Digitalization of speech therapy with AI-based personalized feedback

Eugenia Rykova

University of Eastern Finland
University of Applied Sciences TH Wildau
Catholic University Eichstätt-Ingolstadt
eugenryk@uef.fi

Abstract

This paper introduces an ongoing PhD thesis carried out in the framework of a research project, in which an application for speech and language therapy support of German speakers with aphasia is developed. Work completed in the selection and implementation of ASR solutions, and creating a semantic analysis pipeline is described, and challenges and future work perspectives are discussed.

Index Terms: automatic speech recognition, aphasic speech, speech and language therapy, digital health.

1. Introduction

Aphasia, literally translated from (Ancient) Greek as „speechlessness“ [1], is an acquired language disorder due to a focal brain injury. It affects some or all language modalities, which makes communication difficult and decreases the quality of life. High intensity and duration of speech and language therapy (SLT) bring certain benefits to communication improvements [2-3]. However, not all people with aphasia (PWA) have access to sufficient SLT (e.g., due to lack of specialists or geographical remoteness). Research shows the efficiency of supplementing in-person therapy with independent usage of digital therapy solutions [4].

Various researchers have explored the possibilities of automatic assessment of speech produced by PWA (see, for example, [5-6]). Naming-oriented semantic exercises have been automated with the help of automatic speech recognition (ASR) for Portuguese [7], English [8-9], and German [10-11]. The latter, however, are not in active use yet. When the answer is only rated as correct/incorrect, with no further analysis of users' errors, the reported acceptance/rejection accuracies on PWA's speech range from 75% [8] to 89.5% [9].

The current project [12] focuses on developing a mobile application for German-speaking PWA that will provide personalized detailed feedback in naming and other exercises. In the present PhD research, ASR and further text processing solutions are used for multilevel feedback: phonemic/phonetic, semantic, and grammatical. To build the corresponding pipeline, the following questions need to be addressed.

1. Which existing (open-source) ASR solutions are suitable for the task-specific speech of German-speaking PWA?

2. How can selected ASR solutions be improved and/or adapted for the purposes of SLT?

3. How can a combination of selected ASR solutions and existing tools for semantic and grammatical analysis serve for speech production errors analysis?

4. What are patients' and therapists' attitudes to the proposed digital solution?

2. ASR solutions

2.1. Model selection

Evaluation of the ASR systems consisted of several steps. First, the suitability of more than 50 open-source ASR solutions was assessed with the help of several speech recordings from different corpora, including PWA's speech [13-14]. Based on the ranking of error rates, 13 models were selected for further evaluation. In the absence of necessary data from PWA, test material from other corpora with atypical speech (presenting abnormalities similar to PWA's speech) was considered for further evaluation, namely speech of adult cochlear implant (CI) users and normal-hearing speakers as a counterpart [15], and speech produced under alcohol intoxication the same speakers under no intoxication as a counterpart [16]. Additionally, two small datasets with aphasic speech were used. AvEv recordings (39 single words) were obtained from four PWA who had taken part in an avatar evaluation experiment was used [17]. UniSt recordings (79 single words) had been made during Aachen Aphasia Test (AAT) [18] sessions and were obtained on request from Stuttgart University Institute for Natural Language Processing [19].

Finally, four open-source ASR models were selected for the backend of the app [20-23] based on character error rates, the number of empty outputs, and the number of precisely recognized words. Three of these models [20-22] are to a certain extent independent from pronunciation and language models and are suitable for phoneme-level pronunciation analysis, while the fourth model [23] gives only existing orthographic forms as output, which is more suitable for subsequent semantic and grammatical error analysis. All four models are to a greater or lesser extent robust to speaker gender and age. The experiments suggest that for better single-word recognition the audio samples should be not too short and pronounced neither too slowly nor too fast (i.e. intentionally speeded up).

2.2. Post-hoc implementation of non-standard phonetic features

Although the selected ASR models present a possibility for fine-tuning, the project lacks adequate data for model (re)training. Thus, it was decided to research the possibility of applying the knowledge about non-standard phonetic features post hoc to ASR output. The methodology combines generating alternative pronunciations based on non-standard patterns [24] and using alternatives for evaluation [25].

First, the orthographic ASR output form is phonemized using automatic grapheme-to-phoneme conversion (g2p) [26]. Then, the phonetic transcription is subject to modifications based on the non-standard phonetic features. The first set of features considers aphasic speech: syllabification with greater pauses between syllables, which causes recognition of syllables as separate words; and slow and careful speech production, which causes vowel prolongation. The next set comprises relevant dialect features selected from the Thuringia-Upper Saxon dialect group [27-29] due to the geography of the project and the data available for the experiments. The modified transcription is then compared to the target transcription, and the error rate (ER) threshold is applied. If the ER is lower than the threshold, the error is considered phonemic/phonetic, and semantic otherwise (see example in Figure 1).

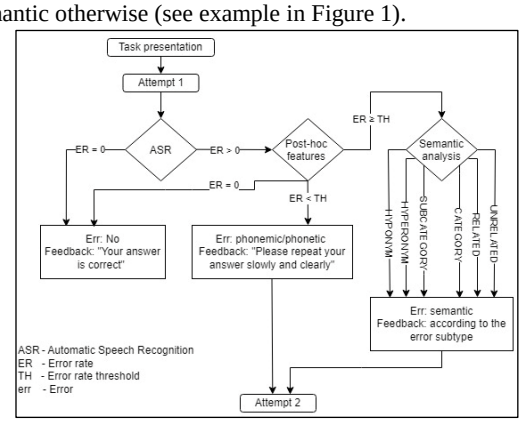

Figure 1: *Error analysis pipeline – attempt 1.*

The proposed method was tested on the 412 single-word recordings made during AAT sessions and obtained on request from the University of Leipzig Medical Center. It has proven to work: after the implementation of each feature set, ASR error rates decrease significantly, general acceptance/rejection accuracy improves, and the accuracy of error attribution also increases.

2.3. Challenges and future work

The selected four models present a possibility to be fine-tuned: to PWA speech or speech of a particular user in a customized version, and to single-word recognition task rather than continuous speech recognition. This requires a certain amount of corresponding data. Obtaining the data from PWA, possible anonymization of such data, and further application for model fine-tuning are seen as the following steps. The complete solution should be tested "live" to take into account the corresponding audio quality and processing time.

3. Semantic and grammatical analysis

3.1. Semantic analysis pipeline

If the answer of the speaker does not pass the ER threshold (i.e. is not recognized as correct or containing phonemic/phonetic errors only), it is subject to further analysis. In particular, it must be compared to the target in terms of their semantic relationship and distance. The current semantic analysis is built upon GermaNet – a semantic network for the German language [30]. It consists of two parts: semi-automatic enrollment of the exercise item into the system and the analysis of a semantic error. The latter includes but should not be limited to the recognition of hyponymy/hypernymy and belonging to the same semantic (sub)category. If the answer is not recognized as an existing word (i.e. contains both semantic and phonemic/phonetic errors), a search for close orthographic matches is performed, and the match that is semantically the closest to the target is subject to the relationship analysis described above.

Current work is concentrated on including further types of semantic relationships in analysis, for example, synonymy, antonymy, and meronymy/holonymy. Close orthographic matches search is extended with close phonemic matches search, using automatic g2p. On the other hand, the search is thought to be limited to the members of the target lexical and semantic categories, while the final assumption is based on both semantic and orthographic/phonemic distance to the target.

3.2. Challenges and future work

The current pipeline, or GermaNet in general, has certain limitations. First, it is mostly suitable for the words of the same lexical category (except for causative relationship, pertainyms, and participles), so that the relationship between "to eat" and "food" would not be recognized. Second, GermaNet takes only lemmas as input, which makes it necessary to implement an additional step with a lemmatizer. Further limitations can arise from a mismatch of the semantic categories in typical SLT tasks or a broader common understanding of language and GermaNet. Exploring other semantic networks (e.g., BabelNet [31]), adding grammatical analysis, elaborating more intuitive semantic categories, and joint implementation with selected ASR solutions will be addressed next.

4. Users' evaluation

The automatic error analysis process includes the following components: ASR, post-hoc phonetic features implementation (if applicable), phonemic/phonetic error analysis (if applicable), semantic and grammatical error analysis (if applicable), and issuing corresponding feedback.

Based on this general pipeline, digital exercises are to be designed and implemented in an app and then tested in SLT practice. A questionnaire is then designed to collect therapists' and patients' opinions on the exercises. Such parameters as, for example, plausibility, clarity, and user-friendliness should be assessed. The answers will be evaluated both quantitatively and qualitatively, and compared between the subsets. Based on the feedback, necessary changes or suggestions can be made to improve the solution (cf. [32]).

5. Limitations

The greatest limitation of the current work is the lack of relevant data. ASR solutions were mostly tested with other atypical speech and to much less extension with aphasic speech. Furthermore, the analyzed data mentioned in this paper are not suitable for ASR model (re)training or adaptation.

On the other hand, few semantic errors are present in the data, and the examples to test the semantic analysis pipeline have to be constructed artificially. The current basis for semantic analysis, GermaNet [30], presents certain limitations on its own, described above in the corresponding section.

The current project is a regional one and therefore is focused on the German language, in particular on the dialects of the Thuringian-Upper Saxon group. However, general principles and pipelines elaborated as the result of the present research can be scaled to other dialects and languages.

6. References

[1] J. Ryalls, "Where does the term "aphasia" come from?" *Brain and Language*, 21, pp. 358-363, 1984.

[2] S. K. Bhogal, R. Teasell, and M. Speechley, "Intensity of aphasia therapy, impact on recovery," *Stroke*, 34, pp. 987-993, 2003.

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

[4] M. Braley, J. S. Pierce, S. Saxena, E. D. Oliveira, L. Taraboanta, V. Anantha, . . . S. Kiran, "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, 2021.

[5] A. Adikari, N. Hernandez, D. Alahakoon, M. L. Rose, and J. E. Pierce, "From concept to practice: a scoping review of the application of AI to aphasia diagnosis and management," *Disability and Rehabilitation*, pp. 1-10, 2023.

[6] G. Pottinger and Á. Kearns, "Big data and artificial intelligence in post-stroke aphasia: A mapping review," *Advances in Communication and Swallowing*, vol. Pre-press, no. Pre-press, pp. 1-15, 2024.

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

[8] K.J. Ballard, N. M. Etter, S. Shen, P. Monroe, and C. Tien Tan C., "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, pp. 818–834, 2019.

[9] D. S. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, C. Doogan, I. Shaw, W. Latham, A. P. Leff, and J. Crinion, "NUVA: A Naming Utterance Verifier for Aphasia Treatment," *Computer Speech & Language*, 69, 101221, 2021.

[10] J. Heide, J. Netzebandt, S. Ahrens, J. Brüsch, T. Saalfrank, and D. Schmitz-Antonischki, "Improving lexical retrieval with LingoTalk: an app-based, self-administered treatment for clients with aphasia," *Frontiers in Communication*, 8:1210193, 2023.

[11] Y. Lin, P. Klumpp, J. Pfab, A. Abdelioua, D. Gebray, and M. Späth, "Eine automatische Sprachbewertung für die neolexon Aphasie-App mithilfe Künstlicher Intelligenz [Automatic language assessment with artificial intelligence. for the neolexon aphasia app]," *Poster session presentation at Sprachtherapie aktuell: Forschung - Wissen – Transfer 9(1): XXXIV. Workshop Klinische Linguistik e2022-11*, April 2022.

[12] TDG - TRANSLATIONSREGION FÜR DIGITALE GESUNDHEITSVERSORGUNG [Translational Region for Digital Healthcare], "*AphaDIGITAL: Entwicklung einer digitalen, dezentralen sprachtherapeutischen Versorgung [Development of digital, decentralized speech therapy solutions"*. Accessed: January 25, 2024. Available: https://inno-tdg.de/projekte/aphadigital/

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

[14] Rhein-Zeitung, Germany. *Am Anfang war das Wort: Zu Besuch bei einem Aphasiker [In the beginning was the word: Visiting a person with aphasia]*. (Oct. 26, 2017). Accessed: May 16, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=Z1ZglYMSx1Y.

[15] V. Neumeyer, "Phonetische Untersuchungender Artikulation von CI-Trägern [Phonetic studies of CI users' articulation] [Master's thesis]," Master Thesis, Ludwig-Maximilians-Universität, München, Germany, 2011.

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

[17] E. Zeuner, J. Pietschmann, S. Voigt-Zimmermann, E. Rykova, and M. Walther, "aphaDIGITAL - Avatar-gestützte digitale Aphasietherapie: Evaluation [aphaDIGITAL: Avatar-supported digital aphasia therapy - evaluation study]," presented as poster 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, September 23-25, 2022. Available at https://www.researchgate.net/publication/371510421_aphaDIGITAL_-Avatar-gestutzte_digitale_Aphasietherapie_Evaluation.

[18] W. Huber, *Aachener aphasie test (AAT) [Aachen Aphasia Test]*. Verlag für Psychologie Hogrefe, Göttingen, Zürich, 1993.

[19] Universität Stuttgart. „Sprache und Gehirn: Ein neurolinguistisches Tutorial [Language and brain: a neurolinguistics tutorial]". Accessed: June 17, 2023. [Online]. Available: https://www2.ims.uni-stuttgart.de/sgtutorial/index.html.

[20] M. Fleck, "*Wav2vec2-large-xls-r-300m-german-with-lm"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/mfleck/wav2vec2-large-xls-r-300m-german-with-lm.

[21] J. Grosman, "*Fine-tuned XLSR-53 large model for speech recognition in German"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german.

[22] O. Guhr, "*wav2vec2-large-xlsr-53-german-cv9"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv9.

[23] NVIDIA, "*NVIDIA Conformer-Transducer Large (de)"*. Accessed: September 12, 2022. [Online]. Available: https://huggingface.co/nvidia/stt_de_conformer_transducer_large.

[24] A. Masmoudi, M. E. Khmekhem, Y. Est`eve, L. H. Belguith, and N. Habash, "A corpus and phonetic dictionary for Tunisian Arabic speech recognition," in *LREC*, 2014, pp. 306-310.

[25] A. Ali, P. Nakov, P. Bell, and S. Renals, "WERd: Using social text spelling variants for evaluating dialectal speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2017, pp. 141–148.

[26] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in Python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.

[27] U. Wallraff, "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, 2007.

[28] M. J. Rocholl, *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]*. Hildesheim, Zürich, New York: OLMS, 2015.

[29] B. Siebenhaar, private communication, Jan. 2024.

[30] B. Hamp, and H. Feldweg, "GermaNet - a Lexical-Semantic Net for German," *Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

[31] R. Navigli, and S. P. Ponzetto, "BabelNet: Building a Very Large Multilingual Semantic Network," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 216–225.

[32] A. Pompili, A. Abad, I. Trancoso, J. Fonseca, and I. P. Martins, "Evaluation and Extensions and of an Automatic Speech Therapy Platform," in *Computational Processing of the Portuguese Language. PROPOR 2020. Lecture Notes in Computer Science*, P. Quaresma, R. Vieira, S. Aluísio, H. Moniz, F. Batista, and T.-Gonçalves, Eds., 12037, Springer Cham, 2020, pp. 43–52.