Loquens 12
December 2025, e116
eISSN: 2386-2637
https://doi.org/10.3989/loquens.2025.e116

Evaluation of German Automatic Speech Recognition solutions in the context of speech and language therapy support of people with aphasia

Eugenia Rykova
University of Eastern Finland, Finland
Technical University of Applied Sciences TH Wildau, Germany
Catholic University Eichstätt-Ingolstadt, Germany
eugenryk@uef.fi, ORCID iD: https://orcid.org/0000-0003-3819-0949

Mathias Walther
Technical University of Applied Sciences TH Wildau, Germany
mathias.walther@th-wildau.de, ORCID iD: https://orcid.org/0009-0001-8451-8290

Submitted: 11/06/2024; Accepted: 31/07/2024; Published online: 30/06/2025

Citation: Rykova, E. & Walther, M. (2025). Evaluation of German Automatic Speech Recognition solutions in the context of speech and language therapy support of people with aphasia. *Loquens*, 12, e116. https://doi.org/10.3989/loquens.2025.e116.

ABSTRACT: Those who suffer from aphasia benefit from digital speech and language therapy solutions, and automatic speech recognition (ASR) has been already used for giving feedback on the correctness of the answers in naming exercises. AphaDIGITAL application is to provide German-speaking users with detailed feedback on phonemic/phonetic and semantic errors, based on automatic speech and language processing. For this purpose, open-source ASR solutions for German were evaluated on different corpora of atypical speech, including two small datasets with aphasic speech samples. Character error rate, the number of precisely recognized items and empty outputs served as evaluation metrics. The four selected models are generally robust to the deteriorated condition of speech and audio quality and consistently outperform commercial models in atypical speech recognition. Applying error acceptance threshold, additional use of phonemic error rate, and other valuable insights for ASR implementation in aphaDIGITAL are discussed.

Keywords: aphasia; automatic speech recognition; speech and language therapy; digital health.

RESUMEN: *Evaluación de soluciones de reconocimiento automático del habla en alemán en el contexto del apoyo a la terapia del lenguaje para las personas con afasia.* Aquellos que sufren de afasia se benefician de soluciones digitales de terapia del lenguaje, y el reconocimiento automático del habla (RAH) ya se ha utilizado para proporcionar retroalimentación sobre la corrección de las respuestas en ejercicios de denominación. La aplicación aphaDIGITAL debe ofrecer a los usuarios germanohablantes una realimentación detallada sobre errores fonémicos/fonéticos y semánticos, basada en el procesamiento automático del habla y lenguaje. A tal fin, se evaluaron soluciones del RAH de código abierto para el alemán con diferentes corpus de habla atípica, incluidos dos pequeños conjuntos de datos con muestras de habla afásica. Se utilizaron como métricas de evaluación la tasa de errores en los caracteres, el número de elementos precisamente reconocidos y las salidas vacías. Los cuatro modelos seleccionados son generalmente robustos frente al deteriorado estado del habla y la calidad del audio, y consistentemente superan a los modelos comerciales en el reconocimiento del habla atípica. Se discuten la aplicación del umbral de aceptación de errores, el uso adicional de la tasa de errores en fonemas y otros conocimientos valiosos para la implementación del RAH en aphaDIGITAL.

Palabras clave: afasia; reconocimiento automático del habla; terapia del lenguaje; software médico.

1. INTRODUCTION

Using automatic speech processing tools, including Automatic Speech Recognition (ASR), in speech and language pathology has become increasingly popular in the last two decades. Such tools provide valuable help in diagnostics and therapy when used by a speech and language therapy (SLT) practitioner (Keshet, 2018), on the one hand, and on the other hand, contribute to more autonomous healthcare (Hönig & Nöth, 2016). In particular, mobile applications to support SLT are becoming popular (Griffel *et al.*, 2019; Vaezipour *et al.*, 2020).

Aphasia is a language disorder that occurs after completed language development due to brain damage, which in 80% of the cases is caused by a stroke. Every year, aphasia affects 25,000 new patients in Germany (Wiehage & Heide, 2016). SLT improves functional communication of those who suffer from aphasia, with certain benefits brought by high intensity and duration of the therapy (Bhogal and Speechley, 2003; Brady *et al.*, 2016). In reality, not everyone has enough access to extensive or even sufficient SLT because of geographical remoteness, lack of specialists, or other reasons. Nevertheless, in-person therapy can be efficiently supplemented with digital therapy solutions used independently (van de Sandt-Koenderman, 2011; Des Roches and Kiran, 2017; Braley *et al.*, 2021), and oral speech production exercises with adequate feedback are highly desired by users (Kitzing *et al.*, 2009). Vaezipour *et al.* (2020) have analyzed SLT apps for English-speaking people with aphasia (PWA), and from those 70 meeting the eligibility criteria only 24% offer exercises on perceiving and producing oral speech, and while some of them provide automatic feedback, it does not necessarily have high quality.

The aphaDIGITAL project (TDG, 2021) focuses on developing a mobile application for German-speaking PWA that is to provide detailed feedback with the help of speech and text processing in a variety of exercises (cf. Griffel *et al.*, 2019). There are different requirements for the speech recognition solution(s) in the framework of the aphaDIGITAL app. First, it must provide certain phonetic precision (reflecting acoustic modeling), in other words, be able to produce output independently (at least partially) from the existing vocabulary and spelling of the language, or pronunciation and language models in terms of ASR (Keshet, 2018). This is needed for the feedback on pronunciation, which incorporates the committed error(s), for example, phoneme deletion or substitution. On the other hand, a pronounced word must be recognized as an existing one (or at least close to the language reality) in order to be passed further in the pipeline for semantic and grammatical analysis (Rykova & Walther, 2024a). The current paper presents the process of evaluating and selecting ASR solutions for the aphaDIGITAL app, answering the following research questions (RQs):

1. Which existing ASR solutions are suitable for the task-specific speech of German-speaking PWA?
2. How do open-source ASR models perform in comparison to commercial solutions?
3. Which aspects should be considered when implementing an ASR solution for SLT support of PWA?

2. BACKGROUND

2.1. Aphasia speech features in the light of ASR

Aphasia could be translated as “speechlessness” from (Ancient) Greek (Ryalls, 1984). It affects all language modalities: reading and listening (comprehension), and speaking and writing (production). There are several typical clinical pictures of the disorder, but some linguistic symptoms can be considered the most noticeable and universal across PWA. Anomia, or word-finding problems, is one of them (Benson, 1988). This deficit is treated with naming-oriented semantic exercises, which can be automated with the help of ASR (see Section 2.2.2).

Aachen Aphasia Test (AAT) (Huber, 1983) is considered the gold standard in Germany for aphasia diagnosis. Assessment at phonetic and phonemic levels includes a mostly qualitative description of fluidity, vocalization, preciseness, speed, and rhythm (articulation and prosody level), and a quantitative evaluation of the phonemic structure correctness: added, dropped, repeated, or shuffled phonemes in speech output.

Contrary to motor speech disorders, phonetic and phonemic errors in aphasia (a language disorder) are mostly inconsistent and unpredictable. Aphasia can be, however, comorbid with motor speech disorders: apraxia of speech (AOS), and, much less frequently, dysarthria. In AOS, the neurologic mechanisms for motor planning and programming are affected, while the motor function itself remains intact (Qualls, 2011). That results in phonemic structure distortions, speech disfluency, prolonged sound duration, and other prosodic/temporal abnormalities (Le *et al.*, 2016, see also Wambaugh *et al.*, 1996). Dysarthria manifests itself in weakness, slowness, poor coordination, and restricted and imprecise movements of muscles that take part in oral speech production. That causes low intelligibility of speech in general, and such particular deviations as, for example, slower speech rate, strained phonation, irregular articulation, and reduction or deletion of word-initial consonants (Caballero Morales and Cox, 2009; Qualls, 2011). The research on ASR for dysarthric speakers is actively ongoing, for example in adapting acoustic models or modeling the errors (Gutz, 2022).

Aphasia generally affects more men than women, and age is another risk factor for stroke and aphasia (Schulz and Werner, 2019; see also Johnson *et al.*, 2022). Furthermore, older individuals tend to recover from aphasia slower and to a lesser extent. Age per se can influence speech production on various linguistic levels, including acoustics. For example, older individuals speak at a 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 *et al.*, 2008).

2.2. ASR in aphasia diagnostics and therapy

2.2.1. PWA's speech assessment

Plenty of studies have explored the potential of ASR systems to automatically assess the intelligibility of pathological speech (for reviews see Jamal *et al.*, 2017; Keshet, 2018; Adikari *et al.*, 2024). In general, deteriorated condition of speech, high variability among speakers, and insufficiency of data make it difficult to use ASR for aphasic speech. The corresponding solutions should be dynamic and flexible, in the best-case scenario allowing personal tailoring (van de Sandt-Koenderman, 2011). It must be noted that commercial systems with excellent results in many applications for typical speakers demonstrate poor performance on the material of impaired speech. In its turn, personalized models can reach very high recognition rates for the latter (Green *et al.*, 2021).

As applied to aphasia, Le and colleagues explore the possibilities of automatic assessment of the continuous speech produced by English speakers with aphasia (Le *et al.*, 2016), the ways of improving the automatic recognition of aphasic speech (Le & Provost, 2016) and consequent detection of phonemic and neologistic paraphasias (Le *et al.*, 2017). Torre *et al.* (2021) set a new benchmark in ASR for PWA in English and provide the first adapted system for Spanish.

Lee *et al.* (2016) evaluate the feasibility and challenges of assessing continuous speech of Cantonese speakers with aphasia with the help of ASR. Chatzoudis *et al.* (2022) propose fine-tuning of ASR models that share cross-lingual speech representation to low-resource languages and present models for detecting aphasia and transcribing PWA's speech for French and Greek.

Another area of ASR in aphasia assessment includes automatic transcription of the PWA's speech and further analysis of text features, possibly in combination with acoustic features. Such work has been done, for example, on the material of English (Fraser *et al.*, 2013) and Cantonese (Qin *et al.*, 2018; Qin *et al.*, 2020). Kohlschein *et al.* (2018) aim at an automatic version of German AAT (Huber, 1983), which uses acoustic features, phonemic structure, and higher-level linguistic features for diagnosing and classifying aphasia.

2.2.2. Automatic feedback in naming exercises

Virtual Therapist for Aphasia Treatment (VIRTHEA) in European Portuguese, introduced in 2011 (Pompili *et al.*, 2011), seems to be the first system that uses ASR (an in-house ASR engine) to process what is said by the user and evaluate whether the answer was correct or incorrect – a verification task, in other words. The system focuses on naming exercises to improve the word-retrieval ability. Abad *et al.* (2013) state that since VIRTHEA is assumed to be used by PWA with very low (or none) motor speech deficits, a general acoustic model can be used without retraining. However, PWA's speech may contain a considerable number of hesitations, repetitions, and other disruptive factors that weaken ASR performance. Therefore, a keyword spotting method is proposed in order to verify that a correct word has been pronounced during the analyzed speech segment. The system demonstrates promising results on a corpus of nomination tests from native Portuguese speakers with different types of aphasia, in particular high correlation between human and automatic naming scores, and high word verification rates – 82% accuracy on average. VIRTHEA is positively perceived by SLT practitioners and is being updated according to the wishes of the latter (Pompili *et al.*, 2020).

Ballard and colleagues (2019) evaluate an open-source ASR engine to provide binary feedback (correct/incorrect) in a picture-naming task for Australian English and reach a mean accuracy performance of 75%. Naming Utterance Verifier for Aphasia Treatment (NUVA), developed by Barbera *et al.* (2021) for British English, reaches a mean accuracy of 89.5% with a smaller range than VIRTHEA and the system for Australian English (see Barbera *et al.*, 2021 for a detailed comparison). In NUVA, a word pronounced by a user with aphasia is compared to two recordings of healthy speakers and classified as correct or incorrect using a verification threshold. Different threshold calibration methods are applied to a proposed ASR model with a phone error rate of 15.85%. This model consequently outperforms Google Cloud Platform speech-to-text service used as an ASR baseline.

Several research teams work on SLT solutions for German-speaking PWA with ASR-based feedback. Nevertheless, to the best of the authors' knowledge, there are currently no such apps in active use. Lin *et al.* (2022) report 83% recognition accuracy of target words pronounced by PWA in the research for neolexon Aphasie-App (2023) and propose that an SLT specialist should posteriorly analyze the problem cases. Dietmar Bothe, project manager of aphavox (2020), presents the app with automatic recognition of PWA's speech and corresponding feedback in an interview (Halling, 2023). RehaLingo (2023) seeks to combine several speech recognizers and model possible erroneous inputs (Hirsch *et al.*, 2023). LingoTalk (LingoLab, 2020) exploits built-in iOS or Android ASR software in naming exercises and reaches 98% accuracy with typical speech, but there is no data on PWA's speech (Netzebandt *et al.*, 2022).

3. MATERIALS AND METHODS

3.1. Process and models overview

In the present research, the ASR selection process consisted of several steps. First, more than 50 open-source ASR solutions, including models available from Alpha Cephei (2022), Mozilla Deepspeech (Xu *et al.*, 2020), and via Hugging Face (2022), were screened for further suitability (Rykova *et al.*, 2022). The screening procedure was also applied to the commercial models. Next, 13 selected open-source models were evaluated with a considerable amount of atypical speech data. They are presented in Table 1. Eleven models were accessed via Hugging Face framework, and ims_0 and ims_35 are modified versions of the original model with language model (lm) weights set to 0 and 0.35, respectively.

Table 1: Thirteen open-source ASR models evaluated after the initial screening.

Model name in the current paper	Author(s)	Description given by the author(s) of the model
andrew	McDowell, 2022	Fine-tuned Facebook's Wav2Vec2-XLS-R-1B model (Babu *et al.*, 2022) on German Common Voice (CV) 8.0 dataset.
ims_0 (lm weight = 0)	Denisov and Vu, 2019	The original IMS model (lm weight = 0.7) was trained using kaldi German ASR recipe and implemented with ESPnet end-to-end speech recognition toolkit. Datasets: Tuda-De, SWC, M-AILABS, Verbmobil 1 and 2, VoxForge, RVG 1, PhonDat1.
ims_35 (lm weight = 0.35)		
jonatas53	Grosman, 2022a	Fine-tuned Facebook's Wav2Vec2-XLSR-53 model (Conneau *et al.*, 2021) on German CV 6.1 dataset.
jonatas1b	Grosman, 2022b	Fine-tuned Facebook's Wav2Vec2-XLS-R-1B model on German using CV 8.0, Multilingual TEDx, Multilingual LibriSpeech, and Voxpopuli datasets.
jsnfly	Jsnfly, 2022	Fine-tuned Facebook's Wav2Vec2-XLS-R-1B model on German CV 8.0 dataset.
marcel	Bischoff, 2022	Fine-tuned Facebook's Wav2Vec2-XLSR-53 model on German using the CV dataset.
maxidl	Idahl, 2022	Fine-tuned Facebook's Wav2Vec2-XLSR-53 model on German using the CV dataset.
mfleck	Fleck, 2022	Fine-tuned Facebook's Wav2Vec2-XLS-R-300M model (Conneau *et al.*, 2021) on German CV dataset.
nvidia1	NVIDIA, 2022a	A "large" version of Conformer model, trained on several thousand hours of German speech data, NeMo toolkit (Kuchaiev *et al.*, 2019).
nvidia2	NVIDIA, 2022b	A "large" version of Conformer-Transducer model, trained on several thousand hours of German speech data, NeMo toolkit.
oliver8	Guhr, 2022a	Fine-tuned Facebook's Wav2Vec2-XLSR-53 on German CV 8.0 dataset.
oliver9	Guhr, 2022b	Fine-tuned Facebook's Wav2Vec2-XLSR-53 on German CV 9.0 dataset.

Lastly, the thirteen open-source models and the commercial ones were tested with the PWA's speech. Four commercial models were subject to comparison, namely Fraunhofer German ASR (fr-hofer), European Media Lab transcription service (eml), Google Speech Cloud ASR (google), and IBM Watson ASR (watson). The outputs were obtained via BAS web services, available for academic purposes (Kisler *et al.*, 2017). One may use these services for a limited amount of data only, therefore the commercial models were not evaluated together with the open-source ones at the previous step.

During the evaluation, a new highly performing ASR model, Whisper (Radford *et al.*, 2023) was released. A screening with PWA's samples showed, however, that this model would not be suitable for aphaDIGITAL purposes because it failed to recognize speech in the given samples.

3.2. Datasets

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. For the screening step, individual recordings were selected from the following German corpora:

- speech of cochlear implants (CI) users and normal-hearing speakers from CI Articulation Corpus (Neumeyer, 2009) – hereinafter CI corpus,

- speech of intoxicated and sober speakers from Alcohol Language Corpus (Schiel *et al.*, 2008) – hereinafter ALC corpus,
- speech of a person with aphasia from AphasiaBank (MacWhinney *et al.*, 2011),
- speech of eight PWA extracted from a YouTube video (Rhein-Zeitung, 2018),
- typical speech from PHONDAT2 (Hess *et al.*, 1995) for comparison (cf. Wirth and Peinl, 2022).

The transcriptions provided together with the audio were used for the recordings from CI, ALC and PHONDAT2 corpora, and AphasiaBank. The YouTube video was transcribed by the speech science students. The annotators followed the principle of phonemic orthography: they transcribed actual pronunciation rather than a standard orthographic form but used the German graphemes as the output form.

In the absence of necessary data from PWA, test material from other corpora with atypical speech was considered for the main evaluation. Thus, the speech of adult CI users can be characterized by decreased vowel exactness and precision of articulatory movements. It is considered deteriorated, especially with a longer period of deafness or pre-lingual onset, which is also reflected in automatic recognition rates (Ruff *et al.*, 2017; Arias-Vergara *et al.*, 2022). The changes in speech production under intoxicated condition include decreased speech rate and weakened speech motor control, which can be captured by both human perception and digital acoustical analysis (Pisoni & Martin, 1989; Tisljár-Szabó *et al.*, 2014). Hence, the selected 13 models were evaluated with the help of material from ALC and CI corpora, which covers female and male speakers of different ages, presented in Table 2.

Table 2: Datasets of atypical speech used for the evaluation of ASR models.

Dataset name	Number of elements	Description
NA_phrases	1274	phrases uttered by sober speakers from ALC corpus
A_phrases	1404	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

For the last evaluation and comparison step, two datasets with aphasic speech, internally named AvEv and UniSt, were used. AvEv is a small dataset obtained from four PWA who took part in the avatar evaluation experiment (Zeuner *et al.*, 2022). While selecting the correct option in a PC-based picture-naming task, the participants incidentally pronounced the corresponding words. The experiment was videotaped. The audio was extracted from the videos and the words were segmented out, which made a set of 39 single words. It must be, however, kept in mind that the quality of these recordings is low. Besides that, a lot of words were pronounced in a manner deviating from the standard pronunciation (e.g., due to dialectal differences or the presence of aphasia). Two speech science students provided separate annotations (based on the principle of phonemic orthography described above) as alternative ground truth in addition to a standard orthographic form of the target words. UniSt is a dataset comprising 61 words uttered by SLT specialists, and 79 recordings of PWA's responses. The recordings had been made during AAT screening sessions (repetition and picture-naming tasks) with six PWA and were obtained on request from University of Stuttgart Institute for Natural Language Processing, where they are used as learning material in neurolinguistics online tutorial (Universität Stuttgart, 2023). These recordings were also transcribed by the speech science students, following the principle described above. Phonemic transcriptions were generated automatically with a slightly modified version of Deep Phonemizer (Schäfer *et al.*, 2023). Additionally, an SLT specialist classified the PWA's responses in UniSt dataset as containing no error, a phonemic/phonetic error, or a semantic error. A phonemic/phonetic error was understood as such a deviation in a segmental structure of the word that would result in a transcription distinct from the standard orthographic form. The answers with no error or phonemic/phonetic error were considered semantically acceptable. The SLT specialist also provided finer classification of errors according to the ICF (International Classification of Functioning, Disability and Health) guidelines (Schneider *et al.*, 2021).

3.3. Measurements

Character Error Rate (CER) was the main accuracy metric to evaluate the ASR systems:

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

S – the number of substitutions,
D – the number of deletions,
I – the number of insertions,
N – the number of characters in the reference (target),
C – the number of correct characters.

If there are too many substitutions and/or insertions in the ASR transcription, the CER value can be higher than 1 (or 100%). For some comparisons, a normalized CER was calculated. In this case, the total number of substitutions, deletions, and insertions is divided by the maximum length of the sequences in question. CER does not only reflect the performance of an ASR system, but is relevant for granular analysis of impaired speech input. It was calculated separately for each of the evaluation material sets (according to the target phrase/word) and then ranked. The HITS measurement was used to assess the number of precisely recognized words. In word sets, the percentage of empty outputs was also taken into consideration in the evaluation process. Thus, each of the metrics was ranked, and the mean rank was calculated for each model/dataset. CER and HITS (correctly recognized words) were computed with the help of the JiWER Python library (Python Software Foundation, 2022).

CER values were subject to Student's t-test (datasets from ALC and CI corpora) and Wilcoxon Rank Sum Test (AvEv and UniSt datasets). It was decided not to use any correction for multiple comparisons to decrease the risk of Type II error. In other words, it was more important to detect the difference between the models' performance when it was insignificant than miss a significant difference. All the analyses were performed in R (R Core Team, 2023) at 95% confidence.

4. RESULTS

4.1. Model selection

Rykova *et al.* (2022) show some preliminary results of the ASR model screening. Table 3 contains mean CER values (in percent), mean ranks, HITS (H), and empty outputs (E) percentage for each of the 13 models obtained with atypical speech from ALC and CI corpora, CER values for typical and atypical speech are in most cases significantly different. Recognition results on the other datasets can be found in the Annex A. The last two columns present the mean (M) rank for all the datasets used for evaluation and its absolute value (ab), respectively. For words datasets, CER values are given ignoring the missing values. The lowest CER values, the highest ranks and HITS are in bold. The CER values that are significantly lower than the others in a pairwise t-test comparison (p-values < 0.05) are marked with an asterisk.

Table 4 contains mean CER values (in percent), HITS (H), and empty outputs (E) percentage for the 13 open-source models and 4 commercial ones obtained with AvEv and UniSt datasets. For the open-source models, the mean (M) rank per dataset is given for the comparison among them only, and the absolute (ab) rank value is given for the comparisons among all 17. The last column presents the absolute rank among the open-source and commercial models for PWA's speech from AvEv and UniSt datasets. For the AvEv dataset, manual transcriptions differ equally from the orthographic target (normalized CER = 26% and 25%) and achieve a 17% normalized CER in comparison to each other. However, there are no statistical differences between CER values, obtained in the three comparisons: ASR output vs target and ASR output vs two manual transcriptions. To avoid any bias, the CER values obtained from comparisons with target orthographic transcriptions are considered. For the UniSt dataset, manual transcriptions achieve a 4% normalized CER in comparison to each other, and there are no significant differences in CER values obtained in the comparisons of ASR output vs manual transcriptions, so the quantitative results are presented for one of the manual transcriptions only. The results for speech therapists' and PWA's speech are treated separately, as the CER values differ significantly. The best results among open-source models are marked in bold.

Three models with the highest ranks, namely jonatas53, mfleck, and oliver9, are selected as those providing phonetic level granularity, on the one hand, and robust to degraded speech and audio quality, on the other. Additionally, nvidia2 is selected as the model that is able to recognize words close to language reality (i.e., in accordance with pronunciation and language models). They are marked with grey in Table 3 and Table 4. Some further details on the performance of these models can be found in Rykova and Walther (2024b).

4.2. Open-source vs commercial models

The results of the screening phase (Rykova *et al.*, 2022) demonstrate that although google, fr-hofer, and, to some extent, watson models show top results in precise word recognition, this performance drops on atypical speech, which includes both speech in deteriorated condition and unusual phrases uttered by typical speakers. Besides that, the mean CER values of commercial models are noticeably higher than those of open-source ones.

Table 3: ASR results for 13 models obtained with atypical speech from ALC and CI corpora.

Model	A_phrases			A_words				CI_words				all datasets	
	CER	H	M rank	CER	H	E	M rank	CER	H	E	M rank	rank	
												M	ab
andrew	7.4	64.3	11.6	12.5	49.4	0	4.4	33.7	37.6	2.4	7.3	7.0	9
ims_0	9.2	69.8	12.1	23.1	37.3	1.2	10.6	55.9	24.0	26.8	10.4	10.4	12
ims_35	8.6	75.1	9.1	25.1	39.7	2.8	11.0	64.0	21.4	38.9	11.9	10.4	13
jonatas53	5.3	76.4	5.6	14.0	44.7	0	5.3	**20.9***	**59.5**	0	**1.3**	4.2	**2**
jonatas1b	5.4	79.5	4.4	16.4	**57.5**	0.3	6.2	37.0	36.9	4.9	8.7	5.6	6
jsnfly	5.8	72.4	8.1	12.5	**58**	0	**2.3**	31.5	47.0	0	4.5	5.2	4
marcel	6.8	70.2	10.0	13.4	46.8	0	5.0	23.8	45.4	0	4.0	7.0	10
maxidl	6.6	73.2	8.4	13.8	49.6	0	4.6	27.5	49.9	0	3.8	6.6	7
mfleck	5.2	77.4	4.1	**10.9***	57.5	0	2.4	24.6	54.3	0	3.2	**2.7**	**1**
nvidia1	**3.8***	**86.9**	2.9	44.1	28.2	19.6	12.9	78.7	17.0	55.5	12.8	8.7	11
nvidia2	**3.7***	**87.0**	**1.3**	21.2	55.6	7.4	8.7	65.0	31.4	27.2	10.7	6.7	8
oliver8	5.7	71.7	8.1	12.7	52.3	0	3.8	23.6	54.6	0	2.7	5.5	5
oliver9	5.2	77.3	5.3	12.1	52.2	0	4.1	24.7	58.1	0.1	4.4	4.3	**3**

CER – character error rate (in percent); H – HITS percentage; E – empty outputs percentage; M – mean value; ab – absolute value.
The lowest CER values, the highest ranks and HITS are in bold. The CER values that are significantly lower are marked with an asterisk.
The models selected for the current app after the evaluation are marked with grey.

Table 4: ASR results for 13 open-source and 4 commercial models obtained with AvEv and UniSt datasets.

Model	AvEv					UniSt therapist					UniSt PWA					all PWA
	CER	H	E	rank		CER	H	E	rank		CER	H	E	rank		ab rank
				M	ab				M	ab				M	ab	
andrew	65.6	5.1	2.6	6.0	8	38.2	16.4	0	6.0	10	49.1	6.3	0	3.3	4	5
ims_0	58.4	2.6	79.5	8.0	12	29.8	19.7	29.5	8.0	14	62.8	3.2	30.4	11	15	14
ims_35	86.5	0	89.7	12.3	16	29.1	21.3	42.6	7.3	13	63.3	5.3	38.0	11	15	16
jonatas 53	67.8	7.7	0	4.3	**3**	41.3	24.6	0	5.0	4	58.5	6.3	0	4	6	4
jonatas 1b	62.2	7.7	2.6	4.7	6	43.5	31.1	0	4.7	4	58.5	**12.6**	0	**2**	**2**	**2**
jsnfly	98.2	0	0	8.7	13	42.7	23.0	0	5.7	7	54.4	8.4	0	3.3	5	8
marcel	70.9	2.6	2.6	8.3	11	52.2	4.9	0	9.0	15	68.7	1.1	0	9	12	13
maxidl	64.9	2.6	10.3	7.7	9	39.7	19.7	0	5.7	7	62.3	4.2	1.3	9.3	12	12
mfleck	54.0	17.9	0	**1.7**	**1**	**28.0**	31.1	0	**1.3**	**1**	**45.1**	**12.6**	0	**1**	**1**	**1**
nvidia1	74.1	5.1	41.0	9.3	14	34.4	29.5	4.9	6.0	7	60.6	10.5	8.9	6.7	7	11
nvidia2	**53.1**	**20.5**	20.5	4.0	**2**	39.9	**32.8**	6.6	6.7	11	63.3	10.5	12.7	8	10	6
oliver8	66.4	5.1	0	4.7	6	43.9	19.7	0	7.3	12	63.6	6.3	0	6.7	9	7
oliver9	69.2	10.3	0	4.3	**3**	38.9	21.3	0	4.7	**3**	58.9	10.5	0	3.3	**3**	**3**
eml	n/a	0	100		17	65.8	4.9	67.2		17	73.8	1.1	77.2		17	17
fr-hofer	56.6	**23.1**	23.1		**3**	38.6	**41.0**	6.6		4	67.1	9.5	12.7		11	9
google	**53.7**	2.6	84.6		10	**26.3**	39.3	21.3		**2**	50.1	9.5	38.0		8	10
watson	78.1	0	66.7		15	43.2	18.0	36.1		16	60.0	6.3	45.6		12	15

CER – character error rate (in percent); H – HITS percentage; E – empty outputs percentage; M – mean value; ab – absolute value.
The lowest CER values, the highest ranks and HITS are in bold.
The models selected for the current app after the evaluation are marked with grey.

From the evaluation of the AvEv dataset, the performance of the fr-hofer model seems to be the best among the commercial models and comparable to the top open-source models. It also has the highest percentage of HITS among all models and a relatively low CER value, yet this holds true for about three-quarters of the initial data only (non-empty output). Although the mean CER value for the google model also is relatively low, it only accounts for less than 16% of the initial data, and the number of HITS is in the lowest range.

The performance of the fr-hofer and google models on therapists' speech from the UniSt dataset is at the top: among all the models, they have the highest number of HITS and the mean CER for google is the lowest (but on about 80% of the data only). The results change extremely when the models deal with PWA's speech. Thus, both the mean CER value and the percentage of empty outputs for google almost double, and the number of HITS decreases more than four times. The drop in the fr-hofer model performance is similar (its mean CER value increases 1.7 times). The mean CER values of the four selected open-source models increase approximately 1.5 times; the highest numbers of HITS (obtained with mfleck and nvidia2) decrease 2.5-3 times; and the empty output of nvidia2 doubles, while the other three selected models produce results on all the data. Furthermore, none of the commercial models recognizes distorted pronunciations as the human transcribers, producing HITS only on canonical transcriptions of existing words, in distinction to jonatas53, mfleck, and oliver9.

4.3. Evaluating the models on PWA's speech

The lowest mean CER value for the whole AvEv dataset – 54% (ranging from 0 to 125%, standard deviation SD = 34.6%) – is achieved with the mfleck model. To compare, a mean CER value of 53.1% (ranging from 0 to 150%, SD = 42%) is achieved with nvidia2, but on 80% of the data (20% is empty output). If this value is taken as a threshold for accepting ASR output text as correct, the total number of the words accepted by any of the four selected models reaches 28, which is 72% of the given dataset.

The lowest mean CER values for both parts of the UniSt dataset are also reached with the mfleck model: 28% on therapists' speech (ranging from 0 to 117%, SD = 29.5%), and 45.1% on PWA's speech (ranging from 0 to 150%, SD = 32.1%). Most of the HITS are obtained on the orthographic form of existing words, even when these represent erroneous speech production. For example, a person says "twist" instead of the target *Zwist* 'dispute', and "twist" is recognized correctly by an ASR model (i.e., it's a HIT), but meanwhile *Twist* 'twist' is an actual word, too. Few non-existing forms, representing different degrees of deterioration, are recognized: "schwern" (target *Stern* 'star') – with oliver9; "schweibmaschine" (target *Schreibmaschine* 'typewriter') – with jonatas53; "schwo" (target *Strumpf* 'stocking'), "losig" (separately pronounced part of target *Verantwortunglosigkeit* 'irresponsibility'), and "poloret" (target *Lotterie* 'lottery') – with mfleck. Considering transcriptions of both transcribers as a possible target, the four selected models together can recognize 54% HITS on the speech of speech therapists, and 24% on the PWA's speech.

It must be noted that German orthography principles include several ambiguities, and the same sounds or sound combinations can be transcribed in different ways, which is especially relevant for non-existing words. For example, spellings "schweibmaschine" (jonatas53) and "schwaibmaschine" (mfleck) correspond to the same pronunciation; or the initial phoneme /ʃ/ is transcribed as "sch" or "s" in "schwern" and "stern", respectively. Thus, an additional comparison of (automatically generated) phonemic transcriptions (i.e., CER for phonemic transcriptions – PER) might be relevant. In the case of the UniSt dataset, such comparison brings one more HIT among non-existing words. In the AvEv dataset, there are two additional words, whose PER is lower than the 54% threshold, which increases the joint acceptance rate to 77%.

The proposed approach was tested using 54% as a CER/PER threshold to accept PWA's answers as semantically correct (the error rate value is below the threshold) or not. First, the human transcriptions were compared to the corresponding orthographic target as if that were an ideal ASR model. For six recordings, there was a mismatch between human and threshold-based answer acceptance. One error, classified as semantic paraphasia by the SLT specialist ("kraftfahr brief" vs target *Kraftfahrzeugschein*, which refer to two different documents), would be automatically attributed to a phonemic/phonetic error because more than half of the word is pronounced correctly. Five of them were classified as a phonemic/phonetic error by the SLT specialist (in particular, phonemic conduite d'approche – "approaching" the target with self-corrections, and phonemic neologisms – phonemic changes in the target that make the latter hard to recognize), but the error rates exceeded 54%. Non-normalized error rates are especially sensitive to insertions, which can be ignored by a human listener to recognize the target word surrounded by extra phonemes (e.g., "likurk" vs target *Kur* 'cure'). In this case, using normalized error rates could be a solution, which reduces the total number of error classification mismatches to five. It must be noted that for two more semantically accepted phonemic neologisms, the CER/PER values were only slightly below the threshold (50%).

Furthermore, PWA might utter extra words together with the target word (e.g., articles or false starts). If the target is pronounced correctly, there will be a HIT, but the CER value will be higher than 0, including higher than the acceptance threshold. If there is a deviation in pronunciation or a flaw in automatic recognition, apart from high error rates, there will be no HIT, although an SLT specialist would accept such an answer. Thus, it seems reasonable not only to look for a target word in the uttered phrase but perform a CER/PER analysis for each recognized word of the output. On the other hand, some PWA speak so slowly and carefully/laboriously that the syllables of one word are recognized as separate words.

That causes a rise in CER/PER values, which might lead to the rejection of an answer that would be accepted by an SLT specialist. In this case, deleting the spaces between the output chunks and treating the whole output as one word might be useful.

Taking into consideration the above-mentioned points and using the 54% normalized CER/PER acceptances threshold, ASR outputs of the four selected models were compared to the target words with a subsequent automatic error classification. The comparison of the manual and automatic error classification is displayed in Table 5. Five cases of error mismatches described above are excluded from the table. The ASR outputs in these cases yield the same results (i.e. mismatches) as the human transcriptions.

Table 5: Manual and automatic error classification on the UniSt PWA dataset.

Manual classification	Automatic classification		
	no error	phonemic/phonetic error	semantic error
no error	12	20 + 3 gained through error rate normalization + 1 gained through separate analysis of each word + 2 gained through deleting the spaces	3
phonemic/ phonetic error	0	16 + 3 gained through separate analysis of each word + 1 gained through deleting the spaces	4
semantic error	0	0	9

As one can see, there are no false positives among the automatically classified errors. From the samples accepted by the SLT practitioner, 10.8% are erroneously classified as semantic errors. In 63% of the fully correct answers, ASR models are only able to reach the level of a phonemic/phonetic error, although the answer would be accepted as semantically correct.

5. DISCUSSION

5.1. Selection of the open-source models for aphasic speech recognition (RQ 1)

Based on the experiments with various speech material in German, including speech samples from PWA and other atypical speech, four open-source ASR models are selected for the backend of the aphaDIGITAL app. Three of these models (jonatas53, mfleck, oliver9) are to a certain extent independent of pronunciation and language models and are suitable for phoneme-level pronunciation analysis, while the fourth model (nvidia2) gives only existing orthographic forms as output, which is more suitable for subsequent semantic and grammatical error analysis. The error-analysis component will use every distinct ASR output for comparison to the target. The selected four models present a possibility to be fine-tuned: to PWA's speech or speech of a particular user in a customized version, and to word recognition task rather than continuous speech recognition.

5.2. Comparison of the selected open-source models to the commercial ones (RQ2)

The selected models have consistently outperformed commercial systems in recognition of atypical speech, which in the current paper is primarily reflected in a high amount of empty outputs in the experiments with PWA's speech, and a great contrast between the results on speech therapists' and PWA's speech samples from UniSt dataset (cf. Green *et al.*, 2021 and Wirth and Peinl, 2022). Results of the screening phase (Rykova *et al.*, 2022) suggest that when Google Speech Cloud ASR and Fraunhofer German ASR recognize the words, they recognize them precisely, but imprecisely recognized words are far from the target (cf. Barbera *et al.*, 2021). Such precise recognition of Google and Fraunhofer models holds true for the data from the AvEv dataset (orthographical form was used as the target in the experiments) and for canonically pronounced words from the UniSt dataset when the target (manual transcription) coincided with the orthographic form of the word. The commercial models cannot recognize distorted pronunciations as such, in contrast to the "rule-independent" open-source models.

5.3. Model performance on PWA's speech (RQ 3)

The experiments with the AvEv corpus suggest that the selected models together can reach a 72% recognition rate if the CER threshold of acceptance is set to 54%. For the UniSt dataset, manual transcription was used as the target, and the four selected models together transcribed 24% of the words uttered by PWA exactly like human experts (cf. 54% of words uttered by speech therapists in the same dataset), including some non-existing forms. Considering the ambiguities of German orthography principles, incorporating phonemic comparisons and an additional PER threshold seems to be reasonable for the proposed application. Introducing PER with the same 54% threshold allows increasing the joint acceptance rate on the AvEv dataset to 77%. Further research will be dedicated to exploring the threshold further, evaluating false positives. On the other hand, a number of dialect variations seen in the data suggest the relevance of incorporating the knowledge on systematic dialect changes into the pipeline (cf. Pompili *et al.*, 2011).

The variability of answers produced by PWA suggests the following ASR implementation concerns. First, to overcome the fact that there might be other words in the answer besides the target, the ASR output should be segmented into separate words for further analysis. To reduce the adverse effect of insertions further, the CER/PER threshold should be used for a normalized comparison, in other words, the distance between orthographic/phonemic transcriptions should be divided not by the length of the target, but by the length of the longest word in the comparison. Finally, due to PWA's laborious speech production, ASR output could be considered as one word, removing the spaces between the segments (if applicable).

The proposed threshold-based acceptance approach has proven to be valid on most of the manual transcriptions of PWA's speech samples from the UniSt dataset. Thus, it has not worked for conduite d'approche type of error, for 40% of the semantically acceptable phonemic neologisms, and for semantic paraphasia as part of a compound word. The transcriptions of ASR models yield the same results on these samples. Detecting and further analysis of these types of errors is a subject for further research.

On the rest of the UniSt PWA data, the four selected models reach 90.3% acceptance accuracy, with 100% specificity and 89.2% sensitivity, which is one of the top results among the existing applications (see Section 2.2.2). However, more than half of PWA's fully correct answers were recognized as containing a phonemic/phonetic error due to the audio quality and flaws of ASR models. Working on the models' improvement and testing them with a designated device are seen as the next steps.

6. CONCLUSIONS

The paper describes the evaluation of open-source German ASR solutions for further use in a mobile SLT application. As a result, four open-source models have been selected. These models fulfill the suitability requirements for both phoneme-level pronunciation analysis and subsequent semantic and grammatical error analysis. They also outperform commercial models in atypical speech recognition, including audio recordings of low quality.

Using 54% as a normalized error acceptance threshold for orthographic/phonemic transcriptions, analyzing ASR output segment per segment, on the one hand, and as one word with no spaces, on the other, allows reaching promising results in the experiments with aphasic speech data. Improving ASR models' performance (e.g. combining several models in a model ensemble or speaker adaptation techniques), making the approach more sensitive to error types, and implementing and evaluating the whole speech analysis pipeline are foreseen for the following stages of the project.

ANNEX A

Recognition results for 13 open-source models on NA_phrases, NA_words, and NORM_words datasets.

Model	NA_phrases			NA_words				NORM_words			
	CER	HITS	M rank	CER	HITS	empty	M rank	CER	HITS	empty	M rank
andrew	6.0	66.5	11.3	12.2	49.7		5.1	24.6	26.6	1.5	7.7
ims_0	7.3	71.1	12.3	21.8	39.0	1.3	10.3	52.4	21.8	20.3	10.5
ims_35	6.8	77.0	9.1	23.1	41.9	2.5	10.7	61.4	19.9	31.6	11.7
jonatas53	3.9	79.2	5.4	12.9	48.1		5.2	**12.9**	**47.1**		**1.3**
jonatas1b	4.0	82.5	4.3	17.6	56.1	0.1	7.3	31.8	29.7	3.5	8.5
jsnfly	4.3	76.4	7.1	12.2	**58.6**		2.8	19.9	31.2		4.6
marcel	5.1	72.8	9.5	12.0	50.1		4.6	18.6	35.9		3.4
maxidl	5.2	76.1	9.1	13.0	51.6	0.1	6.8	19.3	35.9		3.8
mfleck	4.0	80.3	5.0	**10.7**	**58.7**		**2.3**	15.6	39.3		3.2
nvidia1	**2.3**	**90.2**	**2.0**	43.0	27.8	17.9	12.9	70.2	8.7	60.7	12.7
nvidia2	**2.5**	**89.7**	**1.8**	21.3	57.3	8.1	8.5	56.4	23.3	35.7	11.0
oliver8	4.6	74.2	9.1	11.5	54.9		4.0	17.0	41.3		2.9
oliver9	4.0	80.5	5.1	11.0	56.7		3.0	15.5	43.2	0.1	4.3

CER – character error rate (in percent); empty – empty outputs percentage; M – mean value.

The lowest CER values, the highest ranks and HITS are in bold.

The models selected for the current app after the evaluation are marked with grey.

DATA AVAILABILITY

ALC, CI, and PHONDAT2 corpora were downloaded from BAS CLARIN repository (https://clarin.phonetik.uni-muenchen.de/BASRepository/) under free access for scientists.

AphasiaBank (https://aphasia.talkbank.org/) data was accessed with the permission for research and education purposes.

The samples of UniSt dataset were obtained with the permission for research and education purposes from the Institute for Natural Language Processing of the University of Stuttgart (https://www2.ims.uni-stuttgart.de/sgtutorial/index.html).

ACKNOWLEDGMENTS

Elisabeth Zeuner, Aischa Khader-Lindholz, and Ida Luise Kranzfelder from the Speech Science Department of the Martin-Luther-University Halle-Wittenberg annotated the PWA's data from the YouTube video, and AvEv and UniSt datasets. Elisabeth Zeuner classified the errors of PWA in UniSt dataset.

DECLARATION OF COMPETING INTEREST

The authors of this article declare that they have no financial, professional or personal conflicts of interest that could have inappropriately influenced this work.

FUNDING SOURCES

AphaDIGITAL project is sponsored by German Federal Ministry of Education and Research under funding code 03WIR3108A 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.

AUTHORSHIP CONTRIBUTION STATEMENT

Eugenia Rykova: Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Mathias Walther: Conceptualization, Funding Acquisition, Resources, Project Administration, Supervision, Writing – review & editing.

REFERENCES

Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., and Martins, I. P. (2013). Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language, 27*, 1235–1248.

Adikari, A., Hernandez, N., Alahakoon, D., Rose, M. L., and Pierce, J. E. (2024). From concept to practice: A scoping review of the application of AI to aphasia diagnosis and management. *Disability and Rehabilitation*, *46*(7), 1288-1297.

Alpha Cephei, Inc. (2022). *Vosk speech recognition toolkit models*. Retrieved from https://alphacephei.com/vosk/models. Accessed September 12, 2022.

Aphavox. (2020). *Therapieunterstützung für Aphasiker. Eigentraining mit dem Tablet und Feedback durch Spracherkennung [Therapy support for people with aphasia. Self-training with a tablet and feedback through speech recognition].* https://aphavox.de/media/201029_aphavox_faltblatt_web.pdf

Arias-Vergara, T., Batliner, A., Rader, T., Polterauer, D., Högerle, C., Müller, J. Orozco-Arroyave, J.-R., Nöth, E., and Schuster, M. (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.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., and Auli, M. (2022). XLS-R: self-supervised cross-lingual speech representation learning at scale. *Proceedings of Interspeech 2022*, 2278-2282.

Ballard, K. J., Etter, N. M., Shen, S., Monroe, P., & Tien Tan, C. (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*, 818–834.

Barbera, D. S., Huckvale, M., Fleming, V., Upton, E., Coley-Fisher, H., Doogan, C., and Crinion, J. (2021). NUVA: A Naming Utterance Verifier for Aphasia Treatment. *Computer Speech & Language, 69*, 101221.

Benson, D. F. (1988). Anomia in aphasia. *Aphasiology, 2*(3-4), 229-235.

Bhogal, S. K., Teasell, R., and Speechley, M. (2003). Intensity of aphasia therapy, impact on recovery. *Stroke, 34*, 987-993.

Bischoff, M. (2022). *Wav2Vec2-Large-XLSR-53-German*. Retrieved from https://huggingface.co/marcel/wav2vec2-large-xlsr-53-german. Accessed September 12, 2022.

Brady, M. C., Kelly, H., Godwin, J., and Enderby, P. (2016). Speech and language therapy for aphasia following stroke (Review). *Cochrane Database of Systematic Reviews 2016*(6), CD000425.

Braley, M., Pierce, J. S., Saxena, S., Oliveira, E. D., Taraboanta, L., Anantha, V. and Kiran, S. (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.

Caballero Morales, S. O., and Cox, S. J. (2009). Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing, volume 2009*, 308340.

Chatzoudis, G., Plitsis, M., Stamouli, S., Dimou, A.–L., Katsamanis, A., and Katsouros, V. (2022). Zero-shot cross-lingual aphasia detection using automatic speech recognition. *Proceedings of Interspeech 2022*, 2178-2182.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. *Proceedings of Interspeech 2021*, 2426-2430.

Denisov, P., andVu, N.T. (2019). IMS-speech: A speech to text tool. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, 170-177.

Des Roches, C. A., and Kiran, S. (2017). Technology-based rehabilitation to improve communication after acquired brain injury. *Frontiers in Neuroscience, 11*, 382.

Fleck, M. (2022). *Wav2vec2-large-xls-r-300m-german-with-lm.* Retrieved from https://huggingface.co/mfleck/wav2vec2-large-xls-r-300m-german-with-lm. Accessed September 12, 2022.

Fraser, K., Rudzicz, F., Graham, N., and Rochon, E. (2013). Automatic speech recognition in the diagnosis of primary progressive aphasia. *SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies*, 47–54.

Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., and Tomanek, K. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. *Proceedings of Interspeech 2021*, 4778-4782.

Griffel, J., Leinweber, J., Spelter, B., and Roddam, H. (2019). Patient-centred design of aphasia therapy apps:a scoping review. *Aphasie und verwandte Gebiete | Aphasie et domaines associés, 46*(2), 6-21.

Grosman, J. (2022a). *Fine-tuned XLSR-53 large model for speech recognition in German*. Retrieved from https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german. Accessed September 12, 2022.

Grosman, J. (2022b). *Fine-tuned XLS-R 1B model for speech recognition in German.* Retrieved from https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-german. Accessed September 12, 2022.

Guhr, O. (2022a). *Wav2vec2-large-xlsr-53-german-cv8-dropout.* Retrieved from https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv8. Accessed September 12, 2022.

Guhr, O. (2022b). *wav2vec2-large-xlsr-53-german-cv9.* Retrieved from https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv9. Accessed September 12, 2022.

Gutz, S. E. (2022). *Automatic speech recognition as a clinical tool: Implications for speech assessment and intervention* [Doctoral dissertation]. Harvard University.

Halling, S. (2023). Ein Trainingsprogramm für AphasiepatientInnen: aphavox [Training software for aphasia patients: aphavox]. *Logos 31*(1), 46-48.

Hess, W.J., Kohler, K.J., & Tillmann, H.-G. (1995). The Phondat-verbmobil speech corpus. *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, 863-866.

Hirsch, H.-G., Neumann, C., Tiggelkamp, Y., Fiorista, R., Knecht, S., Schnitzler, A., and Frieg, H. (2023). RehaLingo - towards a speech training system for aphasia. In C. Draxler (Ed.), *Proceedings of the 34th conference Elektronische Sprachsignalverarbeitung* (pp. 134-141). TUDpress.

Hönig, F., & Nöth, E. (2016). Automatische Sprachverarbeitung in der Sprachtherapie [Automatic signal processing in speech and language therapy]. In K. Bilda, J. Mühlhaus, & U. Ritterfeld (Eds.), *Neue Technologien in der Sprachtherapie [New Technologies in Speech and Language Therapy]* (pp. 173-184). Thieme.

Huber, W. (1983). *Aachener aphasie test (AAT) [Aachen Aphasia Test].* Verlag für Psychologie Hogrefe, Göttingen, Zürich.

Hugging Face, Inc. (2022). *The AI community building the future*. Retrieved from https://huggingface.co. Accessed December 12, 2022.

Idahl, M. (2022). *Wav2Vec2-Large-XLSR-53-German*. Retrieved from https://huggingface.co/maxidl/wav2vec2-large-xlsr-german. Accessed September 12, 2022.

Jamal, N., Shanta, S., Mahmud, F., and Sha'abani, M. N. (2017). Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review. *AIP Conference Proceedings, 1883*(1), 020028.

Johnson, L., Nemati, S., Bonilha, L., Rorden, C., Busby, N., Basilakos, A., and Fridriksson, J. (2022). Predictors beyond the lesion: Health and demographic factors associated with aphasia severity. *Cortex, 154*, 375-389.

Jsnfly. (2022). *XLS-R-1b-DE.* Retrieved from https://huggingface.co/jsnfly/wav2vec2-xls-r-1b-de-cv8. Accessed September 12, 2022.

Keshet, J. (2018). Automatic speech recognition: A primer for speech-language pathology researchers. *International Journal of Speech-Language Pathology, 20*, 599–609.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*(C), 326–347.

Kitzing, P., Maier, A., and Åhlander, V. L. (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.

Kohlschein, C., Klischies, D., Meisen, T., Schuller, B. W., and Werner, C. J. (2018). Automatic processing of clinical aphasia data collected during diagnosis sessions: Challenges and prospects. In D. Kokkinakis (Ed.), *Proceedings of the LREC 2018 Workshop "Resources and ProcessIng of linguistic, para-linguistic andextra-linguistic Data from people with various forms of cognitive/ psychiatric impairments (RaPID-2)"* (pp. 11-18).

Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., and Castonguay, P. (2019). NeMo: a toolkit for building AI applications using Neural Modules. *ArXiv*, abs/1909.09577.

Le, D., Licata, K., Persad, C., and Provost, E.M. (2016). Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*, 2187-2199.

Le, D., Licata, K., and Provost, E.M. (2017). Automatic paraphasia detection from aphasic speech: A preliminary study. *Proceedings of Interspeech 2017*, 294-298.

Le, D., & Provost, E. M. (2016). Improving automatic recognition of aphasic speech with AphasiaBank. *Proceedings of Interspeech 2016*, 2681-2685.

Lee, T., Liu, Y., Huang, P.-W., Chien, J.-T., Lam, W. K., Yeung, Y. T., and Law, S.-P. (2016). Automatic speech recognition for acoustical analysis and assessment of aCantonese pathological voice and speech. *Proceedings of ICASSP 2016*, 6475-6479. IEEE.

Lin, Y., Klumpp, P., Pfab, J., Abdelioua, A., Gebray, D., and Späth, M. (2022, April). *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.

LingoLab. (2020). *Digitale Lösungen für die Sprachtherapie [Digital solutions for speech and language therapy]*. Retrieved from https://lingo-lab.de/. Accessed June 7, 2023.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology, 25*(11), 1286–1307.

McDowell, A. (2022). *A fine-tuned version of facebook/wav2vec2-xls-r-1b on the MOZILLA-FOUNDATION/COMMON_VOICE_8_0-DE dataset.* Retrieved from https://huggingface.co/AndrewMcDowell/wav2vec2-xls-r-1B-german. Accessed September 12, 2022.

Neolexon. (2023). *Logopädie-Apps [Speech and language pathology apps]*. Retrieved from https://neolexon.de/. Accessed June 7, 2023.

Netzebandt, J., Schmitz-Antonischki, D., and Heide, J. (2022). Hochfrequente Wortabruftherapie mit LingoTalk [High-frequency word retrieval therapy with LingoTalk]. *forum:logopädie 36*(3), 18-24.

Neumeyer, V. (2009). *Phonetische Untersuchungender Artikulation von CI-Trägern [Phonetic studies of CI users' articulation]* [Master's thesis]. Ludwig-Maximilians-Universität München.

NVIDIA. (2022a). *NVIDIA Conformer-CTC Large (de).* Retrieved from https://huggingface.co/nvidia/stt_de_conformer_ctc_large. Accessed September 12, 2022.

NVIDIA. (2022b). *NVIDIA Conformer-Transducer Large (de)*. Retrieved from https://huggingface.co/nvidia/stt_de_conformer_transducer_large. Accessed September 12, 2022.

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

Pompili, A., Abad, A., Trancoso, I., Fonseca, J., and Martins, I. P. (2020). Evaluation and extensions and of an automatic and speech therapy and platform. In P. Quaresma, R. Vieira, S. Aluísio, H. Moniz, F. Batista, & T. Gonçalves (Eds.), *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Portugal* (pp. 43–52). Springer International Publishing.

Pompili, A., Abad, A., Trancoso, I., Fonseca, J., Martins, I. P., Leal, G., and Farrajota, L. (2011). An on-line system for remote treatment of aphasia. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, 1-10.

Python Software Foundation. (2022). *JiWER: Similarity measures for automatic speech recognition evaluation.* Retrieved from https://pypi.org/project/jiwer/ . Accessed December 15, 2022.

Qin, Y., Lee, T., Feng, S., and Hin Kong, A. P. (2018). Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning. *Proceedings of Interspeech 2018*, 3418-3422.

Qin, Y., Lee, T., and Hin Kong, A. P. (2020). Automatic Assessment of Speech Impairment in Cantonese-Speaking People with Aphasia. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 331-345.

Qualls, C. D. (2011). Neurogenic disorders of speech, language, cognition-communication, and swallowing. In D. E. Battle (Ed.), *Communication Disorders in Multicultural and International Populations* (pp. 148–163). Mosby.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, 28492–28518. JMLR.org

RehaLingo. (2023). *Language rehabilitation software for the 21st Century.* Retrieved from https://www.rehalingo.com Accessed June 7, 2023.

Rhein-Zeitung. (2018). *Am Anfang war das Wort: Zu Besuch bei einem Aphasiker [In the beginning was the word: Visiting a person with aphasia]*. Retrieved from https://www.youtube.com/watch?v=Z1ZglYMSx1Y. Accessed May 16, 2022.

Ruff, S., Bocklet, T., Nöth, E., Müller, J., Hoster, E., and Schuster, M. (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 Specialities*, *79*(5), 282-294.

Ryalls, J. (1984). Where does the term "aphasia" come from? *Brain and Language, 21*, 358-363.

Rykova, E., and Walther, M. (2024a). AphaDIGITAL – digital speech therapy solution for aphasia patients with automatic feedback provided by a virtual assistant. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS),* 3385-3394.

Rykova, E., and Walther, M. (2024b). Linguistic and extralinguistic factors in automatic speech recognition of German atypical speech. *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, 358–367. Association for Computational Linguistics.

Rykova, E., Walther, M., and Zeuner, E. (2022, Au-gust). *AphaDIGITAL — Avatar-based digital speech therapy solution for aphasia patients: first evaluation.* Poster session presentation at the 35th Fonetiikan Päivät, Joensuu, Finland. Available at https://www.researchgate.net/publication/364676432_aphaDIGITAL_-Avatar-based_digital_speech_therapy_solution_for_aphasia_patients_first_evaluation

Schäfer, C., Oliznyk, O., and Haller, M. (2023). *Deep Phonemizer: a G2P library in PyTorch.* Retrieved from https://github.com/as-ideas/DeepPhonemizer. Accessed May 17, 2023.

Schiel, F., Heinrich, C., Barfüsser, S., and Gilg, T. (2008). ALC — Alcohol Language Corpus. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08),* 1641-1645.

Schneider, B., Wehmeyer, M., and Grötzbach, H. (2021). Aphasische Symptome und Syndrome [Aphasia symptoms and syndroms]. In B. Schneider, M. Wehmeyer, & H. Grötzbach (Eds.), *Aphasie: ICF-orientierte Diagnostik und Therapie [Aphasia: ICF-based diagnostics and therapy]* (pp. 25-56). Praxiswissen Logopädie · Monika Maria Thiel · Mascha Wanke · Susanne Weber.

Schulz, J. B., and Werner, C. J. (2019). *Statistischer Jahresbericht 2018. [Year 2018 statistics report].* Aphasiestation, Klinik für Neurologie, Uniklinik RWTH Aachen, Germany.

TDG — Translationsregion für digitale Gesundheits-versorgung [Translational Region for Digital Healthcare]. (2021). AphaDIGITAL: Entwicklung einer digitalen, dezentralen sprachtherapeuti-schen Versorgung [Development of digital, de-centralized speech therapy solutions]. Retrieved from https://inno-tdg.de/projekte/aphadigital/. Accessed January 25, 2023.

Tisljár-Szabó, E., Rossu, R., Varga, V., and Pléh, C. (2014). The effect of alcohol on speech production. *Journal of Psycholinguistic Research, 43*, 737–748.

Torre, I. G., Romero, M., and Álvarez, A. (2021). Improving aphasic speech recognition by using novel semi-supervised learning methods on AphasiaBank for English and Spanish. *Applied Sciences, 11*(19), 8872.

Universität Stuttgart. (2023). *Sprache und Gehirn: Ein neurolinguistisches Tutorial [Language and brain: a neurolinguistics tutorial].* Retrieved from https://www2.ims.uni-stuttgart.de/sgtutorial/index.html. Accessed June 17, 2023.

Vaezipour, A., Campbell, J., Theodoros, D., and Russell, T. (2020). Mobile apps for speech-language therapy in adults with communication disorders: review of content and quality. *JMIR mHealth and uHealth*, *8*(10), e18858.

van de Sandt-Koenderman, W. M. (2011). Aphasia rehabilitation and the role of computer technology: Can we keep up with modern times? *International Journal of Speech-Language Pathology, 13*, 21-27.

Vipperla, R., Renals, S., and Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. *Proceedings of Interspeech 2008*, 2550–2553.

Wambaugh, J. L., Doyle, P. J., Kalinyak, M. M., and West, J. E. (1996). A critical review of acoustic analyses of aphasic and-or apraxic speech. *Clinical Aphasiology, 24*, 35-63.

Wiehage, A., & Heide, J. (2016). *Aphasie: Informationen für Betroffene und Angehörige [Aphasia: information for the affected and relatives].* Retrieved from https://www.dbs-ev.de/fileadmin/dokumente/Publikationen/dbs-Broschuere_Aphasie_2016.pdf

Wirth, J., and Peinl, R. (2022). ASR in German - a detailed error analysis. *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, 1-8. IEEE.

Xu, J., Matta, K., Islam, S., and Nürnberger, A. (2020). German speech recognition system using DeepSpeech. *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 102-106.

Zeuner, E., Pietschmann, J., Voigt-Zimmermann, S., Rykova, E., and Walther, M. (2022, September). *aphaDIGITAL - Avatar-gestützte digitale Aphasietherapie: Evaluation [aphaDIGITAL: Avatar-supported digital aphasia therapy - evaluation study].* Poster session presentation at DGSS Annual Conference 2022, Stimme und Geschlecht im Wandel' – Implikationen für Theorie und Praxis in der Sprechwissenschaft und Phonetik ["Voice and Gender in Transition" – Implications for Theory and Practice in Speech Science and Phonetics], Jena, Germany. Available at https://www.researchgate.net/publication/371510421_aphaDIGITAL_-Avatar-gestutzte_digitale_Aphasietherapie_Evaluation.