Towards preserving errors: fine-tuning automatic speech recognition for literal transcription of aphasic speech

Frank Tsiwah | Rutger Luinge

Introduction and aims

Spontaneous speech is a rich source of diagnostic, language processing and language usage data in aphasia, and yet it remains underused in both clinical practice and research. This is because it is time consuming to transcribe and analyse. Time and resources are consistently the primary barriers to connected-speech work (Stark et al., 2021; Bryant et al., 2017), with estimates ranging from 1.5–3 hours of transcription per 10 minutes of speech (Boles, 1998) to 60 minutes of combined transcription and analysis per minute of conversation (Armstrong et al., 2007). This constrains sample sizes for studies using spontaneous speech and impedes a holistic assessment and/or treatment of language impairment in functional communication context.

Automatic speech recognition (ASR) offers a potential of automating spontaneous speech transcription. For example, *Whisper*, a state-of-the-art ASR model achieves near-human accuracy on healthy speech but was trained solely of neurotypical speech data and is engineered to "clean up" transcription outputs by auto-correcting all errors. That is, on aphasic speech, it normalises paraphasias to the nearest real word and has been found to hallucinate content on disfluent input (Koenecke et al., 2024), and thus, removing the exact information (speech and language errors) clinicians and researchers need for profiling people with aphasia (PWA).

A clinically useful ASR system must provide literal transcriptions that preserve errors (e.g., lexical and grammatical errors), neologisms, deviant sounds, and partial words – the type of errors central to diagnosis, aphasia subtype classification, and therapy planning. This study therefore addresses three questions: (1) Can an end-to-end ASR model be fine-tuned to transcribe aphasic speech literally? (2) Does it generalise across aphasia subtypes and different recording settings? (3) How does it compare with Whisper v3 as a general-purpose baseline? To our knowledge, this is the first ASR study explicitly designed and trained to enforce literal, error-preserving transcription of aphasic speech.

Method

We used five English language corpora from AphasiaBank (*NEURAL, NEURAL 2, UNH, Wright, Richardson*), containing picture descriptions (e.g., cookie theft, broken window, and cat rescue), narrative (e.g., Cinderella story) and open-ended question tasks from people with aphasia, with time-aligned transcripts and subtype labels.

Because literal transcription requires closely representing non-words and atypical pronunciation data in the transcripts, we first converted each word to an IPA-based phonemic representation using a grapheme to phoneme model. This step allowed us to treat the transcription problem as predicting phoneme strings rather than orthographic words. The audio and phoneme sequences were then aligned at the sentence level using forced alignment. Three pre-trained versions of *wav2vec 2.0* (Baevski et al., 2020) ASR-model: base monolingual (English), multilingual XLSR, and large monolingual (English), were compared. Each model was fine-tuned with identical hyperparameters and data. The best-performing model was then selected and subsequently trained on increasing amounts of data from all five corpora to examine the effect of training-set size and across-PWA variability. Performance was evaluated per aphasia subtype (Wernicke, Broca, Anomic, Conduction, Transcortical Motor) and across clinics (different data repositories on AphasiaBank), using phoneme error rate (PER) and word error rate (WER). Whisper v3 was evaluated on the same test sets under default settings.

Results

Wav2vec 2.0 large – English monolingual was the best-performing model, reaching PERs of 0.066 (controls), 0.170 (fluent aphasia), and 0.351 (non-fluent aphasia) on 8 hours of training data, outperforming the base and XLSR variants on all groups. Note that the numbers are in probabilities ranging from 0 – 1, where 0 and 1 are equivalent to 0% and100% error rates, respectively. At the aphasia subtype level, performance varied in systematic and clinically interpretable ways. For Wernicke's aphasia, the model achieved a PER of 0.076 and a WER of 0.122. For Anomic aphasia, PER was 0.125 and WER 0.212, while Conduction aphasia showed PER 0.150 and WER 0.255. Transcortical Motor aphasia was in between, with PER 0.105 and WER 0.169. As expected, Broca aphasia was the most challenging: PER was 0.267 and WER 0.373. These results are reported in Figure 1. Error analyses using phoneme level confusion matrices showed that many remaining errors involved predictable confusions between acoustically similar sounds, such as stops differing only in voicing (e.g., /ð/ ↔ /θ/, /s/ ↔ /z/).

Discussion

This study shows that literal, error preserving automatic transcription of aphasic speech is feasible with current end to end ASR technology, even when only modest amounts of carefully curated aphasic data are available. By training wav2vec 2.0 to output phoneme sequences instead of words, the system captures non words, sound substitutions and partial words that are usually lost when off the shelf systems "clean up" the signal. In direct comparison with Whisper v3, the fine-tuned model halves word error rates on aphasic speech while preserving clinically relevant error patterns, suggesting that aphasia adapted ASR can add value on top of much larger general-purpose models.

The subtype pattern is also informative. The model's performance on Wernicke and other fluent aphasias suggests that the model can follow highly fluent but error rich speech without "correcting away" neologisms and paraphasias, whereas higher error rates for Broca and other non-fluent profiles likely reflect shorter utterances, more challenging acoustics and greater inconsistency in the manual reference transcripts. Practically, a phoneme-level transcription that preserves errors should reduce the cost of human post-editing of spontaneous speech relative to transcribing from scratch.

There are nonetheless important limitations. The training data are restricted to English and to a small set of clinics and recording setups, so performance cannot be assumed to generalise to other languages or dialects. Phoneme level models depend on a fixed symbol inventory and may not fully capture gradient articulatory deviations or voice quality changes, and all outputs remain machine generated and require human post-editing. Nevertheless, the present work points to a practical path forward: with a reasonable amount of aphasic training material and a phoneme focused design, ASR can produce literal transcriptions that outperform a widely used general purpose model while preserving the linguistic information that matters most in aphasia research and clinical documentation.

References

Armstrong, L., Brady, M., Mackenzie, C., & Norrie, J. (2007). Transcription-less analysis of aphasic discourse: A clinician's dream or a possibility?. *Aphasiology, 21*(3-4), 355-374.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Boles, L. (1998) Conversational discourse analysis as a method for evaluating progress in aphasia: a case report. *Journal of Communication Disorders, 31*, 261–274.

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. Aphasiology, 31(10), 1105-1126.
Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024, June). Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency* (pp. 1672-1681).
Stark, B. C., Dutta, M., Murray, L. L., Bryant, L., Fromm, D., MacWhinney, B., ... & Sharma, S. (2021). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech-Language Pathology, 30*(1S), 491-502.

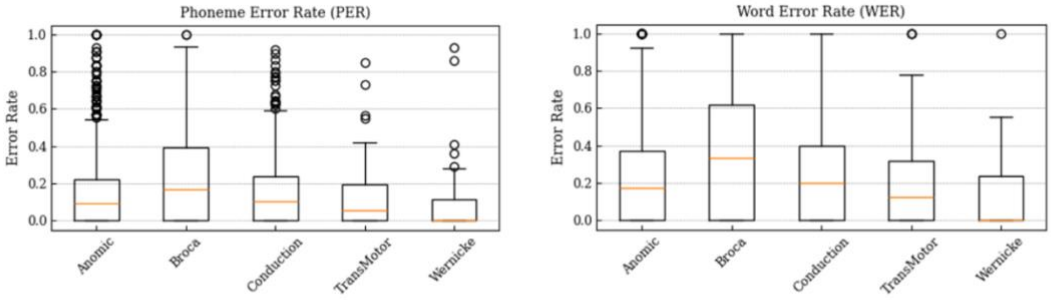


Figure 1. Phoneme error rate (PER, left) and word error rate (WER, right) of the fine-tuned wav2vec 2.0 model for each aphasia subtype (Anomic, Broca, Conduction, Transcortical Motor, Wernicke). The boxplots show how error rates vary across speakers within each subtype, with lower values indicating more accurate, more literal transcriptions.