INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Automatic Speech Recognition of Disordered Speech: Personalized models outperforming human listeners on short phrases

Jordan R. Green[1,5], Robert L. MacDonald[2], Pan-Pan Jiang[2], Julie Cattiau[2], Rus Heywood[2], Richard Cave[3], Katie Seaver[1], Marilyn A. Ladewig[4], Jimmy Tobin[2], Michael P. Brenner[2,5], Philip C. Nelson[2], Katrin Tomanek[2]

[1] MGH Institute of Health Professions, USA
[2] Google LLC, USA
[3] MND Association, UK
[4] Cerebral Palsy Associations of New York State, USA
[5] Harvard University, USA

jgreen2@mghihp.edu, katrintomanek@google.com

Abstract

This study evaluated the accuracy of personalized automatic speech recognition (ASR) for recognizing disordered speech from a large cohort of individuals with a wide range of underlying etiologies using an open vocabulary. The performance of these models was benchmarked relative to that of expert human transcribers and two different speaker-independent ASR models trained on typical speech. 432 individuals with self-reported disordered speech recorded at least 300 short phrases using a web-based application. Word error rates (WERs) were estimated for three different ASR models and for human transcribers. Metadata were collected to evaluate the potential impact of participants, atypical speech characteristics, and technical factors on recognition accuracy. Personalized models outperformed human transcribers with median and max recognition accuracy gains of 9% and 80%, respectively. The accuracies of personalized models were high (median WER: 4.6%) and better than those of speaker-independent models (median WER: 31%). The most significant improvements were for the most severely affected speakers. Low signal-to-noise ratio and fewer training utterances were associated with poor word recognition, even for speakers with mild speech impairments. Our results demonstrate the efficacy of personalized ASR models in recognizing a wide range of speech impairments and severities and using an open vocabulary.

Index Terms: speech recognition, speech disorders, personalized models, automatic speech recognition

1. Introduction

Speech impairments affect millions of individuals, limiting self-expression, vocational opportunities, and access to new technologies. Automatic speech recognition (ASR) could open doors to independence for these individuals by improving access to dictation and home automation, and ultimately facilitating more fluid real-time conversations through machine translation. ASR accuracy has improved significantly over recent years owing to the increased computational power of deep learning systems and the availability of large training datasets. Accuracy benchmarks for commercial systems are now as high as 95% for many speakers and applications [1,2].

Despite these improvements, however, disordered speech recognition is still unacceptably low, rendering the technology unusable for speakers who could benefit the most. Word error rates (WER) for three commercially available ASR systems (IBM Watson, Google Cloud, and Microsoft Azure Bing) for individuals with severely dysarthric speech are high (range: 78-89%), resulting in a low percentage (range: 0-1.2%) of correctly transcribed sentences [3]. The poor recognition is partly due to the diversity and complexity of atypical speech patterns [4] and insufficient training data [5]. To navigate these challenges, researchers have largely focused on small vocabulary recognition from small cohorts of individuals [6,7] using personalized models rather than more generalizable speaker-independent models [8–11].

Benchmarking the state-of-the-art accuracy of disordered speech recognition is, however, difficult because of differences across studies in the (a) settings and devices used to record speech (e.g., a laboratory setting using a headset microphone vs. a cell phone in a noisy home environment), (b) demands of the speaking tasks (spontaneous speech vs. reading, or complete sentences vs. single words), and (c) participants' characteristics in terms of underlying etiologies, speech impairment type, and speech impairment severity. Reporting a high accuracy rate, for example, is less impressive if the participant cohort only includes mildly affected talkers rather than more severely affected talkers [12].

In this paper, we report ASR accuracy from a diverse cohort of speakers with disordered speech in terms of etiology, severity, and atypical speech characteristics. Speech recordings were obtained under conditions that were consistent with the intended use-case - in each user's home environment using their native hardware. Personalized ASR models were evaluated by comparing their recognition accuracies with those of ASR baseline models developed for typical speech and expert human transcribers. Comparing the performances of personalized ASR models with those of human listeners provides a benchmark for evaluating ASR's potential to repair or assist human communication [13]. We also consider the characteristics of the speaker and their speech and technical factors that may degrade recognition accuracy, such as recording quality and training set size. Understanding why ASR works better for some speakers than others will aid ongoing efforts to maximize ASR performance across a broadening range of speech disorders [14,15].

http://dx.doi.org/10.21437/Interspeech.2021-1384

2. Methods

2.1. Participants and speech recordings

Our participant cohort included 432 English speakers with various etiologies, speech impairment types, and speech impairment severities. All of our participants were 18 years or older. Data were collected only after explicit consent had been granted by individuals to provide speech samples for the purpose of research and improving speech recognition products and services. We provided users with clear information pertaining to the purpose of the data collection and scope of research. The speech recordings were obtained remotely using a web-based platform and the participants' native environment and hardware, which could be a cell phone, tablet, or computer, with or without an external microphone. Details of data collection are described in a companion paper by MacDonald et al. [16]. Participants read phrases randomly from a screen prompt. The phrases came from different target domains, including home automation, typical phrases used in communication with a caregiver, and longer conversational phrases.

2.2. Metadata

Participant metadata. Participant metadata are presented in Table 1. Information pertaining to participants' age and sex was not collected owing to privacy reasons, whereas that pertaining to etiology, although unsolicited, was provided by a majority of the participants or else inferred by a licensed speech-language pathologist (SLP). Missing data included instances when labels were difficult to determine or were unavailable.

Speech features metadata. The SLPs graded each participant's speech along ten different speech, voice, and resonance features using a 5-point Likert scale (typical, mild, moderate, severe, and profound). The features included overall severity, intelligibility, fastness, slowness, prosody, consistency, articulatory, phonatory, resonatory, and respiratory. Owing to the small number of participants in the profound speech severity group (n = 5), we combined these subjects' data with that of the severe severity group.

Technical factors metadata. For each participant, we recorded the number of utterances used to train their personalized model and an algorithmic estimate of overall recording quality as indexed by the signal-to-noise ratio (SNR) [17].

2.3. Data sets

Full dataset. The full dataset contained recordings from 432 speakers, all of whom recorded 300 or more utterances. The median number of training utterances was 1530 (min = 304, max = 9830). We split our datasets into training (80%), dev (10%), and test (10%) phrases. Splitting was done at the phrase level, such that each recording of a given phrase was always added to the same split (train, test, or dev) across all speakers. The full dataset was divided into several subsets to address specific experimental questions as described below.

High & Low WER subsets. To identify the factors associated with poor ASR performance, we created a subset of participant data with low and high personalized model WERs based on the 1st and 5th quintiles of the WER distribution, respectively. Speech and technical metadata were compared by calculating the between-group effect sizes (Cohen's D) for each metadata label.

Surprisingly High WER subset (SurpHigh)**.** The High WER subset was further filtered to include only participants with typical speech or mild speech severity labels within the High WER group. These data were used to determine whether specific speech labels or technical factors increased the WER.

Human transcription WER subset. To compare the accuracies of personalized models with those of human listeners, we created a subset that included only the participants (n = 116) for which human transcriptions were obtained. Participants in this subset represented the full range of speech severity. These word-level orthographic transcriptions were performed by experienced listeners (3 SLPs with experience serving persons with dysarthria) and obtained for 30 randomly selected phrases.

Training set sizes for the ASR models were comparable across all severity groups (range across severities = 60 utterances). The median test set size for the ASR models was 75 utterances.

Table 1: *Speaker characteristics with participant count (and percentage) on the full dataset and three subsets of participants: (1) Low WER, (2) High WER, and (3) SurpHigh WER subset.*

	Overall (N=432)	High WER (N=78)	Low WER (N=79)	SurpHigh (N=17)
Sex				
FEMALE	157 (36.3%)	22 (28.2%)	37 (46.8%)	7 (41.2%)
MALE	266 (61.6%)	56 (71.8%)	42 (53.2%)	10 (58.8%)
Missing	9 (2.1%)	0 (0%)	0 (0%)	0 (0%)
Etiology				
AMYOTROPHIC LATERAL SCLEROSIS	172 (39.8%)	24 (30.8%)	41 (51.9%)	6 (35.3%)
PARKINSONS DISEASE	66 (15.3%)	15 (19.2%)	10 (12.7%)	4 (23.5%)
CEREBRAL PALSY	53 (12.3%)	19 (24.4%)	3 (3.8%)	0 (0%)
ATAXIA	21 (4.9%)	5 (6.4%)	4 (5.1%)	2 (11.8%)
OTHER	18 (4.2%)	2 (2.6%)	5 (6.3%)	1 (5.9%)
HEARING IMPAIRMENT	17 (3.9%)	3 (3.8%)	3 (3.8%)	1 (5.9%)
MUSCULAR DYSTROPHY	14 (3.2%)	0 (0%)	3 (3.8%)	0 (0%)
MULTIPLE SCLEROSIS	9 (2.1%)	2 (2.6%)	0 (0%)	0 (0%)
STROKE	8 (1.9%)	2 (2.6%)	1 (1.3%)	0 (0%)
TRAMATIC BRAIN INJURY	5 (1.2%)	0 (0%)	1 (1.3%)	0 (0%)
VOCAL FOLD PARALYSIS	4 (0.9%)	0 (0%)	2 (2.5%)	0 (0%)
CLEFT LIP and CLEFTPALATE	3 (0.7%)	0 (0%)	0 (0%)	0 (0%)
SPINAL MUSCULAR ATROPHY	3 (0.7%)	0 (0%)	1 (1.3%)	0 (0%)
MULTIPLE SYSTEMS ATROPHY	1 (0.2%)	0 (0%)	0 (0%)	0 (0%)
Missing	38 (8.8%)	6 (7.7%)	5 (6.3%)	3 (17.6%)
Speech Disorder				
DYSARTHRIA	294 (68.1%)	63 (80.8%)	56 (70.9%)	10 (58.8%)
WNL	42 (9.7%)	3 (3.8%)	13 (16.5%)	3 (17.6%)
DYSPHONIA	21 (4.9%)	4 (5.1%)	3 (3.8%)	1 (5.9%)
DEAF	13 (3.0%)	2 (2.6%)	3 (3.8%)	1 (5.9%)
SPEECH SOUND DISORDER	7 (1.6%)	1 (1.3%)	0 (0%)	0 (0%)
APRAXIA	6 (1.4%)	0 (0%)	0 (0%)	0 (0%)
CLEFT LIP and PALATE	3 (0.7%)	0 (0%)	0 (0%)	0 (0%)
APRAXIA & DYSARTHRIA	2 (0.5%)	2 (2.6%)	0 (0%)	0 (0%)
OTHER	2 (0.5%)	0 (0%)	1 (1.3%)	0 (0%)
UNKOWN	2 (0.5%)	1 (1.3%)	0 (0%)	1 (5.9%)
HYPERNASALITY	1 (0.2%)	0 (0%)	0 (0%)	0 (0%)
Missing	39 (9.0%)	2 (2.6%)	3 (3.8%)	1 (5.9%)
Speech Severity				
TYPICAL	71 (16.4%)	16 (20.5%)	18 (22.8%)	0 (0%)
MILD	148 (34.3%)	5 (6.4%)	18 (22.8%)	5 (29.4%)
MODERATE	104 (24.1%)	12 (15.4%)	40 (50.6%)	12 (70.6%)
SEVERE	100 (23.1%)	45 (57.7%)	3 (3.8%)	0 (0%)
Missing	9 (2.1%)	0 (0%)	0 (0%)	0 (0%)

2.4. Automatic speech recognition

We evaluated the recognition accuracies of two speaker-independent ASR systems trained and optimized for *typical* speech, and on ASR models personalized to the speech of our 432 participants with disordered speech.

Speaker-independent ASR models. The first speaker-independent ASR model (*SI-1*) was accessed via Google's Speech-to-Text API[1], a commercial, highly accurate, server-

[1] https://cloud.google.com/speech-to-text

based ASR solution. Google's Speech-to-Text is readily available and integrated into third-party products. The second speaker-independent model (*SI-2*) was an end-to-end ASR model based on the well-studied RNN-T architecture [18].

Our encoder network consists of 8 layers and the predictor network consists of 2 layers of uni-directional LSTM cells. Inputs were 80-dimensional log-mel filterbank energies. Outputs were probability distributions over a 4k word piece model vocabulary. Following the procedure described in [19], training was performed using ~162,000 h of typical speech (from Google's internal production dataset) and techniques designed to (a) make the resulting model more robust across various application domains and acoustic scenarios, and (b) generalize well to unseen conditions. SI-1 allowed us to benchmark the personalized model performance relative to a publicly available ASR system. SI-2 allowed us to benchmark the personalized model performance while controlling for model architecture, which was identical to that of the personalized model.

Personalized ASR models. For each of the 432 participants with disordered speech, we create a personalized ASR model (SI-2) from their own recordings. Our fine-tuning procedure was optimized for our adaptation process, where we only have between ¼ and 2 h of data per speaker. We found that updating only the first five encoder layers (versus the complete model) worked best and successfully prevented overfitting [10]. We applied SpecAugment [20] as a regularization method to increase robustness using parameters optimized for dysarthric speech. Specifically, SpecAugment worked best when we reduced the frequency masking and massively increased the time masking settings (as compared with the defaults found in typical speech). These effects are likely explained by the slowness and lower spectral diversity of the disordered speech. We used the Adam optimizer with a low learning rate ($1e^{-5}$).

3. Results

All WERs were calculated on the home automation test sets, with phrases typically containing 2-4 words. WERs were first calculated on a per-speaker level (on each speaker's test set) and then aggregated (median) across speakers.

3.1. Personalized ASR models significantly outperformed speaker-independent models and accurately recognized diverse types of speech impairments.

The overall model performance on disordered speech was measured on the full data set, i.e., across all 432 speakers. As displayed in Figure 1, the two speaker-independent ASR models (SI-1 and SI-2) perform very similarly (correlation coefficient: 0.96) with median WERs of 31.5% and 29.4%, respectively. In contrast, WERs were substantially lower for the personalized ASR models with a median WER of 4.6%. This trend was apparent for all speech severity groups ($p > 0.001$, for all pairwise comparisons). Although the median WER of the personalized ASR models was very low, (a) 30 of the 432 cases (7%) had a WER greater than 24%, which was the statistical outlier threshold based on the 3rd quartile + 1.5 * interquartile range criteria, and (b) approximately 113 of the cases (26%) exceeded the more conservative target WER of 10%, a WER that is likely to result in more wide-spread acceptance among users. Additional analyses were conducted to identify the speaker, speech, and technical factors that negatively affected these individuals' ASR performance. WERs are shown as a function of severity in the bottom panel of Figure 1. The WERs of both speaker-independent ASR models increased significantly with severity ($p > 0.001$). In contrast, the WERs of the personalized ASR models were similar for the normal, mild, and moderate groups (pooled median = 3.8%) but were significantly greater (median = 13%) for the severe group ($p > 0.001$).

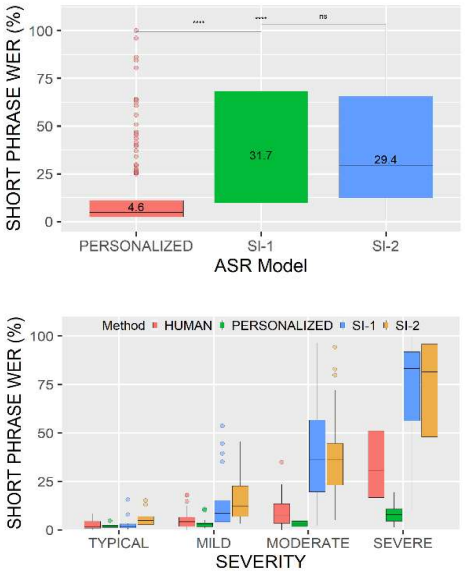

Figure 1: *Top panel: WER for each model. Bottom panel: WER as a function of severity for each model. **** = p<.001*

3.2. Speech characteristics, low SNR, and fewer training utterances adversely affected personalized ASR models

The participant metadata for the Low (1st quintile) and High WER (5th quintile) subsets are displayed in columns 2 and 3 in Table 1. Among the etiologies, cerebral palsy was disproportionately represented in the High WER group ($p < .05$). Severity was lower in the Low WER group than in the High WER group ($p < .01$). Figure 2 shows Cohen's d effect sizes for the speech and technical metadata. The SurpHigh WER group had fewer training utterances and lower SNR compared with the Low WER group ($p < .01$) resulting in large (negative) effect sizes, with all other factors having small effect sizes except fastness. However, the High WER group exhibited medium to large differences across all factors.

3.3. Personalized ASR models outperformed human speech recognition

To compare ASR with human listeners, we calculated machine and human transcription WERs from the human transcription WER subset. The top panel of Figure 3 shows the mean WER and cumulative probability plot for each method. The bottom panel shows the delta between the WERs of the personalized models and the human listeners (left pane) and the personalized models and SI-1 (right pane). Negative values indicate a lower WER for personalized ASR than for human listeners/speaker-independent ASR. The WERs were, on average, lower for personalized ASR models compared with both human transcribers and speaker-independent ASR, with the most notable gains in recognition accuracy for the participants with human WERs greater than 20 %.

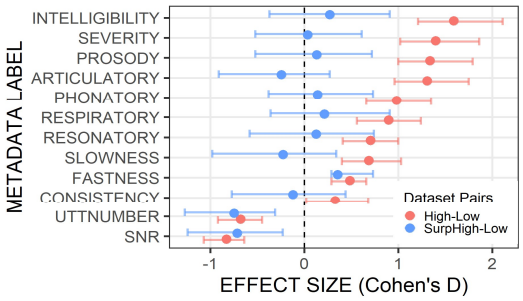

Figure 2: *Speech disorder and technical metadata effect sizes for the High WER - Low WER subset and SuprHigh-Low WER subset. Positive effects indicated that the group values of the High WER group were greater than Low WER groups'.*

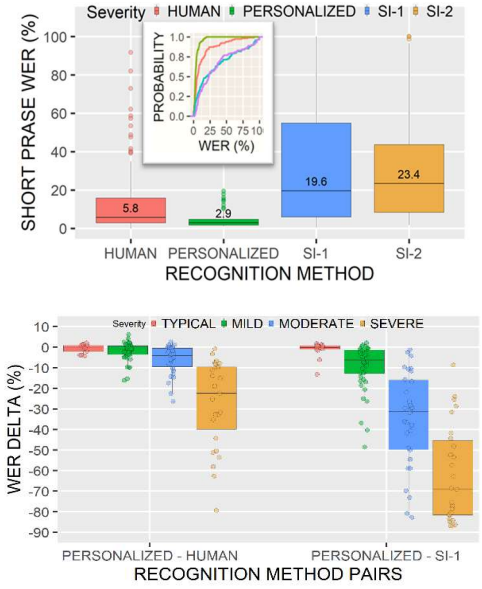

Figure 3: *Top panel: Differences in WER between personalized model vs human transcription and speaker-independent ASR (SI-1). Bottom panel: Differences in WER by severity*

4. Discussion and Conclusions

Our results demonstrate the efficacy of personalized ASR models for recognizing a wide range of speech impairments and severities using an open vocabulary. The accuracy of these models was excellent (i.e., median WER of 4.8%) for most speakers and similar or better than those of expert human listeners. Although personalized WERs were highest for the severe speech impairment group, they were still relatively low at a median of 13%. The relative gains in recognition accuracy of moderately and severely disordered speech afforded by the personalized models were as high as 80%, with a median of 9%, as compared to the human transcriptions. The median gains in accuracy afforded by the personalized ASR models was 49%, compared with those of the speaker-independent models for moderately and severely disordered speech. The inadequacy of commercially available ASR systems for recognizing severely disordered speech is well documented [3,21,22]. Although our findings are encouraging for users of domains that require only short phrases (e.g., home automation), additional research is needed to evaluate ASR performance during more challenging speaking conditions, such as during longer utterances or spontaneous conversations. A follow-up analysis conducted by our group on longer conversational phrases, yielded a relative median WER improvement of 60–77% for participants with mild, moderate, and severe speech impairments. A more comprehensive analysis of this data will be presented in future work.

Despite our encouraging results, WERs were greater than 10% for approximately 26% of the participants. For the speakers with typical to mild severity in this group, recognition accuracies were degraded primarily by technical factors (i.e., smaller training set size and lower SNR) rather than atypical speech features. In contrast, for speakers with moderate to severe severity in this group, degraded performance was due to the combined influence of atypical speech features, a low SNR, and a smaller number of training utterances (see Figure 2). These findings underscore the need for more robust strategies to mitigate the impact of noise on disordered speech recognition [23]. Additional research is also needed to determine if the low SNR is due to extrinsic sources of noise rather than the noisy components of disordered speech, such as aperiodic phonation [24,25] or hypernasality [26].

Intelligibility and severity were the most prominent atypical speech features in the High WER group; however, other speech features also emerged, including abnormal prosody, articulation, and phonation. These speech features are known to degrade overall speech intelligibility [27,28].

Different speech disorder presentations are expected to impact ASR non-uniformly. The distribution of speech disorder types within the High WER group indicates that dysarthria due to cerebral palsy was particularly difficult to model. For these speakers, robust recognition may require a more diverse and larger training set, and perhaps one that contains multiple repetitions of weakly predicted utterances [29].

These results demonstrate the potential of personalized ASR models for improving access to technology and functional communication in persons with speech impairments. The accuracies of the personalized models were as good as or even better than expert human listeners. Despite the promising performance, additional research is needed to improve the robustness of personalized ASR when (a) speech is produced under more demanding conditions such as everyday functional communication, (b) training utterances are difficult to acquire, and (c) speech is characterized by low intelligibility and, possibly, high production variability. This research will be enhanced by continued efforts to obtain larger and more diverse atypical speech training datasets and understand the individual-level factors that degrade ASR.

5. Acknowledgements

We'd like to thank the 1000+ participants who shared their voice samples. We also thank advocacy groups who connected participants with Project Euphonia, including ALS-TDI, Team Gleason, CDSS, CureDuchenne, ALSA, LSVT Global, MND Association, and Cerebral Palsy Association of New York State. We also thank Dimitri Kanevsky and Aubrie Lee for substantial data contributions and guidance on collection processes. We thank Mungkol Sarin, Knot Pipatsrisawat, Alena Butryna, and Joshua Advincula for critical tooling support Brian Richburg, Marc Maffei, and Victoria Bolowsky for assistance with references. Jordan Green's work was partially supported by NIH (K24 DC016312).

6. References

[1] R. Stonjic, R. Taylor, M. Kardas, V. Kerkex, and L. Viaud, "Papers with code - speech recognition.," 2020. https://paperswithcode.com/task/speech-recognition (accessed Oct 5, 2020).

[2] G. Synnaeve, "WER are we?" //github.com/syhw/wer_are_we (accessed Oct 5, 2020).

[3] L. De Russis, F. Corno, L. De Russis, F. Corno, and P. Di Torino, "On the Impact of Dysarthric Speech on Contemporary ASR Cloud Platforms On the Impact of Dysarthric Speech on Contemporary ASR Cloud Platforms people without abnormality in their speech intelligibil-ity," *Springer*, vol. 5, no. 3, pp. 163–172, Sep. 2019.

[4] F. L. Darley, A. E. Aronson, and J. Brown, *Motor speech disorders*. Philadelphia: W.B. Saunders, 1971.

[5] K. Caves, S. Boemler, and B. Cope, "Development of an automatic recognizer for dysarthric speech," in *Proceedings of the RESNA Annual Conference*, Phoenix, AZ., 2007, p. n/a..

[6] S. Fager, D. Beukelman, S. K. Fager, D. R. Beukelman, T. Jakobs, and J.P. Hosom, "Evaluation of a Speech Recognition Prototype for Speakers with Moderate and Severe Dysarthria: A Preliminary Report," *Taylor Fr.*, vol. 26, no. 4, pp. 267–277, Dec. 2014..

[7] I. Calvo *et al.*, "Evaluation of an Automatic Speech Recognition Platform for Dysarthric Speech," *Folia Phoniatr. Logop.*, pp. 1–10, 2020.

[8] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A Comparative Study of Adaptive, Automatic recognition of Disordered Speech," in *13th Annual Conf. of the Int'l Speech Communication Association*, 2012, pp. 1776–1779..

[9] S. R. Shahamiri and S. S. Binti Salim, "Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," *Adv. Eng. Informatics*, vol. 28, no. 1, pp. 102–110, Jan. 2014.1.

[10] J. Shor *et al.*, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," *arXiv preprint*, 2019. http://arxiv.org/abs/1907.13511.

[11] M.J. Kim, J. Yoo, and Kim, H. "Dysarthric Speech Recognition Using Dysarthria-Severity-Dependent and Speaker-Adaptive Models," in *Proc. Interspeech*, *2013*, 25-29.

[12] S. K. Fager and J. M. Burnfield, "Speech Recognition for Environmental Control: Effect of Microphone Type, Dysarthria, and Severity on Recognition Results," *Assist. Technol.*, vol. 27, no. 4, pp. 199–207, Oct. 2015.

[13] A. Jacks, K. L. Haley, G. Bishop, and T. G. Harmon, "Automated Speech Recognition in Adult Stroke Survivors: Comparing Human and Computer Transcriptions," *Folia Phoniatr. Logop.*, vol. 71, no. 5–6, pp. 286–296, Oct. 2019..

[14] M. B. Mustafa, F. Rosdi, S. S. Salim, and M. U. Mughal, "Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker," *Expert Systems with Applic.*, vol. 42, no. 8, pp. 3924–3932, May 15, 2015.

[15] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *AAC Augment. Altern. Commun.*, vol. 16, no. 1, pp. 48–60, 2000.

[16] R. L. MacDonald, P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. Ladewig, J Tobin, M. P. Brenner, P. Q. Nelson, J. R. Green, and K. Tomanek, "Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia", in *Proc. Interspeech*, 2021.

[17] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, *2008*, pp 2598-2601.

[18] A. Graves, "Sequence Transduction with Recurrent Neural Networks,", arxiv.org/abs/1211.3711, 2021.

[19] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," *IEEE Spoken Language Technology Workshop*, pp. 441–447, 2018.

[20] D.S. Park, W. Chan,Y. Zhang, C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, 2613-2617.

[21] F. Ballati, F. Corno, and L. De Russis, "'Hey Siri, Do You Understand Me?': Virtual Assistants and Dysarthria," *Ambient Intell. Smart Environ.*, vol. 23, pp. 557–566, 2018.

[22] K. Hux, J. Rankin-Erickson, N. Manasse, and E. Lauritzen, "Accuracy of three speech recognition systems: Case study of dysarthric speech," *AAC Augment. Altern. Com.*., vol. 16, no. 3, pp. 186–196, 2000.

[23] W. K. Seong, J. H. Park, and H. K. Kim, "Reducing Speech Noise for Patients with Dysarthria in Noisy Environments," *IEICE Trans. Inf. Syst.*, vol. 97, no. 11, pp. 2881–2887, 2014.

[24] R. Chiaramonte and M. Vecchio, "A Systematic Review of Measures of Dysarthria Severity in Stroke Patients," *PM and R*, vol. 13, no. 3. John Wiley and Sons Inc, Mar. 01, 2020.

[25] L. A. Ramig, I. R. Titze, R. C. Scherer, and S. P. Ringel, "Acoustic analysis of voices of patients with neurologic disease: Rationale and preliminary data," *Ann. Otol. Rhinol. Laryngol.*, vol. 97, no. 2, pp. 164–172, 1988..

[26] M. Eshghi, B. Richburg, Y. Yunusova, and J. R. Green, "Instrumental Evaluation of Velopharyngeal Dysfunction in Amyotrophic Lateral Sclerosis," in *Int. Congr. of Phonetic Sciences ICPHS*, 2019, pp. 2996–3000.

[27] J. Lee, K. C. Hustad, and G. Weismer, "Predicting speech intelligibility with a multiple speech subsystems approach in children with cerebral palsy," *J. Speech, Lang. Hear. Res.*, vol. 57, no. 5, pp. 1666–1678, Oct. 2014.

[28] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, "Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behav. Neuro.*, vol. 2015, 2015.

[29] L. J. Ferrier, H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit, "Dysarthric Speakers Intelligibility and Speech Characteristics in Relation to Computer Speech Recognition," *Augment. Altern. Com.*, vol. 11, no. 3, pp. 165–175, 1995.