the prevalence of neurodegenerative disease or speech impairments in their cohorts and that other factors mediated relationships between age and ASR accuracy. Our partial correlations between ASR accuracy and age that accounted for disease duration were not significant, suggesting that healthy ageing does not negatively impact ASR accuracy in our cohort. This suggests that ASR remains a useful communication tool for older populations without impaired speech.

We examined speech from healthy controls, MS, and FA to provide coverage across subtle to severe speech impairments, representing a wide spectrum of speech intelligibility. It is likely that other conditions with more severe influence on speech may produce even lower ASR accuracy. Future studies could examine how other neurodegenerative diseases, and speech disorders more generally, affect ASR accuracy. It could be the case that certain conditions manifest as qualitatively different speech impairments. Understanding how different conditions affect ASR accuracy would aid in deciding the best strategy for optimizing ASR, and assessing which groups require different models when interpreting speech. In turn, this would increase the accessibility of ASR for a broad range of people with speech conditions who stand to benefit from communication aids and voice-activated technologies.

3.1 Limitations and considerations

We examined Australian English speakers and used the AU English language models where available, that is, for Amazon Web Services and Google Cloud but not IBM Watson. It is possible that an AU English model would produce higher accuracy for IBM Watson. Future research could further examine the cross-linguistic effects of neurodegenerative disease on ASR accuracy in other languages, dialects, and accents (cf. Pinto et al., 2017). Neurodegenerative disease may affect tonal languages, such as Cantonese, differently than non-tonal languages (e.g., Wong & Diehl, 1999). How language-specific constraints in dysarthria might influence ASR accuracy remains an open question.

We examined the ASR services that were free, able to transcribe speech into text, could be implemented in Python, and are considered state-of-the-art due to the large corpus of data accumulated by these companies. A variety of other

state-of-the-art ASR services are also available, for example, Dragon (Nuance, 2020), Siri (Apple, 2020), and Cortana (Microsoft, 2020) but these either required software purchases or were bundled with operating systems. Future studies could compare ASR accuracy of these ASR services in the event the APIs become less constrained by the owning companies.

Finally, the lower ASR accuracy shown for people with neurodegenerative disease when using current models points to alternative uses of ASR, such as useful automatic measures for speech intelligibility (e.g., Fontan et al., 2017; Mengistu & Rudzicz, 2011; Schädler et al., 2015). These measures may be able to identify dysarthria within the population and prompt individuals to seek clinical attention (Fontan et al., 2017). ASR technology thus has the potential to benefit society as a virtue of difficulties with recognizing dysarthric speech. It would be prudent to retain legacy ASR models to ensure alternative applications of ASR services can be utilized.

4 Conclusion

Our work highlights the complexities of ASR technology in populations outside of traditional "healthy speaker" models. This poses a challenge for voice assistance devices and speech transcription services, which hold promise for improving access to services for people with physical and/or communication impairments. These services could be optimized for people with dysarthria by training ASR classifiers on large datasets of impaired speech and creating language models that are sensitive to dysarthric speech. Once ASR technology has been optimized for speech from people with neurodegenerative disease, it will have potential to improve communication and benefit society through robust information and communication technology infrastructures. With the demand for telehealth applications increasing due to the recent COVID19 pandemic, optimizing ASR technology for at-risk populations (e.g., people with neurodegenerative disease) should be considered a high priority. Therefore, we advocate that ASR should be improved for people with neurodegenerative disease to facilitate communication.

Appendix

The Grandfather Passage

You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day, he plays skillfully and with zest upon a small organ. Except in the winter, when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

Acknowledgements Authors VSAT and DBG are supported by the ARC Industry Transformational Training Centre IC170100030.

Declarations

Conflict of interest APV is Chief Science Officer of Redenlab, a speech biomarker company.

References

Apple. (2020). *Siri for developers*. https://developer.apple.com/siri/.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379–384.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Blaney, B., & Wilson, J. (2000). Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistics and Phonetics, 14*(4), 307–327.

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

Delatycki, M. B., & Bidichandani, S. I. (2019). Friedreich ataxia-pathogenesis and implications for therapies. *Neurobiology of Disease, 132*, 104606.

Delatycki, M. B., Williamson, R., & Forrest, S. M. (2000). Friedreich ataxia: An overview. *Journal of Medical Genetics, 37*(1), 1–8.

Folker, J., Murdoch, B., Cahill, L., Delatycki, M., Corben, L., & Vogel, A. (2010). Dysarthria in Friedreich's ataxia: A perceptual analysis. *Folia Phoniatrica et Logopaedica*. https://doi.org/10.1159/000287207.

Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., Gaillard, P., Aumont, X., & Füllgrabe, C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research, 60*(9), 2394–2405.

Gibilisco, P., & Vogel, A. P. (2013). Friedreich ataxia. *BMJ, 347*, f7062.

Harding, A. E. (1983). Classification of the hereditary ataxias and paraplegias. *The Lancet, 321*(8334), 1151–1155.

Hothorn, T., Bretz, F., Westfall, P., & Heiberger, R. M. (2008). Multcomp: Simultaneous inference for general linear hypotheses. R Package Version, 0-1.
Jeffreys, H. (1998). *The theory of probability*. OUP.
Kim, S., & Kim, M. S. (2015). Package 'ppcor.' *Communications for Statistical Applications and Methods, 22*(6), 665–674.
Klockgether, T. (2007). Ataxias. *Parkinsonism and Related Disorders, 13*, S391–S394.
Liégeois, F., Morgan, A. T., Stewart, L. H., Cross, J. H., Vogel, A. P., & Vargha-Khadem, F. (2010). Speech and oral motor profile after childhood hemispherectomy. *Brain and Language, 114*(2), 126–134.
Magee, M., Copland, D., & Vogel, A. P. (2019). Motor speech and non-motor language endophenotypes of Parkinson's disease. *Expert Review of Neurotherapeutics, 19*(12), 1191–1200.
MathWorks. (2019). *MATLAB* (9.6.0 (2019b)). The MathWorks Inc.
Mazzoni, D., & Dannenberg, R. (2012). *Audacity® 2.0.0*. Audacity Team.
Mengistu, K. T., & Rudzicz, F. (2011). Comparing humans and automatic speech recognition systems in recognizing dysarthric speech. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. https://doi.org/10.1007/978-3-642-21043-3-36
Microsoft. (2020). *Cortana: Your personal productivity assistant in Microsoft 365*. https://www.microsoft.com/en-us/cortana.
Morey, R. D., Rouder, J. N., & Jamil, T. (2018). *BayesFactor: Computation of Bayes Factors for common designs. R package version 0.9.12-4.2*. https://CRAN.R-project. org/package=BayesFactor. Cited June 30, 2018.
Noffs, G., Boonstra, F. M. C., Perera, T., Kolbe, S. C., Stankovich, J., Butzkueven, H., Evans, A., Vogel, A. P., & van der Walt, A. (2020). Acoustic speech analytics are predictive of cerebellar dysfunction in multiple sclerosis. *The Cerebellum, 19*(5), 1–10.
Noffs, G., Perera, T., Kolbe, S. C., Shanahan, C. J., Boonstra, F. M. C., Evans, A., Butzkueven, H., van der Walt, A., & Vogel, A. P. (2018). What speech can tell us: A systematic review of dysarthria characteristics in Multiple Sclerosis. *Autoimmunity Reviews, 17*(12), 1202–1209.
Nuance. (2020). *Dragon Naturally Speaking software*. https://www.nuance.com/en-au/dragon/support/dragon-naturallyspeaking.html.
Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E. J. (2014). *BayesMed: Default Bayesian hypothesis tests for correlation, partial correlation, and mediation (R package version 1.0.0.)*.
Owens, J. S. (2006). Accessible information for people with complex communication needs. *Augmentative and Alternative Communication, 22*(3), 196–208.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2015). nlme: Linear and nonlinear mixed effects models. R package version 3.1-120. R Package Version, 1-3.
Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R., & Sadat, J. (2017). A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. *Journal of Phonetics, 64*, 156–167.
Poole, M. L., Brodtmann, A., Darby, D., & Vogel, A. P. (2017). Motor speech phenotypes of frontotemporal dementia, primary progressive aphasia, and progressive apraxia of speech. *Journal of Speech, Language, and Hearing Research, 60*(4), 897–911.
Poole, M. L., Wee, J. S., Folker, J. E., Corben, L. A., Delatycki, M. B., & Vogel, A. P. (2015). Nasality in Friedreich ataxia. *Clinical Linguistics and Phonetics, 29*(1), 46–58.
Project Euphonia by Google AI. (n.d.).
R Core Team. (2013). *R: A language and environment for statistical computing*. R Core Team.
Raghavendra, P., Rosengren, E., & Hunnicutt, S. (2001). An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication, 17*(4), 265–275.
Rojas, S., Kefalianos, E., & Vogel, A. (2020). How does our voice change as we age? A systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of Speech, Language, and Hearing Research, 63*(2), 533–551.
Rosen, K. M., Folker, J. E., Vogel, A. P., Corben, L. A., Murdoch, B. E., & Delatycki, M. B. (2012). Longitudinal change in dysarthria associated with Friedreich ataxia: A potential clinical endpoint. *Journal of Neurology, 259*(11), 2471–2477.
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.
Rossum, G. V. (2019). *Python Language Reference, version 3*. Python Software Foundation.
Schädler, M. R., Warzybok, A., Hochmuth, S., & Kollmeier, B. (2015). Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology, 54*(sup2), 100–107.
Shih, W. (2020). Voice revolution. *Library Technology Reports, 56*(4), 5–13.
Stoppler, M. C. (2019). *Multiple sclerosis symptoms, causes, treatment, diagnosis, and life expectancy*. Emedicinehealth. https://www.emedicinehealth.com/multiple_sclerosis/article_em.htm.
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.
Van Riper, C. (1963). *Speech correction principles and methods* (Vol. 7, pp. 176–177). Prentice Hall.
Vogel, A. P., Wardrop, M. I., Folker, J. E., Synofzik, M., Corben, L. A., Delatycki, M. B., & Awan, S. N. (2017). Voice in Friedreich ataxia. *Journal of Voice, 31*(2), 243.e9-243.e19. https://doi.org/10.1016/j.jvoice.2016.04.015.
Wallin, M. T., Culpepper, W. J., Nichols, E., Bhutta, Z. A., Gebrehiwot, T. T., Hay, S. I., Khalil, I. A., Krohn, K. J., Liang, X., & Naghavi, M. (2019). Global, regional, and national burden of multiple sclerosis 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology, 18*(3), 269–285.
Wong, P. C. M., & Diehl, R. L. (1999). The effect of reduced tonal space in Parkinsonian speech on the perception of Cantonese tones. *Journal of the Acoustical Society of America, 105*(2 Pt 2), 1246.
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*. https://doi.org/10.1080/10400435.2010.483646.
Zhang, A. (2017). *Speech recognition (version 3.8)*. May.