Linguistic and extralinguistic factors in automatic speech recognition of German atypical speech

Eugenia Rykova[1, 2, 3] **Mathias Walther**[2]
[1] University of Eastern Finland, Joensuu, Finland
[2] University of Applied Sciences TH Wildau, Wildau, Germany
[3] Catholic University Eichstätt-Ingolstadt, Eichstätt, Germany
eugenryk@uef.fi

Abstract

Automatic speech recognition (ASR) has been already used in speech and language therapy, including diagnostic tasks and practice exercises for people with aphasia (PWA). The lack of relevant data makes it difficult to evaluate the algorithms' suitability for German-speaking PWA. For the current project, four open-source ASR models were selected based on their performance on other types of atypical speech, and the details of their evaluation are presented in this paper. The four selected models are generally robust to speakers' gender and age. The one-word recognition yields better results for words of moderate length. Speech rate should be neither too slow nor too quick for lower error rates both in words and phrases, and the latter should be also of moderate length.

1 Introduction

Automatic speech recognition (ASR) has become part of many everyday services, including digital health. In particular, speech and language therapy (SLT) can benefit considerably from ASR usage – for example, when in-person therapy is supplemented with digital therapy solutions used independently (van de Sandt-Koenderman, 2011; Des Roches & Kiran, 2017; Braley et al., 2021).

Aphasia is a relatively common language disorder that occurs after completed language development because of brain damage, which in 80% of the cases is caused by a stroke (Wiehage & Heide, 2016). People with aphasia (PWA) benefit from high-intensity SLT (Bhogal, Teasell, & Speechley, 2003; Brady et al., 2016) and express the necessity of digitalized speech production exercises with appropriate feedback (Kitzing et al., 2009). However, commercial systems with excellent ASR results in applications for typical speakers demonstrate poor performance when processing impaired speech (Green et al., 2021).

In general, deteriorated condition of speech, high variability among speakers, and insufficiency of data make it difficult to use automatic speech recognition for aphasic speech. Errors in oral speech production, such as imprecise articulation and phonemic structure distortions, are mostly inconsistent and unpredictable, which hinders error modelling (Abad et al., 2013). Aphasia can be also comorbid with motor speech disorders, which bring further disfluencies and decrease speech intelligibility (Qualls, 2012). Besides, age is a risk factor for stroke and aphasia (see Schulz & Werner, 2019), and older individuals tend to recover from aphasia more slowly and to a lesser extent. Age per se can influence speech production on various linguistic levels, including acoustics and prosody (e.g., slower speech rate) (Johnson et al., 2022). Changes in acoustic features are reflected in poorer ASR performance for older speakers, which might be more drastic for female voices (Vipperla, Renals, & Frankel, 2008). On the other hand, aphasia generally affects more men than women (see Schulz & Werner, 2019), and some studies report higher ASR error rates on the speech of males (Adda-Decker & Lamel, 2005), while others note that ASR systems might perform poorer for female speakers because of the deviations from the data on which the systems have been historically trained (see Hirschberg, Litman, & Swerts, 2004).

Slower speech rate, increased duration of the utterances, and hyperarticulation in general – the features typical for aphasic speech – have been reported as factors decreasing the conventional ASR performance on typical speech in various languages, for example, English (Siegler & Stern, 1995; Hirschberg, Litman, & Swerts, 2004; Goldwater, Jurafsky, & Manning, 2008), Japanese (Shinozaki & Furui, 2001) and German (Soltau & Weibel, 1998), and contexts.

While different authors explore the possibility of making ASR systems more suitable for the recognition and assessment of aphasic speech (see for review Adikari et al., 2023; Azevedo et al., 2024; Pottinger & Kearns, 2024), to the best of authors' knowledge, there are currently three systems that use ASR for feedback on correctness/incorrectness in naming-oriented semantic exercises (Abad et al., 2013; Ballard et al., 2019; Barbera er al., 2021). In the apps for German-speaking PWA this option is under research (Lin et al., 2022; Heide et al., 2023). AphaDIGITAL (TDG, 2021) project focuses on developing a solution for German-speaking PWA that will provide detailed phonemic/phonetic and semantic feedback in naming and other exercises (see Rykova & Walther, 2024). For this purpose, four open-source ASR solutions have been selected as the most suitable for PWA's speech recognition based on their performance on other types of atypical speech. In the absence of necessary data from PWA, test material from other corpora with atypical speech was considered for evaluation. This paper presents the analysis of the models' robustness to extralinguistic factors and the effects of linguistic features on recognition rates.

2 Materials and methods

2.1 ASR models

The performance of four open-source ASR models, selected for the future pipeline of PWA's speech analysis, was subject to the current experiments. The models are presented in Table 1.

Name in the current paper: description	Reference
jonatas53: fine-tuned Facebook's Wav2Vec2-XLSR-53 model (Conneau et al., 2021) on German CV 6.1 dataset.	Grosman, 2023
mfleck: fine-tuned Facebook's Wav2Vec2-XLS-R-300M model (Conneau et al., 2021) on German CV dataset.	Fleck, 2023
nvidia2: a "large" version of Conformer-Transducer model, trained on several thousand hours of German speech data, NeMo toolkit.	NVIDIA, 2023
oliver9: fine-tuned Facebook's Wav2Vec2-XLSR-53 on German CV 9.0 dataset.	Guhr, 2023

Table 1: Evaluated open-source ASR models

2.2 Speech corpora

In the absence of necessary data from PWA, test material from other corpora with impaired speech was considered for the present evaluation, namely speech of adult cochlear implants (CI) users from CI Articulation Corpus (Neumeyer, 2009) and speech under intoxicated condition from Alcohol Language Corpus (ALC) (Schiel et al., 2008). The deteriorated features of CI users' and intoxicated speakers' speech resemble those of PWA's. In particular, decreased vowel exactness and precision of articulators' movements characterize the speech of adult CI users, which is also reflected in lower automatic recognition rates (Ruff et al., 2017; Arias-Vergara et al., 2022). Speakers in intoxicated condition demonstrate decreased speech rate and weakened speech motor control, noticeable both for human perception and digital applications (Pisoni & Martin, 1989; Tisljár-Szabó et al., 2014).

Naming-oriented exercises in the existing solutions are oriented on one-word recognition. AphaDIGITAL will include advanced exercises that entail phrase production (e.g., picture description), so the evaluation included both single words and phrases. The analysis included the following audio recordings from ALC and CI corpora:

NA_phrases – 641 phrases uttered by sober speakers from ALC corpus;

A_phrases – 702 phrases uttered by intoxicated speakers from ALC corpus;

NA_words – 1976 words, automatically segmented out of the tongue-twisting lists uttered by sober speakers from ALC corpus;

A_words – 2249 words, automatically segmented out of the tongue-twisting lists uttered by intoxicated speakers from ALC corpus;

NORM_words – 1032 words, automatically segmented out of the sentences uttered by normal-hearing speakers from CI corpus;

CI_words – 1021 words, automatically segmented out of the sentences uttered by CI users from CI corpus.

Due to the requirements of some ASR models, all audio recordings described below were (if

necessary) converted to one channel and resampled to 16 kHz.

2.3 Measurements

Character Error Rate (CER) and HITS measurement (the number of precisely recognized words) were used to evaluate the models' performance. CER values were not normalized, meaning that if there were too many substitutions and/or insertions in the ASR transcription, the CER value could be higher than 1 (or 100%). CER and HITS values were computed with the help of the JiWER Python library (Python Software Foundation, 2023). In word sets, the percentage of empty outputs was also taken into consideration.

The results of recognition were not only compared among the models but also according to the following factors (when applicable):

atypicality: intoxicated/sober condition, usage of cochlear implants;

demographics: gender, age group (young vs old, with 50 years old taken as the division line);

linguistic and speech factors (hereinafter "linguistic"): duration of the analyzed segment in seconds, length of the segment in words or syllables, speech rate measured in words/minute (w/m) or syllables/second (syll/s) – for comparison, intended normal speech rate in German is on average 5.4 syll/s (Dellwo et al., 2006).

CER values according to atypicality were subject to the Student's t-test. Groups based on demographic factors were compared with the help of analysis of variance (ANOVA) with a post-hoc Tukey's Honest Significant Difference (Tukey's HSD) test.

The dependencies between linguistic factors were analyzed via Pearson and Spearman correlation tests. The differences between HITS with respect to linguistic factors were analyzed with a pairwise Wilcoxon signed-rank test. Decision (regression) trees with ANOVA as a fit method (Therneau and Atkinson, 2022) were used to assess the dependency of error rates on linguistic factors. They were created with *rpart* function. The leaf nodes were the mean error rate values for the group of observations selected according to the decision node(s). All the analyses were performed in R (R Core Team, 2023) at 95% confidence.

3 Results

3.1 Atypicality

In phrase recognition, the alcohol intoxication of the speakers affects the performance of all four models, increasing the CER values. In word recognition, alcohol intoxication of the speakers affects the performance of the jonatas53 and oliver9 models, while the CER values of mfleck (the lowest among the four models) and nvidia2 (the highest among the four models) do not change significantly. All four models have lower performance on the speech of CI users. The p-values for the Student's t-test in case of significant difference can be seen in Figure 1.

3.2 Robustness of the selected models to extralinguistic factors

For the four selected models, a graphical representation of the robustness to demographic factors can be seen in Figure 1. The absence of a statistically significant difference in ANOVA and post-hoc Tukey's HSD tests (p-value > 0.05) between CER values of demographic groups is understood under robustness. The significant differences and the corresponding p-values are marked in orange.

In the experiments with NA_phrases, all four models are robust to gender, age, and their interaction. In the experiments with A_phrases, mfleck is not robust to gender: CER values for female speakers are higher.

Mfleck and oliver9 are robust to gender, age, and their interaction in the experiments with NA_words. Jonatas53 is robust to gender, but not to age. Tukey's HSD shows that the underlying difference is CER values for the MO group, which are significantly higher than CER values for both FY and MY groups. In the experiments with A_words, jonatas53, mfleck, and oliver9 are robust to gender, but show significantly higher CER values for the older group. With both datasets, nvidia2 is robust to age, but shows significantly higher CER values for the female group, for A_words, in particular, the difference between FY and MY groups is significant.

Since there is only one normal-hearing young male speaker in the CI corpus and the recognition results for his data do not differ from the corresponding FY group, only age differences are discussed for this dataset. The oliver9 model is robust to age in the experiments with

NORM_words, and jonatas53 and mfleck are robust to age in the experiments with CI_words. In the rest of the comparisons, the CER values for younger speakers are significantly higher.

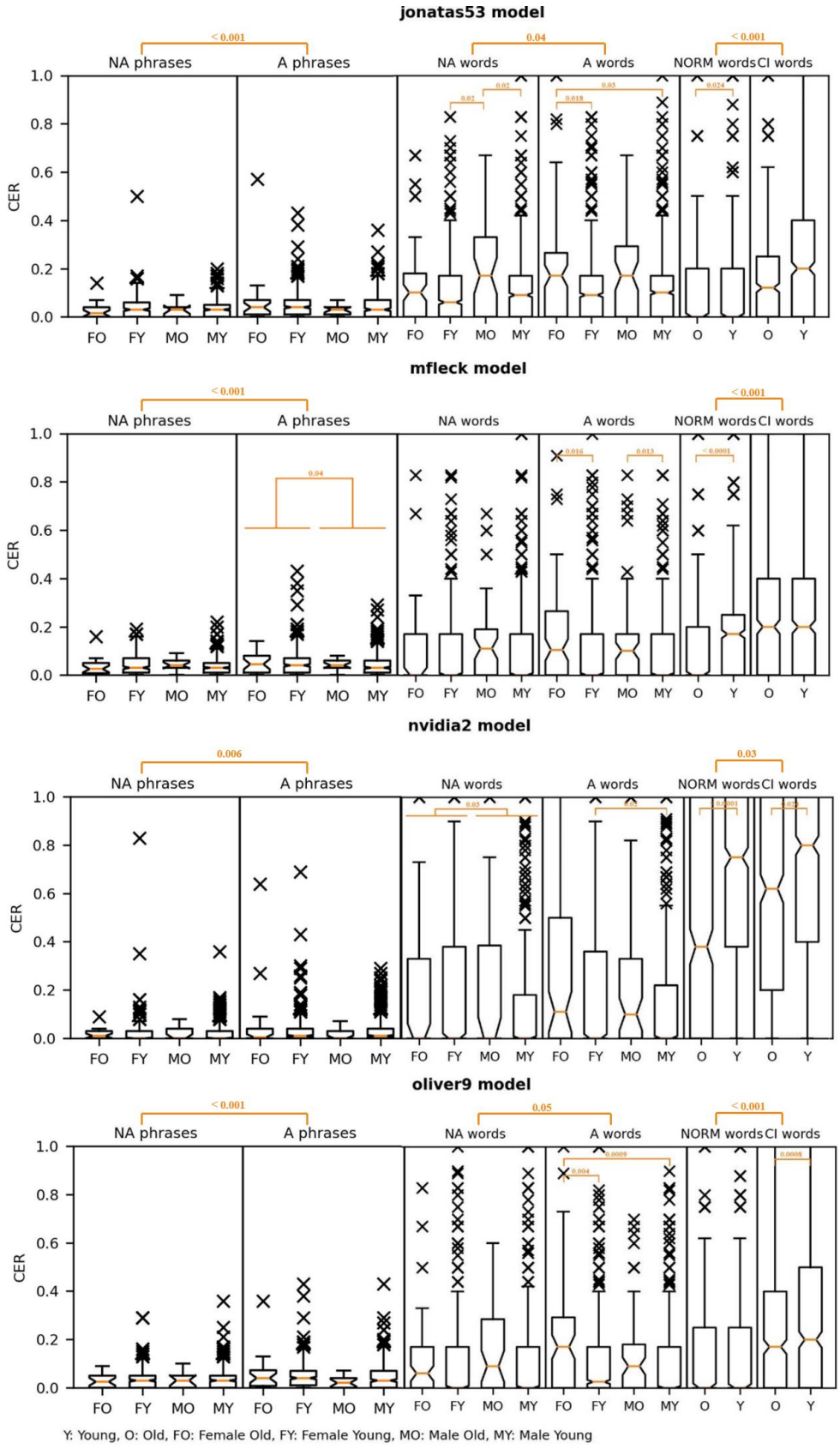

Figure 1: Robustness of the four selected models to gender and age.

3.3 Effect of linguistic factors on the performance of the selected models

3.3.1 Feature description

The numeric values of linguistic parameters are summarized in Table 2. In the ALC phrases datasets, the phrase length correlates strongly with audio duration (ρ = 84%) and speech rate (ρ = 86%). In the ALC words datasets, the word length strongly correlates with audio duration (ρ = 79% for NA_words, ρ = 73% for A_words). In the CI corpus, duration and speech rate have a moderate negative correlation (PCC = 68%). Both alcohol intoxication and CI usage have the following effect on linguistic characteristics: the duration of the same phrases or words becomes longer and the average speech rate becomes slower, but these differences are significant only among young speakers. In the ALC phrases and CI corpus, the speech rate of younger speakers is quicker than that of older speakers, and the duration of the same phrases/words is longer when uttered by the latter.

Dataset	**Duration (s)**				**Speech rate (w/m)**				**Number of words**		
	min	max	M	SD	min	max	M	SD	min	max	med
NA_phrases	1.5	17.5	6.3	2.7	24.2	184.2	89.8	32.8	2	23	10
A_phrases	1.7	17	7	3	22.5	170.4	84.4	31	2	27	10
	Duration (s)				**Speech rate (syll/s)**				**Number of syllables**		
	min	max	M	SD	min	max	M	SD	min	max	med
NA_words	0.3	2.5	0.7	0.3	1.7	7.4	3.9	0.9	1	6	2
A_words	0.3	3	0.8	0.3	1	10	3.8	0.96	1	10	2
NORM_words	0.14	1.1	0.4	0.13	1.8	8.3	4.6	1.2	1	2	
CI_words	0.1	1.6	0.5	0.16	1.2	9.1	4.2	1.3	1	2	

Table 2: Linguistic parameters of the testing datasets

3.3.2 Decision trees for CER

An example of a decision tree for mfleck performance on NA_words can be seen in Figure 2. Following the split according to the speech rate in the root node, and the split according to the duration in the following decision node, the leaf node contains 1705 words, for which the mean CER is the lowest. Combining the tree partitions for several models means choosing the maximum value for greater than and greater than or equal to splitting conditions (> and ≥), and choosing the minimum value for less than and less than or equal to splitting conditions (< and ≤).

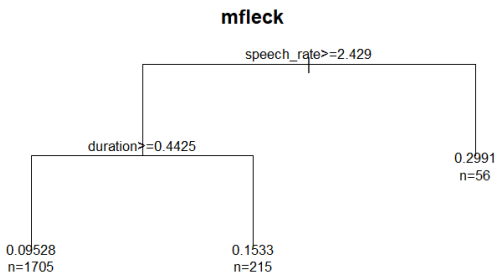

Figure 2: Decision (regression) tree for mfleck performance on NA_words.

Figure 3 presents an extract from the summary of the *rpart* function used for decision tree creation. The condition in the primary split is considered as an alternative one if the difference in the improve between it and the condition chosen for the tree is not greater than 0.01. Thus, in the presented example the number of syllables would be an alternative condition for decision node number 2.

In the experiments with NA_phrases, the most important condition for lower CER (root node) for nvidia2 is speech rate < 60.3 w/m or, alternatively, duration < 4.4 s. For the other three models, it is the phrase length < 6.5 words or, alternatively, sample duration < 4.2 s. In A_phrases, the most important condition is sample duration < 3.2 s (combined for all four models).

Combining the decision trees partition of jonatas53, mfleck, and oliver9 for NA_words and A_words brings out speech rate ≥ 2.9 syll/s and duration ≥ 0.44 s as the two most important conditions, followed by the length of the word ≥ 3.5 syllables for NA_words and word ≥ 4.5 syllables for A_words (this dataset includes longer words). For nvidia2, both speech rate and duration should be higher: > 5 syll/s and ≥ 0.9 s, respectively.

```
Node number 1: 1976 observations,    complexity param=0.03985966
  mean=0.1073684, MSE=0.02690127
  left son=2 (1920 obs) right son=3 (56 obs)
  Primary splits:
      speech_rate < 2.429  to the right, improve=0.03985966, (0 missing)
      n_syllables < 1.5    to the right, improve=0.01170752, (0 missing)
      duration    < 0.4275 to the right, improve=0.01091140, (0 missing)
  Surrogate splits:
      n_syllables < 1.5    to the right, agree=0.972, adj=0.018, (0 split)

Node number 2: 1920 observations,    complexity param=0.01209242
  mean=0.101776, MSE=0.02515304
  left son=4 (1705 obs) right son=5 (215 obs)
  Primary splits:
      duration    < 0.4425 to the right, improve=0.013310100, (0 missing)
      n_syllables < 2.5    to the right, improve=0.009555333, (0 missing)
      speech_rate < 4.528  to the left,  improve=0.005342128, (0 missing)
  Surrogate splits:
      n_syllables < 1.5    to the right, agree=0.895, adj=0.060, (0 split)
      speech_rate < 5.5305 to the left,  agree=0.892, adj=0.037, (0 split)
```

Figure 3: Extract of the rpart function summary for mfleck performance on NA_words.

In the CI corpus, CER values are lower for two-syllable words than for monosyllabic ones. Based on decision trees for NORM_words, the most important condition for jonatas53, mfleck, and oliver9 is duration ≥ 0.27 s in combination with speech rate > 2.1 syll/s. For nvidia2, the duration should be longer than 0.44 s.

For lower CER values in the recognition of CI_words by jonatas53, mfleck, and oliver9, the most important condition is speech rate ≥ 4.3 syll/s, followed by duration ≥ 0.24 s. For nvidia2, the duration should be longer than 0.62 s.

3.3.3 HITS and empty outputs

In the experiments with NA_words and A_words, HITS analysis for jonatas53, mfleck, and oliver9 shows that precisely recognized words are shorter than those with CER > 0. For nvidia2, the empty output is produced for the shortest words uttered at the fastest speech rate.

For the NORM_words, there is a general tendency for precisely recognized words to be longer and uttered at a slower speech rate. The shortest words uttered at the quickest rate are more likely to produce empty output.

The analysis of HITS for CI_words shows a general tendency for precisely recognized words to be longer and uttered at a faster speech rate. The shortest words uttered at the quickest rate are more likely to produce empty output.

4 Discussion

In the experiments with phrases from the ALC corpus, the four models are robust to the gender (as in Goldwater, Jurafsky, & Manning, 2010) and age of the speakers, except for one case: the CER values obtained with mfleck model are higher for female speakers (cf. Vipperla, Renals, & Frankel, 2008). Such results are in contrast with those obtained by Adda-Decker & Lamel (2005), which could be caused by non-natural speech and atypicality in case of intoxication, so that the differences in disfluency, durations, and alternate pronunciations were evened out. Most of the speech material consisted of tongue twisters that had to be pronounced as quickly as possible. The model that performed best for these datasets (nvidia2) yields better results for slower speech rate in NA_phrases. The speech rate is also higher in phrases with more words, and (predictably) the more words a phrase has, the longer its duration is. Thus, one can conclude that extremely high speech rates hinder automatic recognition, which is in line with the studies by Siegler & Stern (1995); Shinozaki & Furui (2001); and one corpus analysed by Hirschberg, Litman, & Swerts (2004). A lower number of words (no more than 6) or shorter duration (generally < 3.2 s) are other decisive factors for better performance in phrase recognition. As stated by Hirschberg, Litman, & Swerts (2004), it is possible that longer phrases just present more space for errors that shorter ones.

In CI_corpus, the speech rate of younger speakers is greater than that of the older speakers, and the duration of the same words is longer when uttered by the latter. That could explain, why in 62.5% of the cases, the models are not robust to age in an unexpected way: lower CER values for older speakers. Excluding nvidia2 (the weakest model for these datasets), ASR systems generally perform

better on audio samples of greater duration (greater than 0.27 s) in combination with speech rates: the lower threshold for normal hearing speakers is 2.1 syll/s, and for the CI users it is 4.3 syll/s. Two-syllable words are recognized with lower CER values on average and are more likely to be recognized precisely, but they should not be uttered too quickly or too slowly, which is in line with the results described by Siegler & Stern (1995), Shinozaki & Furui (2001), and Goldwater, Jurafsky, & Manning (2008).

The experiments with the three models (excluding nvidia2) on words from the ALC corpus confirm that for better single-word recognition the audio samples should be not too short and not too slowly pronounced: duration ≥ 0.44 s and speech rate ≥ 2.9 syll/s (values comparable with NORM_words), correlating with the results on German hyperarticulated speech (Soltau & Waibel, 1998). These datasets contain much longer words than the CI corpus, and there are more relatively shorter words among those that are precisely recognized. The four models are mostly robust to gender, and partially – to age. The CER values for speech samples of older speakers are often higher as in the study by Vipperla, Renals, and Frankel (2008).

Summarizing the above, one can expect that words of moderate length will be recognized better than one-syllable or long words. Speech rate plays an important role in ASR. Thus, in one-word recognition, speech samples uttered at the rates below average of the corresponding datasets, which are lower than intended "very slow" (Dellwo et al., 2006), are more likely to produce higher CER values. Faster speech rates – the maximum values in ALC and CI corpora are higher than intended "very fast" (Dellwo et al., 2006) – also lower the recognition quality, both for words and phrases. For better results with the latter, it is also important to keep the number of words moderate or even low: in the analysed data, the threshold is six words. In the experiments with different datasets, recognition results show inconsistent, and sometimes contrasting, influence of the demographic factors, which might be a consequence of interaction with speech rate. In those datasets, where older speakers speak slower than the younger ones, the CER values of the former are either lower or do not differ from their counterparts. In those with no difference, the CER values for older speakers are higher.

5 Conclusion

The four selected ASR models are generally robust to speakers' gender and age. In fact, the differences might be caused by speech rate rather than by demographic factors per se. Since the models do not necessarily present disadvantages for the speakers of certain gender or age older users, they can be implemented in the error analysis pipeline of the aphaDIGITAL app without a concern that certain users would be treated unfair because of the demographics.

The recognition error rates suggest that words of moderate length are recognized better than one-syllable or long words, which should be taken into consideration when choosing target words for the exercises. Phrase recognition can be included in exercises without drawbacks for the ASR – in fact, phrase recognition might even be more accurate than one-word with the current models.

For better ASR rates, the speech rate of the speaker should be neither too slow (lower than conventional intended "very slow") nor too quick (intentionally speeded up). This knowledge could and should be incorporated into the app instructions ("Please speak at your usual pace") and feedback. For example, if a higher speech rate is detected, the user is asked to speak slower. The findings also suggest that the tasks to produce speech as quickly as possible might not be suitable for assessment with ASR (yet). On the other hand, compensating mechanisms for too slow speech should be elaborated: for example, treating ASR output segments as syllables of one word or adjusting the vowel length and quality.

Ethical Consideration

In the current paper, two speech corpora are explored. Both corpora were downloaded from BAS CLARIN repository (https://clarin.phonetik.uni-muenchen.de/BASRepository/) under free access for scientists.

The app that served as the motivation for current research is viewed as a supplement to in-person SLT and is not to replace SLT practitioners but to allow them to spend more time on complex tasks, which cannot be automatized, during the therapy sessions.

Limitations

The greatest limitation of the current work in general is the lack of relevant data. In the present paper, ASR solutions were tested with atypical speech, but not with the speech of speakers with aphasia.

Acknowledgments

AphaDIGITAL project is sponsored by German Federal Ministry of Education and Research via the TDG innovation ecosystem (Translationsregion für digitale Gesundheitsversorgung [Translational region for digital healthcare]) and „WIR! – Wandel durch Innovation in der Region" [Change through innovation in the region] program.

References

Alberto Abad, Anna Pompili, Ângela Costa, Isabel Trancoso, José G. Fonseca, Gabriela Leal, Luisa Farrajota, and Isabel P. Martins. 2013. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech and Lang*uage, 27(6):1235-1248.

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? *Proceedings of INTERSPEECH 2005*: 2205-2208.

Achini Adikari, Nelson Hernandez, Damminda Alahakoon, Miranda Rose, and John Pierce. 2023. From concept to practice: a scoping review of the application of AI to aphasia diagnosis and management. *Disability and Rehabilitation*, 46(7):1288-1297.

Tomás Arias-Vergara, Anton Batliner, Tobias Rader, Daniel Polterauer, Catalina Högerle, Joachim Müller, Juan-Rafael Orozco-Arroyave, Elmar Nöth, and Maria Schuster. 2022. Adult cochlear implant users versus typical hearing persons: An automatic analysis of acoustic–prosodic parameters. *Journal of Speech, Language, and Hearing Research*, 65(12):4623-4636.

Nancy Azevedo, Eva Kehayia, Gonia Jarema, Guylaine Le Dorze, Christel Beaujard, and Marc Yvon. 2024. How artificial intelligence (AI) is used in aphasia rehabilitation: A scoping review. *Aphasiology*, 38(2):305-336.

Kirrie J. Ballard, Nicole M. Etter, Songjia Shen, Penelope Monroe, and Chek Tien Tan. 2019. Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *American journal of speech-language pathology*, 28(25):818-834

David S. Barbera, Mark Huckvale, Victoria Fleming, Emily Upton, Henry Coley-Fisher, Catherine Doogan, Ian Shaw, William Latham, Alexander P. Leff, and Jenny Crinion. 2021. NUVA: A naming utterance verifier for aphasia treatment. *Computer Speech & Language*, 69(101221).

Sanjit K. Bhogal, Robert Teasell, and Mark Speechley. 2003. Intensity of aphasia therapy, impact on recovery. *Stroke*, 34(4):987-993.

Marian C. Brady, Helen Kelly, Jon Godwin, Pam Enderby, and Pauline Campbell. 2016. Speech and language therapy for aphasia following stroke. *Cochrane database of systematic reviews*, CD000425(6).

Michelle Braley, Jordyn Sims Pierce, Sadhvi Saxena, Emily De Oliveira, Laura Taraboanta, Veera Anantha, Shaheen E. Lakhan, and Swathi Kiran. 2021. A virtual, randomized, control trial of a digital therapeutic for speech, language, and cognitive intervention in post-stroke persons with aphasia. *Frontiers in Neurology*, 12:626780.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of INTERSPEECH 2021*, pages 2426-2430.

Volker Dellwo, Emmanuel Ferragne, and François Pellegrino. 2006. The perception of intended speech rate in English, French, and German by French speakers. In *Proceedings of the 3rd International Conference of Speech Prosody Speech Prosody 2006*, pages 101-104.

Carrie A. Des Roches and Swathi Kiran. 2017. Technology-based rehabilitation to improve communication after acquired brain injury. *Frontiers in Neuroscience*, 11:382.

Michael Fleck. 2023. Wav2vec2-large-xls-r-300m-german-with-lm. https://huggingface.co/mfleck/wav2vec2-large-xls-r-300m-german-with-lm (last accessed 12.09.2023).

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2008. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates. In *Proceedings of ACL-08: HLT*, pages 380-388.

Jordan R. Green, Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, and Katrin Tomanek. 2021. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Proceedings of INTERSPEECH 2021*, pages 4778-4782.

Jonatas Grosman. 2023. Fine-tuned XLSR-53 large model for speech recognition in German. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german (last accessed 12.09.2023).

Oliver Guhr. 2023. Wav2vec2-large-xlsr-53-german-cv9. https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv9 (last accessed 12.09.2023).

Judith Heide, Jonka Netzebandt, Stine Ahrens, Julia Brüsch, Teresa Saalfrank, and Dorit Schmitz-Antonischki. 2023. Improving lexical retrieval with LingoTalk: an app-based, self-administered treatment for clients with aphasia. *Frontiers in Communication, 8:1210193*

Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech communication*, 43(1-2):155-175.

Lisa Johnson, Samaneh Nemati, Leonardo Bonilha, Chris Rorden, Natalie Busby, Alexandra Basilakos, Roger Newman-Norlund, Argye E. Hillis, Gregory Hickok, and Julius Fridriksson. 2022. Predictors beyond the lesion: Health and demographic factors associated with aphasia severity. *Cortex*, 154:375-389.

Peter Kitzing, Andreas Maier, and Viveka Åhlander. 2009. Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phoniatrics Vocology*, 34(2):91-96.

Yuchen Lin, Philipp Klumpp, Jakob Pfab, Abdelaziz Abdelioua, Daniel Gebray, and Mona Späth. 2022. Eine automatische Sprachbewertung für die neolexon Aphasie-App mithilfe Künstlicher Intelligenz [automatic language assessment with artificial intelligence. for the neolexon aphasia app.] Poster session presentation at Sprachtherapie aktuell: Forschung - Wissen – Transfer 9(1): XXXIV. *Workshop Klinische Linguistik e2022-11*, April 2022.

Veronika Neumeyer. 2009. Phonetische Untersuchungen der Artikulation von CI-Trägern [phonetic examination of the CI users' articulation]. Master's Thesis, Ludwig-Maximilians-Universität München, Germany.

NVIDIA. 2023. Conformer-Transducer Large (de). https://huggingface.co/nvidia/stt_de_conformer_transducer_large (last accessed 12.09.2023).

David Pisoni and Christopher Martin. 1989. Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, 13(4):577-587.

Gordon Pottinger and Áine Kearns. 2024. Big data and artificial intelligence in post-stroke aphasia: A mapping review. *Advances in Communication and Swallowing*:1-15.

Python Software Foundation. 2023. JiWER: Similarity measures for automatic speech recognition evaluation. https://jitsi.github.io/jiwer (last accessed 15.12.2023).

Constance Qualls. 2011. Neurogenic disorders of speech, language, cognition-communication, and swallowing. In *Communication Disorders in Multicultural and International Populations*, pages 148–163, Mosby. Elsevier.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Suzan Ruff, Tobias Bocklet, Elmar Nöth, Joachim Müller, Eva Hoster, and Maria Schuster. 2017. Speech production quality of cochlear implant users with respect to duration and onset of hearing loss. *ORL, Journal of Oto-Rhino-Laryngology and its Related Specialties*, 79(5):282–294.

Eugenia Rykova and Mathias Walther. 2024. AphaDIGITAL – Digital Speech Therapy Solution for Aphasia Patients with Automatic Feedback Provided by a Virtual Assistant. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, pages 3385-3394.

Florian Schiel, Christian Heinrich, and Sabine Barfüsser. 2008. Alcohol language corpus: the first public corpus of alcoholized German speech. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1641-1645.

Jörg B. Schulz and Cornelius J. Werner. 2019. *Statistischer Jahresbericht 2018 [Statistical Annual Report 2018]*. Aphasia Station, Neurology Clinic, Aachen University Hospital.

Takahiro Shinozaki and Sadaoki Furui. 2001. Error analysis using decision trees in spontaneous presentation speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'01*:198-201.

Matthew A. Siegler and Richard M. Stern. 1995. On the effects of speech rate in large vocabulary speech recognition systems. In *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1:612-615.

Hagen Soltau and Alex Waibel. 1998. On the influence of hyperarticulated speech on recognition performance. In *Proceedings of the International Conference on Spoken Language Processing-98*, pages 225-228.

TDG - TRANSLATIONSREGION FÜR DIGITALE GESUNDHEITSVERSORGUNG. 2021. *AphaDIGITAL: Entwicklung einer digitalen,*

dezentralen sprachtherapeutischen Versorgung [Development of digital, decentralized speech therapy solutions]. https://inno-tdg.de/projekte/aphadigital/ (last accessed 25.01.2022).

Terry Therneau and Elizabeth J. Atkinson. 2022. *Introduction to recursive partitioning using the rpart and routines.* Technical report, Mayo Foundation.

Eszter Tisljár-Szabó, Renáta Rossu, Veronika Varga, and Csaba Pléh. 2013. The effect of alcohol on speech production. *Journal of Psycholinguistic Research*, 43(6):737-748.

Wilhelmina Mieke E. van de Sandt-Koenderman 2011. Aphasia rehabilitation and the role of computer technology: Can we keep up with modern times? *International journal of speech-language pathology*, 13(1):21-27.

Ravichander Vipperla, Steve Renals, and Joe Frankel. 2008. Longitudinal study of ASR performance on ageing voices. In *Proceedings of INTERSPEECH 2008*, pages 737-748.

Anne Wiehage and Judith Heide. 2016. *Aphasie Informationen für Betroffene und Angehörige. [Information on aphasia for affected individuals and relatives]*. German federal association of academic speech therapists.