

Enhancing ASR-based error analysis for atypical speech: post hoc integration of aphasic and dialectal phonetic features

Abstract

Despite current progress, automatic speech recognition (ASR) often struggles with non-standard speech, for example, influenced by dialectal or pathological features. (Re)training ASR models to accommodate these variations is not always possible due to limited data. This paper proposes applying the knowledge about non-standard (aphasic and dialectal) phonetic features to the ASR transcription post hoc. Using speech data from German speakers with aphasia who speak the Thuringian-Upper Saxon dialect, this study evaluates the impact of these modifications on an ASR-based error analysis pipeline. The approach helps to reduce automatic error rates on the recordings manually labelled as error-free. The performance of the pipeline also improves both in general acceptance or rejection of the responses and error attribution. General acceptance/rejection accuracy reaches the mean of 83.3%, which is considered sufficient to be used in a digital application for speech and language therapy support.

1 Introduction

Digital transformation has led to significant advancements in healthcare, particularly through the integration of artificial intelligence (AI). In the field of speech and language therapy (SLT), AI-driven tools are increasingly used for diagnosing and supporting people with aphasia (PWA) (Adikari, Hernandez, Alahakoon, Rose, & Pierce, 2024; Azevedo et al., 2024; Pottinger & Kearns, 2024). This paper is part of a broader project, *aphaDIGITAL*¹, aimed at developing a mobile application for German-speaking PWA. The project is regional and focuses on Southern Saxony-Anhalt and bordering regions of Saxony and Thuringia.

The app is to provide detailed, personalised feedback in oral exercises using automatic speech recognition (ASR) and subsequent automatic text processing, united into an error analysis pipeline (see Subsection 2.1). Currently, this pipeline includes ASR as well as phonemic/phonetic and semantic analysis components. This paper presents our work on the ASR component of the pipeline, specifically a solution to mitigate the effects of aphasic and dialectal features in the users speech on its automatic recognition, which is performed with selected open-source models.

1.1 ASR in aphasia speech therapy

Aphasia is an acquired language disorder caused by focal brain damage. Although there are several characteristic clinical profiles of the disorder, some linguistic symptoms are particularly salient

¹<https://aphadigital.sprechwiss.uni-halle.de/>

and widespread among PWA. Anomia, or difficulty in word retrieval, is one of the most common deficits Benson (1988), and it is typically treated with naming-oriented semantic exercises.

Naming-oriented semantic exercises have been automated with the help of ASR for Portuguese (Abad et al., 2013; Pompili et al., 2011), Australian English (Ballard, Etter, Shen, Monroe, & Tien Tan, 2019), British English (Barbera et al., 2021), and German (Heide et al., 2023; Hirsch, Tiggelkamp, Neumann, Frieg, & Knecht, 2025; Lin et al., 2022). In all these applications, the feedback is binary – correct or incorrect – without deeper analysis of the users speech errors. Reported accuracy rates of accepting or rejecting PWA’s responses in an oral naming exercise are above the recommended accuracy threshold of 80% for use in SLT (McKechnie et al., 2018) – except for the case described by Ballard et al. (2019). They test the application with the participants suffering from apraxia of speech, comorbid to aphasia, and the application reaches only 75% accuracy. Heide et al. (2023) do not report results on aphasic speech, although the use of ASR is mentioned on the app’s website². The development of the ASR technology for the Neolexon Aphasia-App (Lin et al., 2022) is thoroughly described in Klumpp (2024), but according to the information on the app’s website³, the technology has not yet been implemented. The system developed by Hirsch et al. (2025) is still in the research stage.

Modern ASR systems present distinctive results on typical speech. For example, a word error rate (WER) for English reaches 0.05 (Protalinski, 2017; Xiong et al., 2017) and lower in certain cases (Graham & Roll, 2024; Picovoice, 2025). When tested on a range of datasets, German ASR solutions demonstrate similar results (Picovoice, 2025, see also Wirth & Peinl, 2022). However, their performance on impaired speech in general, and PWA’s speech in particular, remains unreliable (Green et al., 2021; Rykova & Walther, 2025). Aphasic speech is characterised by linguistic searching behaviour and self-corrections, hesitation phenomena with or without vocal utterances, phonemic structure distortions, imprecise articulation or hyperarticulation. The irregularities are often inconsistent and unpredictable, which makes them hard to model – in distinction to motor speech disorders (e.g. dysarthria; see Caballero Morales & Cox, 2009; Mulfari, Carnevale, & Villari, 2023). For the current application, these challenges are compounded by dialectal variations (see Fischer & Jäck, 2023; Pompili et al., 2011), which already present difficulties for ASR systems even with German speakers without speech disorders (Költzsch, 2024). The present paper addresses this challenge in the context of the aforementioned project.

1.2 ASR of dialects

Non-standard language varieties pose significant challenges for the development of specialised ASR systems due to the absence of a standardised written form, internal variation, and the limited availability of training data. WER values of 0.2 and higher are common in dialectal ASR. Evaluation of dialectal speech recognition systems is also not trivial due to the absence of a common benchmark (Hinsvark et al., 2021).

Research has been carried out in several languages, including the dialects of American English (Harris, Mgbahurike, Kumar, & Yang, 2024), Arabic (for a review, see Alsayadi, Abdelhamid, Hegazy, Alotaibi, & Fayed, 2022), Chinese (for a review, see Li, Mai, Wang, & Ma, 2024),

²<https://lingo-lab.de>

³<https://neolexon.de/>

Thai (Suwanbandit, Naowarat, Sangpetch, & Chuangsuwanich, 2023), and German varieties and dialects (e.g., Bystrich, 2023). Hinsvark et al. (2021) discuss promising approaches to dialectal speech recognition, focusing on the challenge of phonetic variation within a language (which the authors refer to as "accented speech"). Phonetic variation consistently leads to degraded ASR performance, even in the absence of dialect-specific vocabulary (Harris et al., 2024).

To improve ASR performance on dialectal speech, it is common to train the corresponding models from scratch with deep learning or using more traditional methods like Gaussian mixture models and hidden Markov models, which can achieve a WER below 0.19 (see Nigmatulina, Kew, & Samardžić, 2020). Another solution involves applying transfer learning or fine-tuning to models pretrained on a standard variety (Bystrich, 2023; Elmahdy, Hasegawa-Johnson, & Mustafawi, 2014; Harris et al., 2024; Suwanbandit et al., 2023). However, both methods require a considerable amount of audio data and corresponding transcripts.

One possible way to circumvent the obstacle of limited audio data is to extend the pronunciation dictionary of the ASR system, which maps orthographic words to their pronunciation(s). Masmoudi, Khmekhem, Estève, Belguith, and Habash (2014) formulate rules for Tunisian Arabic dialect patterns and apply these rules to generate such extensions, achieving a WER of 0.23 with a trained-from-scratch ASR model. Bystrich (2023) explores phonetic mapping rules to supplement the phonetic lexicon of a Standard German ASR system with new Bavarian dialectal pronunciations for existing words and transcribe the spoken dialect into orthographic Standard German (cf. Nigmatulina et al., 2020 and Ulasik et al., 2021). Bystrich (2023) finds that the mapping of some phones is more effective than others, while some mappings have a negative impact.

Ali and colleagues proposes mining non-standard spellings of Egyptian Arabic from social media and applying them to dialect ASR in two ways. One method involves including the non-standard spellings into the language model, which reduced the WER of a trained-from-scratch Egyptian Arabic ASR model from 0.6 to 0.45 (Ali, Mubarak, & Vogel, 2014). The other method uses the non-standard spellings as targets in the evaluation of ASR performance (Ali, Nakov, Bell, & Renals, 2017). The authors call the resulting metric WER for dialects (WERd) and argue that WERd is a more appropriate metric for evaluating dialect ASR. Evaluated on a corpus of Egyptian dialect, an Arabic ASR system achieves a mean WERd of 0.36 compared to a WER of 0.46.

Swiss German

Swiss German – a continuum of Swiss German dialects, to be more precise – differs significantly from Standard German (of Germany). There are variations in phonology, lexicon, morphology and syntax (Scherrer & Rambow, 2010). Furthermore, the dialects differ from each other, primarily in phonology and vocabulary. Swiss German dialects are mainly used for oral and informal written communication and do not have a standardised orthography, while Standard German is used in more rather written contexts (Plüss et al., 2023). This makes automatic speech processing for Swiss German especially challenging. Additionally, publicly available annotated data remain scarce. STT4SG-350, a large corpus of Swiss German speech recordings transcribed into Standard German and balanced across dialectal regions was released only recently (Plüss et al., 2023).

There are other corpora, smaller and less balanced, from which some are transcribed in Swiss German (e.g., Samardžić, Scherrer, & Glaser, 2016; see also Plüss et al., 2023).

Researchers have explored different approaches to ASR for Swiss German dialects. Nigmatulina et al. (2020) develop the first multi-dialectal Swiss German ASR framework based on ArchiMob corpus (Samardžić et al., 2016) and experiment with two types of output: a dialectal phonemic spelling and a normalised transcription closely resembling Standard German. For ASR systems with dialectal output, the authors also introduce an alternative evaluation metric, FlexWER. It generally follows the idea of WERd (Ali et al., 2017), but makes use of the word-level mapping between the dialectal and the normalised representations: an output word is considered correct if it maps to the same normalised transcription as the corresponding word in the ground truth. The lowest WER score of 0.29 is achieved by an ASR system with normalised output. However, a corresponding ASR system with dialectal output achieves a lower FlexWER score (0.21) and also performs slightly better in terms of character error rate (CER): the lowest score is 0.146, with a difference of up to 0.015.

Ulasik et al. (2021) focus on the transcription of Swiss German speech into text in Standard German, treating the task as speech translation (ST). ST combines ASR and machine translation: spoken utterances in a source language are translated into a target language, either as text or speech. The translation can be performed stepwise – first ASR in the source language, then automatic translation into the target language (cascaded approach) – or with the input in the source language mapped directly to the output in the target language (end-to-end approach). Since the authors use the test corpus of Swiss German transcribed into Standard German (Plüss, Neukom, & Vogel, 2021), they employ an end-to-end ST approach for their experiments. BLEU (Bilingual Evaluation Understudy), a standard metric for automatic evaluation of machine translation, is used to assess model performance. BLEU scores range from 0 to 100, with higher scores indicating better performance.

Ulasik et al. (2021) train three types of models using various methods and corpora, select the best model of each type, and combine them into a single system, applying an ensemble method of majority voting both with and without perplexity weighting. The three models ensembled with simple majority voting achieve the highest BLEU score of 38.7 (cf. mean BLEU score of 75 achieved by the model fine-tuned with a large STT4SG-350 corpus in Plüss et al., 2023). Ulasik et al. (2021) experiment with a supervised spelling-correction model that identifies likely errors in the automatic transcriptions and suggests corrections. However, the four tested spelling-correction algorithms decrease overall system performance.

Finally, Kresic and Abbas (2024) present a method to normalise Swiss German written speech to Standard German, fine-tuning a multilingual large language model (mT5). The authors suggest integrating this method with speech recognition for increasing the effectiveness of the latter.

1.3 Research questions

In the absence of adequate data for (re)training ASR models, the current research is based on applying knowledge about non-standard (aphasic and dialectal) phonetic features post hoc to the ASR output. The implementation of the dialectal features adapts the method of using spelling alternatives for standard ASR evaluation introduced by Ali et al. (2017). In the present study,

dialectal pronunciation alternatives are generated based on non-standard pronunciation rules, similarly to Masmoudi et al. (2014) and Bystrich (2023). The main research questions are as follows:

- *Can a post hoc implementation of non-standard speech features improve the performance of an ASR-based error analysis pipeline?*
- *What are the ways to improve automatic error attribution?*

2 Materials & Methods

This work introduces modifications to the initial error analysis pipeline described in Rykova and Walther (2024a). The modified pipeline, which includes a module for post hoc phonetic features, is tested using existing speech recordings from PWA.

2.1 Error analysis pipeline

The original error analysis pipeline for naming exercises was presented in Rykova and Walther (2024a). In a naming exercise, the user is presented with a picture or text stimulus and is expected to utter a target word or phrase as a response. In the presented app, the user's response is transcribed with the ASR. The ASR output is then compared to the target transcription, and an error rate (ER) threshold is applied. If $ER = 0$, the response is error-free. The error is considered phonemic/phonetic if $0 < ER < ER \text{ threshold}$, and semantic if $ER \geq ER \text{ threshold}$.

ER is defined as the minimum of the character error rate (CER) between the orthographic forms of the ASR output and the target, and the phoneme error rate (PER) between their respective phonemic transcriptions. These transcriptions are generated automatically using the grapheme-to-phoneme (g2p) converter *bootphon espeak phonemizer* (Bernard & Titeux, 2021). Both CER and PER are normalised to range from 0 to 1. Based on previous findings (Rykova & Walther, 2025) and discussions with speech therapists, the ER threshold was set to 0.5.

For example:

- Target: *Gürtel* /gyʁtəl/ ('belt')
ASR output: "schnalle" /ʃnalə/ ('buckle')
CER = 0.875, PER = 0.834 → ER = 0.834
ER > 0.5 → **semantic error**
- Target: *Gürtel* /gyʁtəl/ ('belt')
ASR output: "gürtelm" /gyʁtəlm/
CER = 0.143, PER = 0.286 → ER = 0.143
ER < 0.5 → **phonemic/phonetic error**

Four ASR models (Fleck, 2022; Grosman, 2021; Guhr, 2022; NVIDIA, 2022) previously shown to be suitable for aphasic speech (Rykova & Walther, 2025) were used. In one-word recognition tasks involving atypical speech, the lowest mean CER among these models is 0.11, observed in speech from alcohol-intoxicated speakers.

The pipeline evaluates each ASR models output both as a whole and as separate words, removes duplicates, and selects the output with the lowest ER for further analysis and feedback (cf. ensembled use of ASR models in Ulasik et al., 2021). Phonemic/phonetic errors are analysed with a phonemic/phonetic analysis component, and semantic errors are analysed with a semantic analysis component. The user has three attempts to produce the target correctly. After the first two attempts, they receive detailed feedback based on the error type. After the third attempt, the pipeline only checks whether the answer is correct (error-free) or incorrect, regardless of the error type.

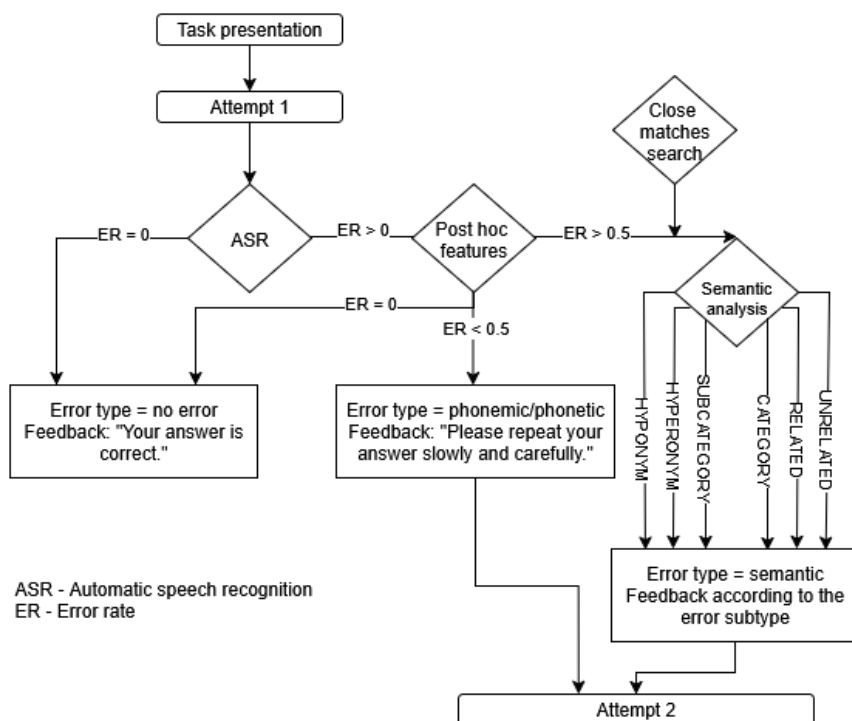


Figure 1: Modified error analysis pipeline section for attempt 1. Adapted from Rykova and Walther (2024a), p. 3387.

Upon internal discussion with a project team, the original Rykova and Walther's (2024a) pipeline has been modified:

- The ER threshold remains at 0.5, but if a half of the target is produced correctly, it is now considered a phonemic/phonetic error (previously – semantic error): phonemic/phonetic errors correspond to $0 < ER \leq 0.5$, and semantic errors correspond to $ER > 0.5$ (cf. Klumpp, 2024 and Redrovan-Reyes et al., 2019).
- Detailed analysis of phonemic/phonetic errors, previously executed after the first two attempts, is now removed after the first attempt. Phonemic/phonetic errors are analysed in details after the second attempt only.

This paper presents a new post hoc features component to be executed in every attempt after the ASR component when $ER > 0$. Figure 1 illustrates the modified pipeline for the first attempt. In a modified version, if the $ER > 0$ after the first comparison of the ASR output to the target (performed within the ASR component), non-standard features are introduced to the ASR output. The ER is then recalculated and, if it is greater than zero (the response is not error-free), compared to the threshold to determine the error type. If $0 < ER \leq 0.5$, the user is asked to repeat their response, thus initiating the second attempt. If $ER > 0.5$, the response is subject to semantic analysis with the corresponding pipeline component(s). For details on semantic analysis, please see Rykova and Walther (2023).

2.2 Post hoc features

For the research purposes, the analysis of test recordings was conducted in three steps, uniting aphasic features under step 2:

- **Step 1:** Original ASR output.
- **Step 2:** Implementation of aphasic speech features:
 - Syllabification and hyperarticulation: long pauses between syllables may cause them to be recognised as separate words. In such cases, spaces between ASR output segments are removed (as shown effective in Rykova & Walther, 2025);
 - Vowel prolongation: slow, careful speech may result in prolonged vowels. The vowel length symbol is removed, and vowel quality is adjusted (e.g. /i:/ → /i/ – if such changes help to reduce PER).
- **Step 3:** Implementation of dialectal features: according to the project's geographic focus and available data, features from the Thuringian-Upper Saxon dialect group were selected (Rocholl, 2015; Siebenhaar, January, 2024; Wallraff, 2007). Based on these features, phonemic transformation rules were formulated. These rules of phonemic changes were applied to the target phonemic transcription in order to generate alternative targets and compare the phonemic transcription of the ASR output to them.

Table 1: Steps of non-standard phonetic features implementation.

Step	Description	Example
1	Separate words from ASR output	Target: <i>Buch</i> ASR outputs: ein buch, en woch, von wo Result: buch is accepted as an error-free answer
2	Aphasic features: Removing whitespaces Reducing vowel length (removing vowel length symbol)	Target: <i>Schuhanzieher</i> ASR output: schuh anzieher, CER = 0.07 Output corrected: schuhanzieher, CER = 0 Result: the answer accepted as error-free Target: <i>Zigarre</i> /tsi:garə/ ASR output: zigare /tsi:garə/, PER = 0.125 Output corrected: /tsi:garə/, PER = 0 Result: the answer is accepted as error-free
3	Dialectal features: Generation of alternative target transcriptions based on phonemic transformation rules	Target: <i>Haustür</i> /haʊstʏ:r/ ASR output: haus dir /haʊs di:r/ Output corrected (step 2): hausdir /haʊsdi:r/ (Vowel length is not reduced at step 2 because it does not bring a reduction in PER) Alternative target generated based on rules t→d and y:→i: /haʊsdi:r/ Result: the answer is accepted as error-free

Due to technical reasons, in the app backend, removing whitespaces is implemented at the very first stage of the ASR output evaluation, corresponding to the ASR component in Figure 1 (see Rykova & Walther, 2025). Vowel prolongation and dialect features are addressed successively in the post hoc features component. Examples of the post hoc features implementation are provided in Table 1. When CER-based analysis suffices, only orthographic forms are shown. For PER-based analysis, automatic g2p transcriptions are included (please note that these transcriptions are intrinsically imperfect).

A full list of dialect-based phonemic transformation rules is provided in the Appendix. To address limitations of the g2p algorithm (Bernard & Titeux, 2021), which is constrained to Standard German phonemes, some dialectal features were approximated using existing phonemes or combinations (e.g., single consonants for long consonants, /o:/ for /ɔ:/). Additionally, all sounds that correspond to an orthographic *r* were unified as /r/ to avoid mismatches between the target and ASR output transcriptions caused by g2p inconsistencies.

2.3 Test recordings

A total of 412 audio recordings obtained from PWA during the Aachen Aphasia Test (AAT) (Huber, 1993) served as the test material. The recordings were provided by the Clinic of Cognitive

Neurology, University of Leipzig Medical Center. Access to the data was granted under an ethical agreement for research purposes. The dataset includes 219 recordings from the repetition task and 193 from the naming task. The naming task was performed by ten speakers (three female), seven of whom (two female) also completed the repetition task. All speakers were from Saxony and Saxony-Anhalt. No additional information about the participants was available.

All recordings were converted from mp3 to WAV format and resampled to 16 kHz to meet the requirements of the ASR models. The test material was extracted from longer recordings of the test sessions. The following (sub)tasks were selected for analysis:

- Naming task:
 - objects – one-stem nouns (e.g., *Gürtel* 'belt', *Zigarre* 'cigar')
 - objects – multi-stem nouns (e.g., *Staubsauger* 'vacuum cleaner', *Kühlschrank* 'refrigerator')
 - colours – adjectives (e.g., *rot* 'red', *orange* 'orange')
 - situations – sentences (e.g., *Hund* 'dog', *Pfeife* 'pipe')
- Repetition task:
 - monosyllabic words (e.g., *Ast* 'branch', *Glas* 'glass')
 - loan- and foreign words (e.g., *Telefon* 'phone', *Schokolade* 'chocolate')
 - compound words (including compounds with lexical morphemes)
(e.g., *Kraftfahrzeugschein* 'vehicle license', *Verantwortungslosigkeit* 'irresponsibility')

For the naming subtask 'situations – sentences', several individual words were selected from the expected multi-word responses. Although the other subtasks required single-word responses, speakers occasionally produced multi-word utterances (e.g., inserting an article before the target word).

The AAT naming task is used to assess word retrieval and verbal expression. It consists of words with one to five syllables long. The repetition task is used to assess auditory processing and speech motor control. It contains words of complex phonemic structures, one to eight syllables long. It also includes less frequent words than the naming task.

The extracted recordings were orthographically transcribed by the second author of the paper. Based on these transcriptions, the number of syllables in each of the utterances was determined. Using this number of syllables and the duration of the utterance, speech rate of a particular participant was calculated in syllables per second (syl/s).

2.4 Pipeline's performance analysis

The recordings were analysed auditorily by the first two authors of this paper and labelled as containing either no errors, phonemic/phonetic errors, or semantic errors. The second author, an SLT specialist, applied principles of SLT and dialect knowledge, while the first author, a clinical linguist, took into account ASR performance and technical considerations (cf. Klumpp, 2024). The final label was determined by consensus among the annotators (cf. Saz et al., 2009).

The label no errors was given to those utterances that would be fully accepted by an SLT specialist. These utterances might still contain false starts, self-corrections, syllabification, and dialectal features. The label "phonemic/phonetic error" refers to deviations in pronunciation that still render the word recognisable. The label "semantic error" refers to responses that are either unrelated to the target (including incorrect words or non-words) or are incomprehensible. It is important to note that the repetition task recordings contained very few semantic errors, as most participants were able to repeat the target words to some extent.

For the purpose of evaluating the pipeline's performance in terms of error rates, only the 249 recordings (120 from the naming task and 129 from the repetition task) that were labelled as error-free were used. All 412 recordings were used for analysing error attribution.

2.5 Statistical analysis

All statistical analyses were conducted in R (R Core Team, 2023) at a 95% confidence level. Error rates were analysed using the Students t-test, or the Wilcoxon Rank Sum Test when the sample size was too small for parametric testing.

Significance in the pipeline's accuracy changes was tested with the help of McNemar's Chi-squared test, designed for paired nominal data (McNemar, 1947). An exact binomial test (Hollander, Wolfe, & Chicken, 2014) was used to verify whether the results were statistically above the chance level.

3 Results & Discussion

This section presents the results of testing the error analysis pipeline using recordings from people with aphasia. It discusses the automatic error rates (ERs) on error-free recordings, the general accuracy of the pipeline in accepting or rejecting responses, its performance in attributing errors and disadvantages of different error misclassification scenarios. It also examines individual differences observed in the selected participants' data.

3.1 Error rates

A series of t-tests revealed no statistically significant differences in ERs between the naming and repetition tasks when analysing error-free recordings:

- step 1: $t(232) = 0.63, p = 0.532$
- step 2: $t(230) = 0.77, p = 0.442$
- step 3: $t(233) = 0.86, p = 0.388$

As a result, the two subsets were merged for further analysis. The ER decreased progressively across the three steps, as shown in Figure 2. Paired t-tests confirmed that these reductions were statistically significant:

- step 1 vs. step 2: $t(248) = 4.77, p < 0.01$

- step 2 vs. step 3: $t(248) = 3.76, p < 0.01$
- step 1 vs. step 3: $t(248) = 5.87, p < 0.01$

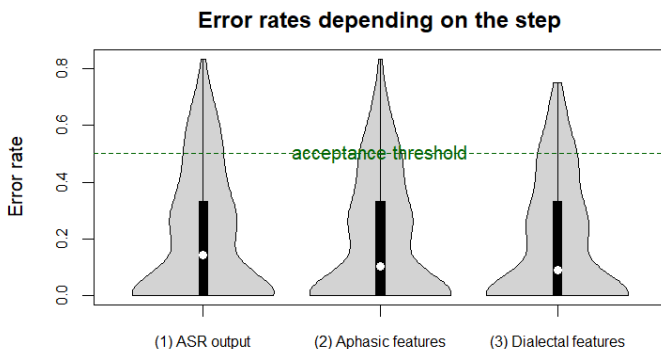


Figure 2: Decrease of the error rates on error-free recordings after non-standard features implementation.

Table 2 presents the changes in the central tendency and data dispersion of the ERs occurred at each step of the analysis.

Table 2: Changes in the error rates' central tendency and data dispersion after non-standard features implementation.

	measure	step 1	step 2	step 3
	mean	0.2	0.192	0.184
	median	0.143	0.105	0.091
	standard deviation	0.225	0.218	0.211
	range	0-0.83	0-0.83	0-0.75
	interquartile range	0-0.33	0-0.33	0-0.33
	percentage of the values equal to zero	44.18%	44.98%	45.38%

While the improvements in ERs on error-free recordings were statistically significant, the absolute differences in average ER values remained modest. The post hoc implementation of aphasic features contributed more to reducing ERs to zero than dialectal features. However, both types of features contributed equally to reducing the mean ER values and their spread in relation to the mean (measured with standard deviation).

It is worth noting that the quality of the original recordings had a substantial impact on ASR performance. The data had been collected for different purposes and stored in mp3 format. In some cases, poor audio quality rendered the post hoc corrections ineffective (e.g., reducing an ER from

0.83 to 0.75 still results in rejecting the response as semantically incorrect). The proposed method shows potential in reducing the impact of aphasia and dialectal variations on ASR performance, but improving the ASR models and the quality of input audio remains essential.

3.2 Pipeline's general accuracy

Evaluated solely on whether users' responses are semantically accepted ($ER \leq 0.5$) or rejected ($ER > 0.5$) – without considering phonemic/phonetic errors (cf. Klumpp, 2024) – the pipeline's accuracy improved at each processing step: 78.2% → 81.1% → 83.3%. Results of the McNemar's Chi-squared test demonstrated statistical significance of the improvements:

- step 1 vs. step 2: $\chi^2(1) = 10.08$, $p = 0.001$
- step 2 vs. step 3: $\chi^2(1) = 7.11$, $p < 0.001$
- step 1 vs. step 3: $\chi^2(1) = 19.05$, $p = 0.008$

Precision for both accepted and rejected responses, as well as recall for accepted responses, increased from step to step. The recall for rejected responses remained unchanged. At the final step, the pipeline correctly identified 83.9% of semantically correct responses and 77.8% of semantically incorrect responses.

3.3 Pipeline's error attribution

Overall pipeline's accuracy in error attribution increased from 50.5% after step 1 to 53.9% after step 3, which is statistically significant according to the McNemar's Chi-squared test: $\chi^2(1) = 12.07$, $p < 0.001$. The accuracy of 53.9% is above chance level of 33.3%: $p < 0.001$ in an exact binomial test.

Due to the nature of the tasks, semantic errors are rare in the repetition task, while phonemic/phonetic errors occur more than twice as often in the naming task. According to the manual labelling of the given data, there were 6 semantic errors and 84 phonemic/phonetic errors in the repetition task, and 39 semantic errors and 34 phonemic/phonetic in the naming task. Because of these differing error distributions, the pipeline's error attribution was analysed in detail separately for each task – see Figure 3.

In the error attribution, the pipeline achieved accuracy score of 58% in the naming task and 50.2% in the repetition task, which are both above the chance level (p -values < 0.001 in an exact binomial test). There are not enough discordant pairs in each task to verify the statistical significance of the differences in accuracy scores between the steps.

In the confusion matrices in Figure 3, one can notice that there are not that many individual changes from step to step, and general picture remains similar. Numerically the greatest changes occur in the recall of phonemic/phonetic errors in the naming task: it increases from 55.9% after step 1 to 67.6% after step 3. The increase in the recall of phonemic/phonetic errors in the repetition task from 52.4% after step 1 to 60.7% after step 3, and the increase in the precision of identifying semantic errors in the naming task from 50.8% after step 1 to 58.5% after step 3 can be also

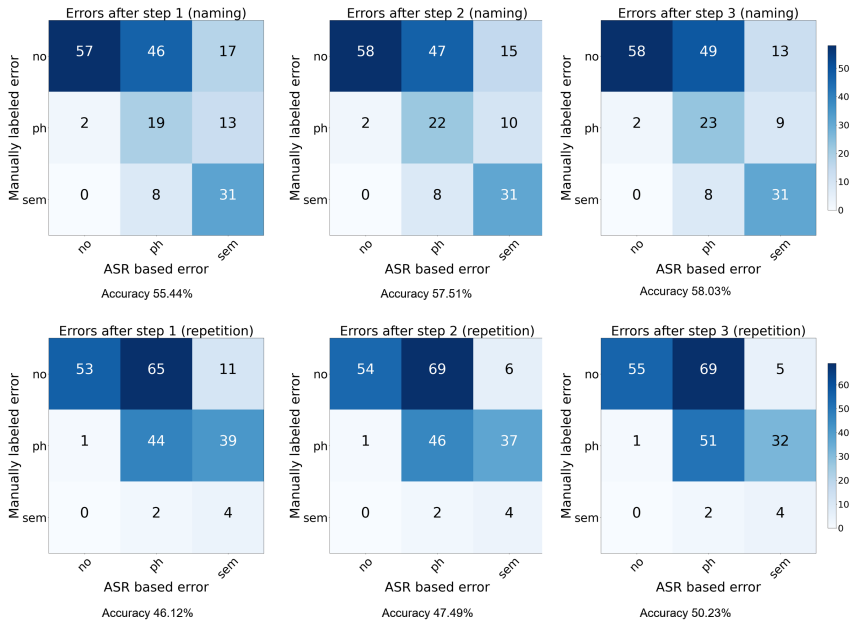


Figure 3: Confusion matrices for the pipeline’s error attribution in naming and repetition task at each step.

viewed as numerically considerable. Otherwise, the increase in precision and recall of the errors did not exceed 3%. In the rest of this subsection, only the values after step 3 are discussed.

The pipeline demonstrates similar precision in identifying error-free responses across both tasks: 96.7% in the naming task, and 98.2% in the repetition task. Only three recordings together from both tasks were misclassified as error-free, while containing phonemic/phonetic errors according to manual labelling. For example, user’s response *peife* was recognised as *pfeife* with target *Pfeife* pipe. Such misclassifications do not depend on the non-standard features and happen due to the ASR systems reliance on language and pronunciation models, which may favour plausible word forms.

Overseeing a phonemic/phonetic error is not viewed as a critical scenario because the application’s primary goal is to train vocabulary and not pronunciation. This type of pipeline’s errors could be addressed by training ASR models on pseudowords or actual mispronunciations with corresponding transcriptions.

The recall of error-free responses is, however, below 50% in both tasks: 48.3% in the naming task, and 42.6% in the repetition task. Due to the imperfections of ASR, error-free responses are frequently erroneously misclassified as phonemic/phonetic errors (and less frequently – as semantic errors), which makes the precision in identifying true phonemic/phonetic errors

relatively low: 28.8% in the naming task, and 41.8% in the repetition task. For the same reason, phonemic/phonetic errors are misclassified as semantic errors, which is most noticeably reflected in precision in identifying semantic errors in the repetition task: 9.8%. Probably, as soon as the target word was that difficult for the speaker that it caused them commit phonemic/phonetic errors, it became extremely difficult for the ASR. Nevertheless, the recall of phonemic/phonetic errors in both tasks and the precision in identifying semantic errors in the naming task are the areas that benefit the most from implementing the non-standard features (see above).

Misclassifying an error-free response as a phonemic/phonetic error is considered less disadvantageous for the user than misclassifying it as a semantic error. The feedback on phonemic/phonetic errors suggests that the response is semantically correct and encourages practising pronunciation, while the feedback on a semantic error would assume a semantically wrong response and the hints would confuse the user. Following the same logic, misclassifying phonemic/phonetic errors as semantic errors, which frequently occurred in the repetition task, is rather a harmful scenario.

The recall of semantic errors is relatively high and did not change with implementation of the non-standard features: 79.5% in the naming task, and 66.7% in the repetition task. In the erroneous scenarios, semantic errors were misclassified as phonemic/phonetic errors. It occurred when a semantically incorrect user's response was orthographically or acoustically similar to the target. The response could be a real word: *zigarette* cigarette vs. target *Zigarre* cigar, or a non-word: *schirrach* /ʃɪrɐx/ vs. target *Spruch* /ʃprʊx/.

Misclassifying a semantic error as a phonemic/phonetic error is not considered a harmful scenario because the user would still receive adequate, although not as tailored to the error type, feedback from the pipeline. Thus, after the first attempt, the pipeline would prompt the user to repeat the word slowly and clearly. After the second attempt, it would provide a phonemic/phonetic hint (see Figure 1 and Rykova & Walther, 2024a). Incorporating lexical analysis earlier in the pipeline could help reduce the acceptance of such semantically incorrect responses. Additionally, analysing syllabic structure could further aid in distinguishing between phonemic/phonetic and semantic errors, complementing the existing ER threshold.

3.4 Interspeaker variability

Given the variability in recording quality, speaker diagnoses, and individual speech characteristics, it seems reasonable to examine the results on a per-speaker basis. For this analysis, data from those seven speakers who completed both the naming and repetition tasks were used.

Table 3 presents the following information for each speaker:

- N: total number of recordings;
- N+R: number of recordings from the naming and repetition tasks;
- E-F: percentage of the recordings manually labelled as error-free;
- M dur: mean duration of recordings (s);
- M sp.r.: mean speech rate (syl/s);

- ER:M mean of the error rate (ER) on error-free items;
- ER:SD standard deviation of the ER on error-free items;
- Acc: general acceptance ($ER \leq 0.5$)/rejection ($ER > 0.5$) accuracy;
- AccN: error attribution accuracy in the naming task;
- AccR: error attribution accuracy in the repetition task.

Additionally, the SLT specialist noted that P4 exhibited laborious articulation, P6 showed signs of dysarthria, and P7 showed signs of apraxia of speech.

Table 3: Results presented per speaker

Speaker information								
	P1	P2	P3	P4	P5	P6	P7	
N items	57	72	61	39	35	67	49	
N+R	32+25	42+30	30+31	9+30	4+31	16+61	29+20	
E-F	56%	74%	89%	67%	94%	15%	39%	
M dur	1.71	0.87	1.54	2	0.82	1.15	2.5	
M sp.r.	1.85	3.15	2.12	1.54	3.45	1.86	0.85	
Automatic analysis								
step 1	ER:M	0.08	0.15	0.17	0.24	0.25	0.3	0.54
	ER:SD	0.16	0.2	0.19	0.2	0.23	0.2	0.19
	Acc	86%	87.5%	93.4%	89.7%	85.7%	58.2%	38.8%
	AccN	78.1%	57.1%	41.4%	77.8%	25%	68.8%	13.8%
	AccR	64%	53.3%	63.6%	33.3%	38.7%	41.2%	26.3%
step 2	ER:M	0.08	0.14	0.16*	0.24*	0.24*	0.29	0.53
	ER:SD	0.16	0.2	0.18	0.2	0.22	0.19	0.19
	Acc	89.5%	87.5%	98.4%	92.3%	91.4%	61.2%	42.9%
	AccN	78.1%	59.5%	41.4%	77.8%	25%	75%	20.7%
	AccR	72%	56.7%	63.6%	33.3%	38.7%	41.2%	26.3%
step 3	ER:M	0.07	0.14*	0.16*	0.23*	0.23**	0.27	0.5**
	ER:SD	0.16	0.2	0.18	0.19	0.21	0.18	0.16
	Acc	89.5%	88.9%	98.4%	92.3%	94.3%	64.2%	53.1%
	AccN	78.1%	59.5%	41.4%	77.8%	25%	75%	24.1%
	AccR	76%	60%	63.6%	33.3%	38.7%	45.1%	36.8%

* - statistically significant difference from step 1
 ** - statistically significant difference from step 1 and step 2

A closer look at the results confirms the heterogeneity of speech difficulties among PWA. For four of the seven speakers (P4-P7), the mean ER on error-free samples was above the mean of the

whole dataset at each step (0.2, 0.192, and 0.184, respectively). Nevertheless, mean ERs showed a statistically significant decrease following the implementation of non-standard features for five speakers. From the remaining two, P1 already had the lowest mean ER ($M = 0.08$ after step 1, and $M = 0.07$ after step 3), while P6 had very few error-free recordings ($n = 10$), limiting the reliability of their results.

General acceptance/rejection accuracy of the pipeline increased with the implementation of the non-standard features. After step 3, the pipeline achieved above 88% general acceptance/rejection accuracy for five speakers (PIP5), with a maximum of 98.4% for P3. For P6, the results were above chance level (cf. minimum accuracy of 65.1% in Ballard et al., 2019), and for P7, the results were at chance level. P7's speech rate is extremely low (0.85 syl/s), which is known to present difficulties for one-word ASR (Rykova & Walther, 2024b).

The three speakers with higher ERs and low error attribution accuracy in the repetition task (P4, P6, P7) were also those identified by the SLT specialist as having articulation difficulties. Motor speech disorders, comorbid to aphasia, challenge ASR systems even more. P6 and P7 supposedly suffer from motor speech disorders, comorbid to aphasia, which challenges ASR even more. comparably, the application described by Ballard et al. (2019), reaches the lowest mean acceptance/rejection accuracy than the other similar ones, including the one presented in this paper, probably because the participants of the study have apraxia of speech. However, dysarthric speech patterns are often predictable and can be modelled (Caballero Morales & Cox, 2009; Mulhari et al., 2023). Other articulatory difficulties may also be modelled upon further analysis.

Surprisingly, for P5, who produced 94% of their responses without errors, the mean ER was relatively high, and the error attribution accuracy was relatively low. This may be due to an imbalanced distribution of recordings across the tasks and very short recordings with P5's utterances. The latter aligns with findings from previous research (Rykova & Walther, 2024b).

4 Conclusions

This study presents a post hoc approach for incorporating aphasic and dialectal phonetic features into an ASR-based error analysis pipeline for people with aphasia. The proposed method is evaluated using recordings from the German Aachen Aphasia Test, specifically from naming and repetition tasks. Post hoc correction of ASR output based on knowledge of input speech characteristics shows promising results and might be applied as an alternative to spelling correction models trained on ASR errors, which do not necessarily improve ASR performance (Ulasik et al., 2021). The idea of using alternative reference targets in ASR performance evaluation (Ali et al., 2017; Nigmatulina et al., 2020) has also proven effective.

The results demonstrate that subsequent integration of non-standard phonetic features – first aphasic, then dialectal – leads to a statistically significant reduction in error rates on error-free recordings. Although the absolute improvements are modest (cf. results for speech under alcoholic intoxication), the enhancements are consistent across speakers and tasks, and the results are comparable to CER scores of an ASR system for dialectal speech with no disorders (Nigmatulina et al., 2020).

The general acceptance/rejection accuracy of the pipeline increases by 5%, reaching 83.3% (and up to 98.4% on an individual level). Such accuracy score is comparable to other SLT applications

that use ASR in automation of naming tasks and above the recommended threshold for SLT purposes (McKechnie et al., 2018).

In addition to improved acceptance/rejection accuracy, the pipeline also shows better performance in error attribution, reaching the overall accuracy of 53.9% and 58% accuracy in the naming task – both above the chance level. Since this is the first known application to offer automated error classification for PWA's speech, there are no directly comparable benchmarks.

Pipeline errors seem to stem mainly from limitations in the ASR component itself, which handles dysarthric or otherwise atypical articulatory patterns inadequately. This could be addressed with retraining, speaker-adaptation, and additional modelling of dysarthric and other articulatory difficulties.

To improve the recognition of phonemic/phonetic errors, (re)training or adapting ASR models using pseudowords or actual mispronunciations could be of help. Incorporating syllabic structure and lexical analysis earlier in the pipeline should strengthen differentiating phonemic/phonetic and semantic errors and decrease misclassification of semantically incorrect but orthographically/acoustically similar responses.

On an individual level, besides qualitative articulatory difficulties, the pipeline demonstrates sensitivity to speech rate and utterance duration. Extremely slow speech and short utterance durations can lead to poorer performance of the pipeline (in line with Rykova & Walther, 2024b).

5 Limitations and future work

The most significant difficulty faced during this study – and the broader aphaDIGITAL project – was the scarcity and poor quality of relevant data. The pipeline was tested using existing recordings of aphasic speech, which were collected under conditions different from those intended for the application. The dataset includes only ten speakers and covers a limited number of target words (cf. low numbers of participants in Abad et al., 2013; Ballard et al., 2019; Barbera et al., 2021). These constraints in both data quality and quantity mean that the current evaluation should be considered a feasibility study rather than a full validation. Nevertheless, the results demonstrate the viability of the proposed method and are comparable to those reported by other researchers working with pathological or dialectal speech.

Although the pipelines overall acceptance/rejection accuracy is similar to that of ASR-assisted SLT applications in other languages, its performance in error attribution remains insufficient for deployment in real-world therapeutic contexts. The majority of misclassifications appear to stem from limitations in the ASR component, which must be addressed upon collection of relevant data.

The current pipeline is designed for German-speaking PWA, particularly those who speak dialects from the Thuringian-Upper Saxon group. However, the pipeline's modularity allows for relatively easy substitution of dialectal features upon formulating the corresponding rules of phonemic changes, making it adaptable to other regional varieties or languages. With appropriate modifications, the pipeline could be extended to other therapeutic or language training contexts.

To address the limitations of this study, future work should focus on:

- live testing of the application with PWA in relevant scenarios;
- systematic data collection aligned with the apps intended use;

- expanding the dataset to include more speakers, a broader range of speech tasks, and greater dialectal and clinical diversity - and further analysing the impact of these factors.

References

- Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., . . . Martins, I. P. (2013). Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, 27(6), 1235–1248.
- Adikari, A., Hernandez, N., Alahakoon, D., Rose, M. L., & 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.
- Ali, A., Mubarak, H., & Vogel, S. (2014). Advances in dialectal Arabic speech recognition: a study using Twitter to improve Egyptian ASR. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Papers* (pp. 156–162).
- Ali, A., Nakov, P., Bell, P., & Renals, S. (2017). WERd: Using social text spelling variants for evaluating dialectal speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 141–148).
- Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., Alotaibi, B., & Fayed, Z. T. (2022). Deep investigation of the recent advances in dialectal Arabic speech recognition. *IEEE Access*, 10, 57063–57079.
- Azevedo, N., Kehayia, E., Jarema, G., Le Dorze, G., Beaujard, C., & Yvon, M. (2024). How artificial intelligence (AI) is used in aphasia rehabilitation: a scoping review. *Aphasiology*, 38(2), 305–336.
- 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(2S), 818–834.
- Barbera, D. S., Huckvale, M., Fleming, V., Upton, E., Coley-Fisher, H., Doogan, C., . . . 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.
- Bernard, M., & Titeux, H. (2021). Phonemizer: text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68), 3958.
- Bystrich, T. (2023). *Data-driven and rule-based approaches to improving Bavarian speech recognition* [Bachelor's Thesis]. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. (Available at <https://publica.fraunhofer.de/entities/publication/05257697-0aca-4ffc-aa1d-c953d56de790>)
- Caballero Morales, S. O., & Cox, S. J. (2009). Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing*, 2009, 1–14.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Elmahdy, M., Hasegawa-Johnson, M., & Mustafawi, E. (2014). Development of a TV broadcasts speech recognition system for qatari Arabic. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 3057–3061). European Language Resources Association (ELRA).
- Fischer, M., & Jäck, A. (2023). Einsatz KI-gestützter Diagnostik von Sprach- und Sprechstörungen bei neurodegenerativen Erkrankungen [Use of AI-supported diagnostics of speech and language disorders in neurodegenerative diseases]. *Nervenheilkunde*, 42(09), 626–634.

- Fleck, M. (2022). *XLS-R-300 Wav2Vec2 German with language model*. <https://huggingface.co/mfleck/wav2vec2-large-xls-r-300m-german-with-lm>. (Accessed on January 24, 2024)
- Graham, C., & Roll, N. (2024). Evaluating Openai's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2), 025206.
- Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., ... Tomanek, K. (2021). Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Proceedings of Interspeech 2021* (pp. 4778–4782).
- Grosman, J. (2021). *Fine-tuned XLSR-53 large model for speech recognition in German*. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german>. (Accessed on January 24, 2024)
- Guhr, O. (2022). *wav2vec2-large-xlsr-53-german-cv9*. <https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv9>. (Accessed on January 24, 2024)
- Harris, C., Mgbahurike, C., Kumar, N., & Yang, D. (2024). Modeling gender and dialect bias in automatic speech recognition. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 15166–15184). Miami, Florida, USA: Association for Computational Linguistics.
- Heide, J., Netzebandt, J., Ahrens, S., Brüsch, J., Saalfrank, T., & Schmitz-Antonischki, D. (2023). Improving lexical retrieval with LingoTalk: an app-based, self-administered treatment for clients with aphasia. *Frontiers in Communication*, 8, 1210193.
- Hinsvark, A., Delworth, N., Rio, M., McNamara, Q., Dong, J., Westerman, R., ... Jette, M. (2021). Accented speech recognition: A survey. *ArXiv, abs/2104.10747*. Retrieved from <https://api.semanticscholar.org/CorpusID:233347110>
- Hirsch, H.-G., Tiggelkamp, Y., Neumann, C., Frieg, H., & Knecht, S. (2025). Evaluating the user interface of the rehalingo speech training system with aphasic patients. In S. Gravunder (Ed.), *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2025* (pp. 61–68). TUDpress, Dresden.
- Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric statistical methods*. John Wiley & Sons.
- Huber, W. (1993). *Aachener Aphasie Test (AAT)*. Göttingen, Zürich: Verlag für Psychologie Hogrefe.
- Klumpp, P. (2024). *Phonetic transfer learning from healthy references for the analysis of pathological speech* (Doctoral dissertation, Friedrich-Alexander-Universität Erlangen-Nuernberg, Germany). Open FAU. (Available at <https://open.fau.de/items/d0c6b800-e217-4049-ab1f-a746fc9b3966>)
- Kresic, M., & Abbas, N. (2024). Normalizing swiss german dialects with the power of large language models. *Procedia Computer Science*, 244, 287–295. (6th International Conference on AI in Computational Linguistics)
- Költzsch, T. (2024, January). *Sprachcomputer der Sparkasse Nürnberg kann kein Fränkisch*. Retrieved from <https://www.golem.de/news/kommunikationspanne-sprachcomputer-der-sparkasse-nuernberg-kann-kein-fraenkisch>

-2401-181689.html

- Li, Q., Mai, Q., Wang, M., & Ma, M. (2024). Chinese dialect speech recognition: a comprehensive survey. *Artificial Intelligence Review*, 57(2), 25.
- Lin, Y., Klumpp, P., Pfab, J., Abdelioua, A., Gebray, D., & Späth, M. (2022). Eine automatische Sprachbewertung für die neolexon Aphasie-App mithilfe Künstlicher Intelligenz [Automatic language assessment with artificial intelligence for the neolexon aphasia app]. In *Sprachtherapie aktuell: Forschung-Wissen-Transfer 9 (1): XXXIV, Workshop Klinische Linguistik e2022-11*.
- Masmoudi, A., Khmekhem, M. E., Estève, Y., Belguith, L. H., & Habash, N. (2014). A corpus and phonetic dictionary for Tunisian Arabic speech recognition. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 306–310). European Language Resources Association (ELRA).
- McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for childrens speech production: a systematic literature review. *International journal of speech-language pathology*, 20(6), 583–598.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Mulfari, D., Carnevale, L., & Villari, M. (2023). Toward a lightweight ASR solution for atypical speech on the edge. *Future Generation Computer Systems*, 149, 455–463.
- Nigmatulina, I., Kew, T., & Samardžić, T. (2020). ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 15–24).
- NVIDIA. (2022). *NVIDIA Conformer-Transducer Large (de)*. https://huggingface.co/nvidia/stt_de_conformer_transducer_large. (Accessed on January 24, 2024)
- Picovoice. (2025, February). *Speech-to-text benchmark*. Retrieved from <https://github.com/Picovoice/speech-to-text-benchmark/blob/master/README.md>
- Plüss, M., Deriu, J., Schraner, Y., Paonessa, C., Hartmann, J., Schmidt, L., ... Cieliebak, M. (2023). STT4SG-350: A speech corpus for all Swiss German dialect regions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 1763–1772). Toronto, Canada: Association for Computational Linguistics.
- Plüss, M., Neukom, L., & Vogel, M. (2021). Swisstext 2021 task 3: Swiss german speech to standard german text. In *Proceedings of the swiss text analytics conference* (Vol. 2021).
- Pompili, A., Abad, A., Trancoso, I., Fonseca, J., Martins, I. P., Leal, G., & Farrajota, L. (2011). An on-line system for remote treatment of aphasia. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies* (pp. 1–10).
- Pottinger, G., & Kearns, Á. (2024). Big data and artificial intelligence in post-stroke aphasia: a mapping review. *Advances in Communication and Swallowing*, 27(1), 41–55.
- Protalinski, E. (2017, May). *Googles speech recognition technology now has a 4.9% word error rate*. Retrieved from <https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word>

-error-rate/

- R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Redrovan-Reyes, E., Chalco-Bermeo, J., Robles-Bykbaev, V., Carrera-Hidalgo, P., Contreras-Alvarado, C., Leon-Pesantez, A., ... Olivo-Deleg, J. (2019). An educational platform based on expert systems, speech recognition, and ludic activities to support the lexical and semantic development in children from 2 to 3 years. In *2019 IEEE Colombian Conference on Communications and Computing (COLCOM)* (pp. 1–6).
- Rocholl, M. J. (2015). *Ostmitteldeutsch–eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum [East-Central German a modern regional language? An investigation into constancy and change in the Thuringian-Upper Saxon language area]*. Georg Olms Verlag.
- Rykova, E., & Walther, M. (2023). Concept for semantic error analysis in a mobile application for speech and language therapy support. In C. Draxler (Ed.), *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023* (pp. 127–133). TUDpress, Dresden.
- Rykova, E., & Walther, M. (2024a). 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 (HICSS)* (pp. 3385–3394). University of Hawai'i at Mnoa.
- Rykova, E., & Walther, M. (2024b). Linguistic and extralinguistic factors in automatic speech recognition of German atypical speech. In P. H. Luz de Araujo, A. Baumann, D. Gromann, B. Krenn, B. Roth, & M. Wiegand (Eds.), *Proceedings of the 20th conference on natural language processing (KONVENS 2024)* (pp. 358–367). Association for Computational Linguistics.
- Rykova, E., & Walther, M. (2025). Evaluation of German ASR solutions in the context of speech and language therapy support of people with aphasia. *Loquens*, 12, e116.
- Samardžić, T., Scherrer, Y., & Glaser, E. (2016). ArchiMob - a corpus of spoken Swiss German. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 4061–4066). European Language Resources Association (ELRA).
- Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W. R. (2009). Tools and technologies for computer-aided speech and language therapy. *Speech communication*, 51(10), 948–967.
- Scherrer, Y., & Rambow, O. (2010). Natural language processing for the Swiss German dialect area. In *Semantic approaches in natural language processing – proceedings conference on natural language processing 2010 (konvens)*.
- Siebenhaar, B. (January, 2024). Personal communication.
- Suwanbandit, A., Naowarat, B., Sangpetch, O., & Chuangsuwanich, E. (2023). Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition. In *Proceedings of Interspeech 2023* (pp. 4069–4073).
- Ulasik, M. A., Hürlimann, M., Dubel, B., Kaufmann, Y., Rudolf, S., Deriu, J., ... Cieliebak, M. (2021). ZHAW-CAI: ensemble method for Swiss German speech to standard german text.

In *Swiss text analytics conference–swisstext 2021, online, 14-16 june 2021*.

- Wallraff, U. (2007). *Ausgewählte phonetische Analysen zur Umgangssprache der Stadt Halle an der Saale [Selected phonetic analyses of the colloquial language of the city of Halle an der Saale]* (Doctoral dissertation, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany). Universitäts- und Landesbibliothek Sachsen-Anhalt. (Available at <https://opendata.uni-halle.de/handle/1981185920/9620>)
- Wirth, J., & Peinl, R. (2022). Automatic speech recognition in German: a detailed error analysis. In *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)* (pp. 1–8).
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., ... Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), 2410–2423. doi: 10.1109/TASLP.2017.2756440

Appendix

Rules of phonemic changes based on the characteristics of the Thuringian-Upper Saxon dialect group

The presentation of the context roughly follows the notations used for phonological rules (Chomsky & Halle, 1968), where C stands for consonant, V stands for vowel, # stands for 'no sound' and is also used to indicate the boundaries of words. The transcriptions were generated automatically using the grapheme-to-phoneme converter *bootphon espeak phonemizer* (Bernard & Titeux, 2021) with some modifications (see Section 2.2).

Schwa (ə) deletion and following assimilation of place (if applicable) could be a feature of informal speech or a dialectal pronunciation. The discussion of the reasons for this phenomenon is outside the scope of the current work, and the corresponding cases are presented as dialectal features following Wallraff (2007).

phonemic change	context (if applicable)	example: standard transcription → alternative transcription
a → ɔ	___ r	Markt /markt/ → /mɔrkt/
a: → o:	___ r	Nahrung /na:rʊŋ/ → /no:rʊŋ/
ɪ → ʏ	___ r, ___ m, ___ ʃ, r ___	irgendwie /ɪrgəndvi:/ → /ʏrgəndvi:/
i: → ɪ	___ C(ə)l, ___ C(ə)r, ___ C(ə)n	Zwiebel /tʃvi:bəl/ → /tʃvɪbəl/
ʏ → ɪ		Brücke /brʏkə/ → /brɪkə/
y: → i:		Blüte /bly:tə/ → /bli:tə/
ʊ → ɔ	___ r	Turm /tʊrm/ → /tɔrm/
œ → ɛ		Knöchel /knœçəl/ → /knɛçəl/
ø: → e:		Vögel /fø:gəl/ → /fe:gəl/
e: → ɛ:		Meer /me:r/ → /mɛ:r/
aʊ → o:		auch /aʊx/ → /o:x/
aɪ → e:		Fleisch /flaɪʃ/ → /fle:ʃ/
ɔɪ → e:		Leute /lɔɪtə/ → /le:tə/
ə → #	___ n, ___ l, ___ m	besen /be:zən/ → /be:zn/
p(ə)n → bm	___ #	umstülpen /ʊmfʏtlpən/ → /ʊmfʏtlbm/
b(ə)n → bm	___ #	leben /le:bən/ → /le:bm/
k(ə)n → gŋ	___ #	schlucken /ʃlʊkən/ → /ʃlʊgŋ/
nən → n	___ #	brennen /brɛnən/ → /brɛn/
mən → m	___ #	kommen /kɔmən/ → /kɔm/
ŋən → ŋ	___ #	fangen /faŋən/ → /faŋ/
pf(ə)n → pfm	___ #	hüpfen /hʏpfən/ → /hʏpfm/
p → b		Papier /papi:r/ → /babi:r/

phonemic change	context (if applicable)	example: standard transcription → alternative transcription
t → d		Teufel /tʊɔfəl/ → /dʊɔfəl/
k → g		Kalt /kalt/ → /galt/
b → p	___ r, ___ l	Blüt /bly:t/ → /ply:t/
d → t	___ r, ___ l	Droge /dro:ɡə/ → /tro:ɡə/
g → k	___ r, ___ l	Glocke /ɡlɔkə/ → /klɔkə/
nd → n		finden /findən/ → /finən/
k → ç	V [+front] ____, l ____, r__	Krieg /kri:k/ → /kri:ç/
k → x	V [-front] __ C [+alveolar]	Jagd /jakt/ → /ja:xt/
g → j	# __ V, V __ V	genau /ɡənaʊ/ → /jənaʊ/
g → ç	V [+front] __ V	Ziege /tsi:ɡə/ → /tsi:jə/
g → ʃ	V [+front] __ V	Ziege /tsi:ɡə/ → /tsi:ʃə/
g → x	V [-front] __ V	Beluga /belu:ɡa/ → /belu:xa/
ç → ʃ		Schicht /ʃiçt/ → /ʃiʃt/
pf → f	# ____, m__ #	Pfeife /pfaiə/ → /faiə/ Kampf /kampf/ → /kamf/
pf → b	__ V	Apfel /apfəl/ → /abəl/
pf → p	__ #	Zopf /tsɔpf/ → /tsɔp/
ts → s	__ #	Holz /hɔltʰs/ → /hɔls/


Correspondence

Eugenia Rykova 

Catholic University of Eichstätt-Ingolstadt &
University of Eastern Finland
eugenia.rykova@ku.de

Elisabeth Zeuner

Martin Luther University Halle-Wittenberg

Susanne Voigt-Zimmermann 

Martin Luther University Halle-Wittenberg

Mathias Walther 

Technical University of Applied Sciences TH Wildau