Cross-lingual transfer learning does not improve aphasic speech recognition

Conference Paper · May 2025

DOI: 10.35096/othr/pub-8051

4 authors, including:

Norina Lauer
Regensburg University of Applied Sciences
100 PUBLICATIONS **175** CITATIONS

SEE PROFILE

CROSS-LINGUAL TRANSFER LEARNING DOES NOT IMPROVE APHASIC SPEECH RECOGNITION

Sara Mühlhausen, Sarah Gomez, Norina Lauer, Timo Baumann
OTH Regensburg
sara.muehlhausen@st.oth-regensburg.de, timo.baumann@oth-regensburg.de

Abstract: In addressing the particular linguistic challenges posed by patients suffering from aphasia, a language disorder, this paper proposes a fine-tuning approach to enhance the speech recognition capabilities of existing models. The available aphasic research data in German is highly limited. To address this constraint, we propose a cross-lingual transfer approach to utilize English data to improve performance in German. This advancement aims to support the development of a therapy platform tailored for patients with aphasia. For the base speech recognition model, we choose to use OpenAI's Whisper model, and for fine-tuning, we make use of TalkBank's AphasiaBank. The experimental findings demonstrate that the transcription of aphasic audio with Whisper is less successful than non-aphasic audio. However, fine-tuning the transcription in the respective language resulted in an enhancement of its quality. In contrast, fine-tuning the transcription in another language and expecting a transfer of the learned aphasic speech properties led to a deterioration in its quality.

1 Introduction

The prevalence of aphasia, a language disorder predominantly caused by brain damage from strokes [1], is estimated to be 100,000 patients in Germany. A third of the patients having a stroke are diagnosed with aphasia [2]. With the number of patients increasing with demographic change, the need for speech and language therapy often exceeds the resources [3], especially in rural areas. In addition to regular in-person meetings with speech and language therapists, in which spoken interactions are trained (such as buying bread in the bakery or medicine in the local pharmacy), patients may benefit from automated platforms for such interaction trainings that lead to a higher training load (and, similarly, higher training outcomes). Existing platforms [4, 5, 6] do not offer interactive communicative exercises or only linear dialogues. We propose a setting to train dialog interaction as shown in Figure 1 and have tested it in a Wizard-of-Oz (WOZ) setting, which we intend to fully automate in the future.

Full automation of the system, of course, raises the question of whether state-of-the-art speech recognition can handle our patients' speech, specifically the special forms of aphasia symptoms, such as phonemic and semantic paraphrasis, neologisms and meaningless sentences. Additionally, we investigate if and how modern speech recognition can be fine-tuned towards the specifics of aphasic speech. Corpora of aphasic speech data are scarce, especially for German, most likely due to data privacy concerns with this vulnerable group of people. In practice, there are very little available data to fine-tune on. We assess whether cross-lingual transfer learning can help overcome this problem by employing English aphasic speech data for our purpose. We find that cross-lingual transfer learning based on English aphasic speech for German aphasic speech is inferior to the baseline model. However, using just the little aphasic speech data that are available for German already helps to improve recognition even in the cross-speaker and cross-domain condition.

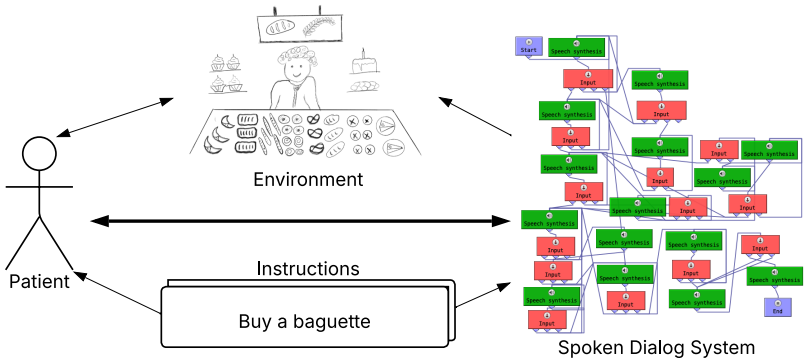

Figure 1 – Exemplary structure of the planned training platform.

2 Data

This section outlines the data used in the experiments reported below, including a description of the overall setup of our ultimate dialogue training platform, the steps taken to acquire the preliminary data, and the methods for transcription and annotation. We also describe the data that we use from other sources.

2.1 Data Collected for This Study

The setting for our ultimate system is depicted in Figure 1 in which a patient interacts with a dialog system and is also presented with a visualization of a task environment (in this case: a bakery). The patient also receives instructions about the task that is supposed to be trained (in this case: buying a baguette). During interaction, the dialog system may alter the display of the environment (e. g., highlight some product). The dialog system will also have access to the task. While this could greatly simplify the interaction, the educational nature of the system implies that it will not fully simplify the interaction but use the task information to keep the complexity of the dialogue interaction at a level that it deems optimal wrt. the patient's training objective.

2.1.1 Therapist and Wizard-of-Oz Dialogues

We performed a data collection with three aphasia patients over multiple experimental rounds a few weeks apart. The patients were tasked with performing two different everyday dialogue scenarios. In the first experimental round, participants worked through their tasks with a speech and language therapist (SLT) (the second author), and another researcher working the audio and video equipment. In the second round, we used the Wizard-of-Oz paradigm [7] to explore the functionality of our intended system with target group users (specifically, the dialogue interaction) before the full system is ready (in particular: high-quality speech recognition for aphasic speech). The SLT in this case guided the users in cases of trouble and the technical researcher (the first author) worked the WOZ system. The prototype of the speech therapy platform was developed using the DialogOS software [8].

Unfortunately, our IRB terms do not allow us to publicly share the speech data from our pilot study. However, we will amend those terms to be able to publish data for a main study in which we hope to investigate the automated system with a broader user population.

Table 1 – Descriptive statistics of the three corpora used.

	our	**Stark**	**Eng-all**
language	DE	DE	EN
total length	00:20:29	03:15:25	441:15:34
length sentence level	00:14:42	01:04:00	208:56:37
# IPUs	215	1,019	210,928
# speakers with aphasia	3	1	504
# transcribed words	887	9,338	1,527,636

2.1.2 Transcription and Annotation

We recorded video and audio of the interactions, as well as relevant log events in the dialog system toolkit. We manually performed speaker diarization, fully transcribed the speech, and annotated the patients' speech for aphasia phenomena using Praat [9], an open-source software that is widely used in the field of speech analysis and phonetics. The annotations were exported to TextGrid. Below, we focus on the patients' speech and transcripts only, which we extract from the TextGrid files and the corresponding audio based on timestamps of the individual speakers (therapist/system and patient).

2.2 AphasiaBank

AphasiaBank [10], a subsidiary of TalkBank [11], is an initiative to make aphasic speech data available to researchers. The majority of the data that it contains is in English. We scraped all English data as well as the only German data, the *Stark corpus*.

Only parts of the data are transcribed (including timestamps) and therefore there is a large difference between the raw audio duration and speech duration after breaking the data into sentence-like inter-pausal units (IPUs) from the time-stamped transcripts. The inter-pausal units were detected using the accompanying transcripts and the metadata provided by the used CHILDES format.

Some descriptive corpus statistics are presented in Table 1. As can be seen, our corpus is by far the smallest of the three but, in contrast to the Stark corpus, contains speech from multiple speakers. Utterances in our own corpus contain fewer words (about 4 words per utterance, as compared to 7-8 for the two other corpora). This may stem from the more rigid interaction paradigm of our setting as compared to the settings of the other corpora.

We randomly split off 1000 IPUs from the English corpus as a conveniently sized test set (*Eng-1000* below); additionally, we employ train/validation splits of 80 %/20 % during fine-tuning (for English: excluding the eng-1000 data).

3 Speech Recognition Experiments

We base our experiments on the Whisper [12] speech recognizer which uses an encoder-decoder transformer [13] architecture to directly convert log Mel-scaled 80-channel spectrogram representations of the audio and decodes those to a sequence of BPE-based text tokens. Whisper models are available as open source and have been trained in multiple model sizes, both uni- and multi-lingually (including German); the multi-lingual models are trained on roughly 680k hours of speech (about ⅔ English). Below, we use the 'large' version in its pre-trained form, and the 'small' version for fine-tuning.

With the end-to-end trained encoder-decoder network architecture, it is not trivial to say

which part of the network learns what part of the task. However, in auto-regressive decoding, the decoder likely learns some form of conditioned language model of the output language and the encoder learns the conditioning of said model on the speech signal.

We use Whisper as provided via the HuggingFace library[1], which is based on PyTorch [14].

3.1 Fine-tuning

Many commonly used models are either pre-trained exclusively for English or, when multilingual, tend to underperform compared to their monolingual counterparts [15]. To address this gap, the concept of transfer learning across languages has gained relevance, offering a promising approach to improve multilingual model performance.

Whisper models, although they can be used as-is, can easily be fine-tuned to better suit a specific task or language and the original paper reported that data on the order of tens of minutes may already be sufficient [12].

We are not aware that Whisper has previously been fine-tuned to optimize recognition performance on continuous aphasic speech. Therefore, we are lacking recipes in how to do this. We compare our task with fine-tuning for dialects (given that aphasic speech sounds different), and for new languages (given that words may deviate from the norm and sentence structure is often interrupted).

Pekarek Rosin and Wermter find that a limited fine-tuning (specifically: limiting fine-tuning of the encoder or decoder) yields higher performance when optimizing Whisper for German ASR when only limited data is available [16]. Liu et al. provide a thorough comparison of fine-tuning strategies to optimize Whisper for low-resource ASR using languages from FLEURS [18] (with about 20 hours of data each) [17]. Here, they find that limiting the fine-tuning to parts of the encoder or decoder typically negatively impacts performance and that regular fine-tuning works best in their case. Torgbi et al. fine-tune Whisper to enhance performance for dialectal speech, again using tens of hours and standard fine-tuning [19].

We take inspiration from Abad et al. who worked on zero-resource domain adaptation [20] in which fine-tuning material in one language (and domain) is used to boost performance in that domain in another language. We believe that there are aspects of aphasic speech that may well hold cross-linguistically (such as hesitation phenomena, syllabic mistakes, . . .) and we test the approach of zero-resource adaptation for German aphasia speech using English aphasia speech data.

We therefore try to improve performance on our aphasic speech corpus in multiple ways:

- we fine-tune on the Stark dataset, although it is only about one hour and single-speaker;
- we fine-tune on the English Aphasia data from AphasiaBank in the hope that this picks up on aphasia-related speech phenomena (hopefully without too much catastrophic forgetting of the model's German recognition ability);
- we fine-tune on the English Aphasia data as above but only the encoder or the decoder, respectively, under the assumption that it could be aspects of the decoder or encoder that should primarily be adapted for aphasic speech;
- we combine the English and German data for fine-tuning.

Fine-tuning is performed on whisper-small, which has 244 million parameters and therefore makes computations more tractable than whisper-large-v3 with its 1.5 billion parameters.

[1] `https://huggingface.co/openai/whisper-large-v3`

Table 2 – Recognition results for our German aphasia corpus and the acquired aphasia corpora on pre-trained models. All numbers are percentages after text normalization.

Model	Dataset	n. WER	n. CER
whisper-large-v3	DE: our	22.8	14.2
	aphasia patient	41.2	25.0
	therapist	10.8	6.6
	DE: Stark	36.2	25.9
whisper-small	DE: our	29.7	17.3
	aphasia patient	53.8	32.7
	therapist	16.3	10.0
	Eng-1000	42.7	28.4
	DE: Stark	60.1	42.8

3.2 Evaluation

The evaluation measurements used in our experiments are Word Error Rate (WER) and Character Error Rate (CER), the most commonly used performance indicators for ASR. The transcripts are evaluated with these metrics after text normalization for which we used the BasicTextNormalizer[2], an implementation of Whisper's text normalization algorithm. This ensures that normalization between corpus transcripts and Whisper expectations match.

4 Results

We present our findings in this section, detailing the speech recognition experiments conducted, the fine-tuning procedures applied, and the evaluations performed on the resulting models.

4.1 Off-the-shelf Pre-trained Speech Recognition

The results of our speech recognition experiments are indicated in Table 2.

Performance on aphasic speech yields word error rates between 22–30 % (our) and 35–60 % (Stark). This is noticeably worse than on common datasets such as *Common Voice 15* (WER: 5.7 %; numbers after text normalization for whisper-large-v3) or FLEURS (WER: 4.9 %) [12]. Performance on German data is somewhat worse compared to the English corpus (Eng-1000), which has a normalized WER of 42.7 % with the whisper-small model. We note that our results for the Stark corpus are in line with other results reported for German aphasia speech using Wav2Vec 2.0-based speech recognition [21]. Finally, we note that our corpus appears to yield better numbers than those found on AphasiaBank.

4.2 Fine-tuned ASR

The results of fine-tuning are presented in Table 3. First off, we note that fine-tuning works as intended when applied in-domain and in-language: we find a relative error reduction of about 16 % for English and about 45 % for German aphasia data. Interestingly, in-domain WERs are even lower when both English and German aphasia data are combined for fine-tuning. This may imply that with both languages together, the fine-tuning is better able to single in on aphasia-related phenomena in the speech, which could be taken as an argument in favour of aphasia-related phenomena sounding similarly (for ASR) across languages and a potential for cross-lingual transfer learning.

[2] `https://github.com/kurianbenoy/whisper_normalizer`

Table 3 – Evaluation results for the aphasia-only part of our collected dataset with different fine-tunings of the whisper-small model onto aphasia speech data. All numbers are percentages after text normalization; val. WER refers to the fine-tuned performance for the validation set part of each fine-tuning corpus.

Fine-tuning Corpus	val. WER	n. WER	n. CER
no fine-tuning	–	53.8	32.7
DE: Stark	35.8	51.2	27.6
Eng-all	32.8	67.5	39.0
decoder-only	36.1	77.9	48.2
encoder-only	37.2	1146.2	613.3
Eng-all + Stark	26.9	68.4	35.5

With respect to optimizing performance on our corpus, we find that fine-tuning with in-language data (Stark) yields slight improvements in WER and CER although the fine-tuned small model does not outperform the large off-the-shelf model.

However, our hypothesis that fine-tuning could be used for cross-lingual transfer learning does not seem to hold, as can be seen by the strong deterioration of performance on German data when fine-tuning on English data. This is independent of the kind of cross-lingual fine-tuning that we attempted, with the encoder-only fine-tuning working spectacularly badly (although it still leads to improvements for English).

5 Conclusions

We have described an application of a spoken dialogue system for aphasia patients that aims to improve spoken interaction skills through training. We have tested the setup with the Wizard-of-Oz setup and have used the resulting patient speech to test existing speech recognition models as speech recognition quality is arguably the bottleneck of a spoken dialogue system that interacts with aphasia patients.

We find that the Whisper model, as a state-of-the-art model, handles aphasic speech input moderately well, similarly for our data as for pre-existing data from AphasiaBank.

We also fine-tune the Whisper model. We find an improvement when in-language fine-tuning is performed with German data. However, our expectation that a model that is fined-tuned with English aphasia data leads to a cross-lingual transfer learning of aphasia phenomena turns out to be wrong, or at least possible improvements in aphasia-related phenomena are drowned in the model attempting to hear English speech.

Fine-tuning with both English and German data lead to best results in-domain (i.e., for the corpora involved) which we take as indication that multi-lingual fine-tuning for aphasia phenomena may indeed work. While at least some speech data has been collected for Dutch [22], the third big West-Germanic language and arguably closer to German than English, we have not been successful in accessing this data. It may also be worthwhile to investigate more closely if fine-tuning different parts of the model (encoder vs. decoder) with different data (German non-aphasic for the decoder, any aphasic speech for the encoder) leas to better outcomes.

Even if we cannot substantially improve speech recognition quality for German aphasic speech, dialog training will still be possible. For one, we may integrate approaches of semantic similarity which is useful for understanding of aphasia patients [23]. Furthermore, our system will often know what a patient is supposed to say given that it knows the patient's task instructions. Therefore, error-free speech recognition is not a strict requirement for our approach.

Acknowledgements

We thank the 3 participants in our study who not only eagerly worked on the tasks as described above. One of them even asked if he could continue and buy more pastries in our toy bakery, indicating quite successful gamification of the training task.

References

[1] SINANOVIĆ, O., Z. MRKONJIĆ, S. ZUKIC, M. VIDOVIĆ, and K. IMAMOVIĆ: *Post-stroke language disorders*. *Acta clinica Croatica*, 50, pp. 79–94, 2011.

[2] ENGELTER, S. T., M. GOSTYNSKI, S. PAPA, M. FREI, C. BORN, V. AJDACIC-GROSS, F. GUTZWILLER, and P. A. LYRER: *Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis*. *Stroke*, 37(6), pp. 1379–1384, 2006. doi:10.1161/01.STR.0000221815.64093.8c.

[3] *Fachkräfteengpassanalyse 2023*. Tech. Rep., Bundesagentur für Arbeit, Statistik/Arbeitsmarktberichterstattung, 2024.

[4] HEIDE, J., 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, 2023. doi:10.3389/fcomm.2023.1210193.

[5] LIN, Y., P. KLUMPP, J. PFAB ET AL.: *Eine automatische Sprachbewertung für die neolexon Aphasie-App mithilfe künstlicher Intelligenz. XXXIV. Workshop Klinische Linguistik. Sprachtherapie aktuell: Forschung - Wissen – Transfer*, 9, pp. e2022–11, 2022.

[6] SWELLER, J.: *Cognitive load theory, learning difficulty, and instructional design*. *Learning and Instruction*, 4, pp. 295–312, 1994.

[7] SCHLÖGL, S., G. DOHERTY, and S. LUZ: *Wizard of oz experimentation for language technology applications: Challenges and tools*. *Interacting with Computers*, 27(6), p. 592–615, 2014. doi:10.1093/iwc/iwu016.

[8] KOLLER, A., T. BAUMANN, and A. KÖHN: *DialogOS: Simple and extensible dialog modeling*. In *Proceedings of Interspeech,*, vol. Show and Tell Session. Hyderabad, India, 2018. doi:10.22028/D291-36154.

[9] BOERSMA, P.: *Praat, a system for doing phonetics by computer*. *Glot international*, 5(9/10), pp. 341–345, 2002.

[10] MACWHINNEY, B., D. FROMM, M. FORBES, and A. HOLLAND: *Aphasiabank: Methods for studying discourse*. *Aphasiology*, 25, pp. 1286–1307, 2011.

[11] MACWHINNEY, B.: *The talkbank project*. In *Creating and digitizing language corpora: Volume 1: Synchronic databases*, pp. 163–180. Springer, 2007.

[12] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. 2022. doi:10.48550/ARXIV.2212.04356.

[13] VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, and I. POLOSUKHIN: *Attention is all you need*. 2023. doi:10.48550/arXiv.1706.03762.

[14] PASZKE, A., S. GROSS, F. MASSA, A. LERER, J. BRADBURY, G. CHANAN, T. KILLEEN, Z. LIN, N. GIMELSHEIN, L. ANTIGA, A. DESMAISON, A. KÖPF, E. YANG, Z. DEVITO, M. RAISON, A. TEJANI, S. CHILAMKURTHY, B. STEINER, L. FANG, J. BAI, and S. CHINTALA: *PyTorch: An imperative style, high-performance deep learning library*. 2019. doi:10.48550/arXiv.1912.01703.

[15] OSTENDORFF, M. and G. REHM: *Efficient language model training through cross-lingual and progressive transfer learning*. 2023. doi:10.48550/arXiv.2301.09626.

[16] PEKAREK ROSIN, T. and S. WERMTER: *Replay to remember: Continual layer-specific fine-tuning for German speech recognition*. In L. ILIADIS, A. PAPALEONIDAS, P. ANGELOV, and C. JAYNE (eds.), *Artificial Neural Networks and Machine Learning – ICANN 2023*, pp. 489–500. Springer Nature Switzerland, Cham, 2023.

[17] LIU, Y., X. YANG, and D. QU: *Exploration of Whisper fine-tuning strategies for low-resource ASR. EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1), p. 29, 2024. doi:10.1186/s13636-024-00349-3.

[18] CONNEAU, A., M. MA, S. KHANUJA, Y. ZHANG, V. AXELROD, S. DALMIA, J. RIESA, C. RIVERA, and A. BAPNA: *FLEURS: Few-shot learning evaluation of universal representations of speech*. 2022. doi:10.48550/arXiv.2205.12446.

[19] TORGBI, M., A. CLAYMAN, J. J. SPEIGHT, and H. TAYYAR MADABUSHI: *Adapting Whisper for regional dialects: Enhancing public services for vulnerable populations in the United Kingdom*. In Y. SCHERRER, T. JAUHIAINEN, N. LJUBEŠIĆ, P. NAKOV, J. TIEDEMANN, and M. ZAMPIERI (eds.), *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 29–38. Association for Computational Linguistics, Abu Dhabi, UAE, 2025. URL https://aclanthology.org/2025.vardial-1.4/.

[20] ABAD, A., P. BELL, A. CARMANTINI, and S. RENAIS: *Cross lingual transfer learning for zero-resource domain adaptation*. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6909–6913. 2020. doi:10.1109/ICASSP40776.2020.9054468.

[21] RYKOVA, E., M. WALTHER, and E. ZEUNER: *aphaDIGITAL – avatar-based digital speech therapy solution for aphasia patients: first evaluation*. In *Finnic Phonetics Symposium. Book of Abstracts*, p. 27. Joensuu, Finland, 2022. doi:10.13140/RG.2.2.14038.93765.

[22] WESTERHOUT, E. and P. MONACHESI: *A pilot study for a corpus of Dutch aphasic speech (CoDAS)*. In N. CALZOLARI, K. CHOUKRI, A. GANGEMI, B. MAEGAARD, J. MARIANI, J. ODIJK, and D. TAPIAS (eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy, 2006. URL https://aclanthology.org/L06-1417/.

[23] RYKOVA, E. and M. WALTHER: *Concept for semantic error analysis in a mobile application for speech and language therapy support*. In C. DRAXLER (ed.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, pp. 127–133. TUDpress, Dresden, 2023. URL https://www.essv.de/pdf/2023_127_133.pdf.