Uncovering the Potential for a Weakly Supervised End-to-End Model in Recognising Speech from Patient with Post-Stroke Aphasia

Giulia Sanguedolce
Department of Computing,
Department of Electrical &
Electronic Engineering,
Department of Brain Sciences,
Imperial College London

Patrick A. Naylor
Department of Electrical &
Electronic Engineering,
Imperial College London

Fatemeh Geranmayeh
Department of Brain Sciences,
Imperial College London

Abstract

Post-stroke speech and language deficits (aphasia) significantly impact patients' quality of life. Many with mild symptoms remain undiagnosed, and the majority do not receive the intensive doses of therapy recommended, due to healthcare costs and/or inadequate services. Automatic Speech Recognition (ASR) may help overcome these difficulties by improving diagnostic rates and providing feedback during tailored therapy. However, its performance is often unsatisfactory due to the high variability in speech errors and scarcity of training datasets. This study assessed the performance of Whisper, a recently released end-to-end model, in patients with post-stroke aphasia (PWA). We tuned its hyperparameters to achieve the lowest word error rate (WER) on aphasic speech. WER was significantly higher in PWA compared to age-matched controls (10.3% vs 38.5%, $p < 0.001$). We demonstrated that worse WER was related to the more severe aphasia as measured by expressive (overt naming, and spontaneous speech production) and receptive (written and spoken comprehension) language assessments. Stroke lesion size did not affect the performance of Whisper. Linear mixed models accounting for demographic factors, therapy duration, and time since stroke, confirmed worse Whisper performance with left hemispheric frontal lesions. We discuss the implications of these findings for how future ASR can be improved in PWA.

1 Introduction

Aphasia is a language impairment that causes difficulties in speaking, understanding and/or writing coherent and meaningful sentences. This deficit negatively impacts numerous daily activities, such as working, shopping or participating in community and leisure experiences. As a consequence, patients with aphasia report high levels of depression, passiveness, social exclusion and a general decline in their quality of life (Spaccavento et al., 2014). Overall, there are at least 2 000 000 people in the USA (National Aphasia Association) and more than 350 000 people in the UK with aphasia (Stroke Association). Roughly 45% of aphasic disorders arises following a stroke (Ali et al., 2015). Stroke cases, mortality and morbidity have increased substantially over the last two decades, with 70% increase in incident strokes, 43% deaths from stroke, and 143% DALYs Feigin et al. (2022). Consequently, the incidence of aphasia has also increased. Importantly, the presence of aphasia per se worsens the overall stroke outcomes (Lazar and Boehme, 2017; Geranmayeh et al., 2016). Therefore, due to the psycho-social burden and the current increase in stroke cases, early diagnosis and treatment of aphasia need to be addressed.

The mainstay treatment of aphasia is speech and language therapy; it entails practices with language exercises for improving language ability, as well as adjusting to new ways of communicating (Palmer et al., 2018). According to the results of different meta-analyses, higher intensity speech therapy treatment is strongly associated with greater treatment efficacy (Robey 1998; Bhogal et al. 2003; Kelly et al. 2010; Breitenstein et al. 2017). Providing ongoing efficient treatment, however, can be challenging due to limited resources, which can make face-to-face speech therapy costly and difficult to achieve for every patient need (Palmer et al., 2012; Le et al., 2018). The situation became worse especially after the COVID-19 pandemic crisis, that led to the suspension or the slowdown of non-urgent care, including speech and language therapies (Chadd et al., 2021).

A solution for these issues might be the use of speech recognition models, able to remotely and automatically transcribe long pieces of conversation to easily analyse patients language profiles and to give tailored treatments. Nevertheless, even though Automatic Speech Recognition (ASR) tools have been already explored in research, until now these

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 182–190
July 14, 2023 ©2023 Association for Computational Linguistics

have been slow to catch up with the performance obtained in healthy speech (Abad et al., 2013; Le et al., 2016; Jamal et al., 2017; Le et al., 2018). Indeed, the models trained on healthy data struggle to achieve high accuracy in metrics like *Word Error Rate* (WER) or *Phoneme Error Rate* (PER), mostly due to the features of aphasic speech.

Speech from PWA is largely thought to have semantic (meaning) and phonological (speech sound) errors, as well as dysfluencies, each with independent recovery trajectories (Stefaniak et al., 2022). Furthermore, aphasic speech has characteristics that might include: slow and hesitant elocution with episodes of agrammatism (e.g. absence or improper use of function words and verbs - Damico et al. 2010), word-finding problems that affect mostly nouns and picturable action words, frequent stammer, as well as an overall flow of speech that is often fragmented, choppy, unintelligible and/or awkwardly articulated (Abad et al., 2013). These aspects can be influenced also by motor control problems like apraxia and dysarthria, frequently present in aphasia, which may also produce articulation distortion and aberrant prosody (Le et al., 2016). Hence, the challenges that these models need to address include the high variability of speech errors, both between and within aphasic individuals, as well as the lack of satisfactory training datasets.

We tested the performance of a state-of-art sequence-to-sequence ASR transformer model that, to our knowledge, has not been used yet on clinical data. This model, released by Open AI (Radford et al., 2022), is named *Whisper*[1] and it is known for its superior performance in healthy speech when compared with other notable commercial and open-source ASR systems. The feasibility of this model in clinical practice is supported by the low WER in healthy speakers, and the powerful large multilingual weakly supervised dataset on which it was trained. Moreover, the ability to run Whisper locally, will help to preserve the privacy of patients' sensitive data and allow testing in compliance with local and continental regulations. For the purpose of this study, the Whisper testing is done on a novel database of speech of PWA that we have created.

We fine-tuned Whisper parameters relevant for aphasic speech, detailed in Section 3.2. This led us to retrieve the best model according to the lowest WER to test on speech audio. We then compared the patients' WER to an aged-matched control group that performed the same speech production task. After correlation analyses, we created linear mixed-effects models and observed interesting and significant relations with the average performance of the ASR (see Section 4). According to these results, our analysis offers useful insights to consider for our next steps, from which other researchers can also benefit. We expect that our study will advance the work of ASR for PWA, enriching and inspiring the research of the natural language processing community applied in the healthcare framework.

2 Related Work

Since the introduction of ASR technology in clinical studies, algorithms have had to deal with several challenges. The variability and complexity of disordered speech, sometimes unintelligible, has led researchers to move forward with the creation of novel ASR trained with pathological speech data. Nevertheless, an additional difficulty they have to face is the scarcity of datasets of such disordered speech, limiting the accuracy and/or the generalisability of the results. An example of this is the work of Peintner and colleagues (2008), which extracted language features from their corpus for distinguishing different frontotemporal lobar degeneration, one of which includes progressive non-fluent aphasia. Although the study demonstrated encouraging outcomes, it was conducted on a comparatively limited dataset, and no examination was performed regarding the reliability of the features extracted using ASR.

Similarly, Fraser et al. (2013) attempted to differentiate and diagnose primary progressive aphasia (PPA) and two of its sub-types, semantic dementia (SD) and progressive non-fluent aphasia (PNFA) extracting 58 lexical and syntactic features. Using a reduced dataset, an optimized support vector machine (SVM) and random forests (RF) classifiers, Jin and colleagues (2022) tried to face the problem of the dataset with data augmentation on a recognition model for patients with dysarthria. They reached an overall WER of 27.8% on the *UASpeech* test set, underlining that the lowest published WER on the subset of speakers with "Very Low" unintelligibly was of 57.3%.

Differently, Kohlschein and colleagues (2017) used the speech elicited in the *Aachen Aphasia*

[1] The model name comes from the acronym of WSPSR standing for Web-scale Supervised Pretraining for Speech Recognition

Test assessment as a database to train their algorithm. They built a model that automatically analysed pathological speech to identify patients' aphasia type and severity based solely on acoustic features. *AphasiaBank*, a large database open to members, was used by Le and colleagues (2018) to successfully detect medically-relevant quantitative measures to predict aphasia with WER (Word Error Rate) of 39% in spontaneous aphasic speech. Their previous work, with a WER of 45% (Le and Provost, 2016; Le et al., 2017), established the first ASR baseline on AphasiaBank, showing that this dataset can guide the understanding of aphasic speech recognition.

In Le et al. (2017), the authors used an acoustic modelling architecture of multi-task DBLSTM-RNN (double bidirectional long short-Term memory recurrent neural network) with four hidden BLSTM with 2 diverse language models for decoding. The authors investigated features based on speech duration, the quality of pronunciation, phone edit distance, and dynamic time warping on phoneme posteriorgrams. On the other hand, Le et al. (2018), even though using a similar pipeline, chose to investigate lexical diversity and complexity, posteriorgram-based dynamic time warping, pairwise variability error, dysfluency and information density in aphasic speech. Lastly, like the work of Qin et al. (2016), a Cantonese version of AphasiaBank has been implemented by Liu et al. (2018), together with the *CUSENT* and *CanPEV* Cantonese corpora. In this case, as an evaluation metric they used a Syllable Error Rate (SER) with the AphasiaBank and a multilayer time delay neural network (MT-TDNN) with a bidirectional long short-term memory (BLSTM) model structure. This obtained 18.5% of WER for unimpaired speech and 42.4% for impaired speech.

An alternative strategy is to employ ASR that already exists as per study by Mahmoud and colleagues (2023), where the authors customised existing ASR for a specific research goal, selecting Microsoft Azure Speech-to-Text API or Google Speech-to-Text API. In this study we are adopting a similar approach: given the impressive performance of Whisper on healthy speech, largely due to its training dataset being several orders of magnitude larger than preceding ASR, we expect Whisper accuracy to be similar to aforementioned models trained on aphasic data.

Table 1: Sample Characteristics

	Control ($N = 23$)	**Patients** ($N = 23$)
	Mean (Standard Deviation)	
Age (*months*)	59.96 (11.24)	61.45 (10.98)
Gender		
Male	10	14
Female	13	9
Grammatical Complexity*	15.17 (3.59)	9.60 (4.33)
Productivity	127.83 (59.87)	91.41 (53.66)
Lexical Diversity*	60.65 (21.85)	37.01 (16.64)
Fluency***	136.81 (39.66)	68.31 (39.13)
Flawed Syntax (%)**	3.53 (8.60)	37.08 (34.16)

$^{*}: p < 0.05$; $^{**}: p < 0.01$; $^{***}: p < 0.001$

3 Methods

3.1 Dataset

For our study, we used the SONIVA (*Speech recOgNItion Validation in Aphasia*) database, a comprehensive validation database that we are creating for training automated aphasic speech recognition in the research and clinical setting. SONIVA is composed of speech recordings derived from PWA taking part in the IC3[2] study (Imperial Comprehensive Cognitive Assessment in Cerebrovascular Disease; Gruia et al. 2022), and PLORAS study (Predicting Language Outcome and Recovery After Stroke; Seghier et al. 2016). The SONIVA database aims to be a large and comprehensively annotated speech database including quantitative measures of speech and English as well as IPA transcriptions. With this dataset we are producing quantitative summary measures from the *Comprehensive Aphasia Test* (CAT; Swinburn et al. 2004). To understand the various relations with the WER, we included into statistical models the patients' CAT-derived summary measures, quantitative measures of spontaneous speech, size and location of stroke lesion, and demographic factors.

We used as input to Whisper the data of 46 participants, divided into an aged-matched controls group ($N = 23$) and PWA (Patients with aphasia; $N = 23$). For patients, audio speech was collected across multiple time-points since their stroke, resulting in a total of 38 audio files. The speech is recorded during the picture description task from the CAT assessment (Swinburn et al., 2004).

The audio was transcribed verbatim by a speech

[2] https://www.ic3study.co.uk

therapist and 3 trained postgraduate students, with excellent inter-rater reliability (73% overall word-level match). The text is in CHAT format (Codes for the Human Analysis of Transcripts; MacWhinney 2014), managed and analyzed through the CLAN software. Using CLAN, the following measures were generated: *grammatical complexity* (mean length of utterance in morphemes), *productivity* (number of total words), *lexical diversity* (number of different words), *fluency* (words per minute) and *flawed syntax* (incorrect utterances that do not have at least one verb, copula, modal, or participle). All these measures are included in the sample characteristics in table 1, together with the Mann-Whitney tests results in case of significant differences between groups.

3.2 End-to-end Transformer

With 680 000 hours of training on noisy data, of which approximately 20% is derived from non-English languages, its performance on healthy speech has been near human-level with respect to accuracy (Radford et al., 2022). In addition to Whisper's large training dataset, its superior performance is enhanced by the weakly supervised transcription. Its labels are not fully precise or complete, but rather are noisy or partial, because the authors used an ASR to create the labels, which are not perfect and prone to errors. Nevertheless, in order to improve the labels' quality, any text that seemed to be created automatically was discarded. This included the elimination transcriptions that had only upper- or lower-case letters or lacked punctuation, as these were probably generated by machines rather than people. Once they created this dataset, the original version of Whisper was trained and used to understand what was wrong with the data (through error rating metrics) for manually inspecting the low-quality parts and creating an iterative training process.

The model architecture is a sequence-to-sequence transformer, commonly used since 2017 (Vaswani et al., 2017) for its reliability. The audio chunks are initially transformed into an 80-channel, 25 ms window, 10 ms stride Mel spectrogram. The features are scaled between -1 and 1 with a mean of 0 throughout the sample. Interestingly, their multi-task training set has special tokens as task specifiers or classification targets (such as language identification or timestamp tokens). Whisper uses the same byte-level BPE text tokenizer used in GPT-2 (Sennrich et al. 2015; Radford et al. 2019) for the English-only models, as they have both English-only and multi-language models, released in different sizes (from 39M parameters for tiny model to 1.55B parameters for large model).

3.3 Hyperparameters Fine-Tuning

We conducted a grid search fine-tuning, choosing the best performing model based on the WER. Therefore, we took into account the following hyperparameters: 1) the *model size* (base, small, medium, with 74 M, 244 M and 769 M parameters respectively); 2) *'compression_ratio_threshold'* (2.0, 2.4, 2.8, 3.2) and 3) *'logprob_threshold'* (-1.5, -1.0, -0.5, -0.25). These parameters were chosen as they were close to the default values, which are '2.4' for the compression_ratio and '-1.0' for the logprob_threshold.

The 'compression_ratio_threshold' regulates the degree of audio compression on the input speech. In case of PWA speech, low pitch is very frequent so modulating this normalisation parameter may be useful. Whisper used this compression rate during decoding as a criterion for adjusting its temperature parameter, increasing it when the generated text had a compression rate higher than 2.4 (Radford et al., 2022).

On the other hand, the 'log prob_threshold' regulates the required probability to add a new token to the vocabulary of the ASR. This fine-tuning is particularly helpful when in the PWA might appear frequent neologisms (newly coined word). Lower log-probability thresholds could lead to a bigger vocabulary and more accurate compression, but may also increase computational complexity. Also here Whisper used the average log probability over generated tokens as a criterion for adjusting the temperature during decoding, increasing the temperature when the average log probability fell below -1.0. By selecting values of -1.5, -1.0, -0.5, and -0.25 for 'log prob_threshold', it is possible to evaluate how these thresholds impact the balance between exploring alternative options and maintaining reliability in the generated text.

3.4 Evaluation metrics

The evaluation of the ASR performance was done with the WER based on string edit distance, calculating the least number of steps necessary to convert one string from Whisper output to the string from the actual manual transcription. However, since the WER penalizes also innocuous differences, we

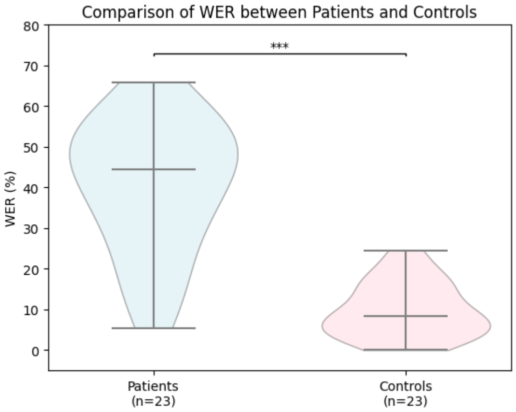


Figure 1: WER distribution density of the *Whisper* model for patients ($N = 23$) and age-matched controls ($N = 23$).
$^{***}: p < 0.001$

had to pre-process the human transcript in CHAT format, similarly to the work of Torre et al. (2021). This procedure is justified as well by the special symbols used that tag phenomena like semantic inconsistencies, repetitions, retracing or sound fragments prevalent in speech from PWA. The CHAT symbols that mark such phenomena have been removed, eliminating also all punctuation.

In the case of neologisms, if the word was not particularly clear, the transcribers would write the literal phonetic alphabet version. For the evaluation of these non-words, they were transformed into the latin alphabet leaving their phoneme sequence unchanged. It should also be noted that human transcriptions included false starts and unique symbols for filler words like "uhuh", "um", and other isolated sounds or interjections, which we decided to preserve since it is a peculiarity of speech from PWA. For the group comparison, we extracted only the participants lines for both the human and ASR transcription, deleting the assessor or carer speech.

3.5 Statistical Analysis

Before modelling the data, to understand the performance difference of Whisper across groups, we compared the WER of patients and controls. In case of repeated measurements, to derive descriptive statistics, we averaged over sessions and then over participants to obtain group characteristics. Instead, the models considered all available information without losing any variability of the data. All the summary outcomes took into account the specific observation weights (e.g. the length of the speech in each audio sample). Due to the non-normality of the distributions and the fact that the samples were independent, we used a Mann-Whitney test.

In addition, a correlation analysis was conducted to establish significant relationships between the WER and CAT scores, as well as the lesion features. Due to the continuous variables considered, we used Pearson correlation coefficients. Furthermore, to pinpoint the associations between our main variables of interest, we used linear mixed-effects models, able to take into account the characteristics of the samples such as repeated measurements and unbalanced data, as well as adjusting the results for potential confounders (Fitzmaurice et al., 2012).

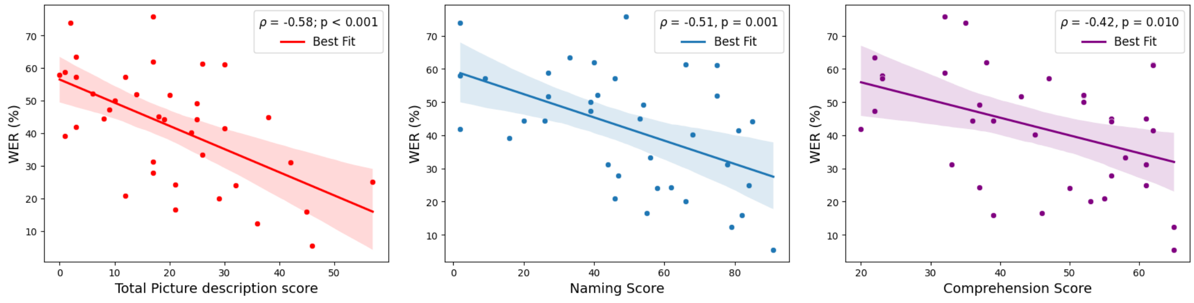
4 Results

Through the grid search optimisation, we generated a total of 48 models, obtained by the combinations of the three aforementioned parameters. The model that performed best, according to the lowest WER, was the one that used the *medium* model, with *compression_ratio_threshold* at 2.0, and *logprob_threshold* at -1.5. The WER differed significantly between controls and patients ($U = 497, \quad p < 0.001$, fig.1), patients had an almost four-fold increase in WER than the control group (38.5% vs 10.3%).

Considering the correlation analysis, CAT scores and WER associations were all found to be significant and they are shown in figure 2. All the three scores of CAT showed a negative relation with the outcome, reflecting in general a worse precision of the ASR in the case of patients with more severe aphasia. As far as the stroke lesion volume is concerned, no significant correlation was found.

Since we wanted to adjust results for potential confounders and find significant and meaningful relations, we modelled the data with mixed-effect models and reported the outcomes in table **??**. In total, four main clusters of models were run to evaluate the effects of *lesioned hemisphere*, *lesion presence* considered singularly, *CAT scores*, and *lesion volume* on the abilities of the ASR to transcribe correctly the speech. All the models were adjusted for socio-demographic (age, gender and years of education) and aphasia-related information (time of test since stroke and hours of speech therapy).

Comparing patients with lesions in left and right hemispheres, the ASR performed worse in terms of WER in left *temporal* and *frontal* lobes, as well as in the left parietal lobe, although this last as-



(a)

(b)

(c)

Figure 2: Association between Word Error Rate of the ASR and the patients' CAT results for (a) *Total Picture description score*, (b) *Naming Score* and (c) *Comprehension Score*.

sociation was not statistically significant. Testing individual lobes confirmed that ASR performance is worse in patients with left frontal lobe lesions, linking it to the localization of expressive language. Moreover, even considering these models, the lesion volume features did not show any significant result.

Finally, CAT derived *Total Picture description score*, *Naming Score* and *Comprehension Score* all had negative relations with the WER, representing higher errors when patients performed poorly in the tests.

5 Discussion

The evaluation of the ASR using the WER metric allowed us to understand how well Whisper, a model trained on a very large healthy speech dataset, performs on PWA speech. We were able to optimise the performance of Whisper based on three hyperparameters, observing similar outcomes in terms of WER when comparing the performance of fine-tuned Whisper model with the performances of previously described ASR systems tailored for PWA.

Using two measures of overt speech production (CAT naming and CAT Total Picture description score for spontaneous speech production) and a measure of speech comprehension (CAT comprehension score), we were able to show that ASR performances is related to the severity of aphasia. These results were confirmed by the mixed-effect models when adjusting for confounding factors such as demographics, time since stroke or duration of therapy. Our findings are in keeping with the study by Torre et al. (2021) that reported 55.5% WER in severe and 22% in mild cases of aphasia.

Furthermore, we showed for the first time that stroke lesion location is related to the performance of the ASR. Speech from patients with left lateralised lesions, and more specifically in the left frontal lobe, was the hardest to recognise using Whisper. This result is consistent with the known localisation of spoken language processing in the brain. Specifically, frontal lobes, together with other parts of the language network, are thought to be primarily implicated in higher-order language functions, such as sentence comprehension, production, speech planning and overt speech production (Geranmayeh et al., 2014). Temporal lobes are essential for language processing and retrieval of semantic information during overt naming (Binder et al. 2020; Binney et al. 2010). Future studies can use information about stroke lesion or brain anatomy to improve ASR training and performance in PWA.

Qualitatively, we noted in some cases Whisper was capable of transcribing filler words (such as "hum", "umm"), frequently observed in PWA. Despite this, the WER occasionally increased as a result of the frequent usage of fillers. False starts (e.g. 'The k- kit- umm... the kitty') were rarely detected and transcribed correctly. There were cases when some words were uttered with low speech volume and were not detected at all, as well as unintelligible words that were skipped altogether by Whisper. These qualitative observations need to be validated with quantitative analysis on larger aphasia-specific datasets to identify PWA speech features that contribute to the worse performance of ASR in PWA. The 'confidence' of the ASR in detecting these aspects can accordingly be reduced, and more specific ASR training can be performed on speech encompassing these specific features.

	WER %				
	Estimate	s.e.	p	95% CI	$\sigma^2_{group}(s.e.)$
Hemisphere Lesion (Left vs Right)					
Temporal (Left)	26.72	12.98	**0.040**	[1.28, 52.16]	137.28 (11.18)
Parietal (Left)	19.79	16.70	0.236	[-12.95, 52.53]	162.68 (12.43)
Frontal (Left)	38.22	13.32	**0.004**	[12.12, 64.32]	128.92 (10.44)
Brain Lobe Lesioned (Yes vs No)					
Temporal					
Left (Yes)	6.96	8.21	0.397	[-9.14, 23.05]	178.74 (12.56)
Right (Yes)	-25.98	11.95	**0.030**	[-49.41, -2.55]	126.47 (9.37)
Parietal					
Left (Yes)	15.54	9.37	0.097	[-2.83, 33.91]	130.29 (10.58)
Right (Yes)	-16.47	15.80	0.297	[-47.44, 14.50]	165.17 (11.86)
Frontal					
Left (Yes)	28.56	10.23	**0.005**	[8.52, 48.61]	92.86 (8.37)
Right (Yes)	-25.98	11.95	**0.030**	[-49.41, -2.55]	126.47 (9.37)
Language Assessments					
Total Picture description score	-0.81	0.23	**0.000**	[-1.26, -0.37]	88.37 (6.98)
Naming Score	-0.32	0.15	**0.037**	[-0.62, -0.02]	109.48 (8.33)
Comprehension Score	-0.64	0.31	**0.041**	[-1.25, -0.03]	165.84 (11.12)
Lesion Volume					
Left Hemisphere Lesion	0.25	0.28	0.378	[-0.30, 0.80]	171 (12.1)
Right Hemisphere Lesion	-1.59	1.28	0.215	[-4.09, 0.92]	161.24 (11.52)
Total Volume	0.20	0.30	0.512	[-0.39, 0.79]	173.1 (12.38)

Table 2: Results of Linear Mixed-Effect regressions on *Hemisphere Lesioned*, the exact *location* of the lesion, *Language Assessments*, and *Lesion Volume*. The models are adjusted for socio-demographic factors (age, gender, and years of education) and aphasia-related information (time of test since stroke and hours of speech therapy).

6 Conclusion and Future Work

This study evaluated the performances of the Whisper end-to-end ASR model on speech derived from patients with post-stroke aphasia. The results highlight the importance of taking lesion location and stroke severity into account when developing speech therapy diagnostics or interventions for PWA using ASR models. Our findings require verification in larger speech databases derived from patients with post-stroke aphasia and their generalisability needs to be assessed in cases of aphasia resulting from other conditions, such as neurodegenerative dementias, which may have different characteristics.

Despite fine-tuning the in-built Whisper parameters to optimise the model performance in this clinical population, we demonstrated that even though Whisper has a competitive performance compared to existing aphasia-specific ASR, it still lacks sufficient clinical diagnostics accuracy. Furthermore, additional ASR metrics such as the confidence of the ASR transcription or the Phoneme Error Rate could be adopted in future research. A further limitation of this work is the small speech database used in this paper. We are actively building a detailed annotated large speech and language database from hundreds of patients with post-stroke aphasia, with the aim of training and developing ASR for pathological speech. We expect that such work will promote greater confidence in the use of AI and specifically NLP for healthcare intervention.

Acknowledgments

G.S. is supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. F.G. is funded through MRC P79100. The authors would like to thank Sophie Brook, Joanna Friedland, Katie Stephenson and Chloe Leong for labelling the data, and the reviewers for their valuable feedback.

References

Alberto Abad, Anna Pompili, Angela Costa, Isabel Trancoso, José Fonseca, Gabriela Leal, Luisa Farrajota, and Isabel P Martins. 2013. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, 27(6):1235–1248.

Myzoon Ali, Patrick Lyden, and Marian Brady. 2015. Aphasia and dysarthria in acute stroke: recovery and functional outcome. *International journal of stroke*, 10(3):400–406.

Sanjit K Bhogal, Robert Teasell, and Mark Speechley. 2003. Intensity of aphasia therapy, impact on recovery. *Stroke*, 34(4):987–993.

Jeffrey R Binder, Jia-Qing Tong, Sara B Pillay, Lisa L Conant, Colin J Humphries, Manoj Raghavan, Wade M Mueller, Robyn M Busch, Linda Allen, William L Gross, et al. 2020. Temporal lobe regions essential for preserved picture naming after left temporal epilepsy surgery. *Epilepsia*, 61(9):1939–1948.

Richard J Binney, Karl V Embleton, Elizabeth Jefferies, Geoffrey JM Parker, and Matthew A Lambon Ralph. 2010. The ventral and inferolateral aspects of the anterior temporal lobe are crucial in semantic memory: evidence from a novel direct comparison of distortion-corrected fmri, rtms, and semantic dementia. *Cerebral cortex*, 20(11):2728–2738.

Caterina Breitenstein, Tanja Grewe, Agnes Flöel, Wolfram Ziegler, Luise Springer, Peter Martus, Walter Huber, Klaus Willmes, E Bernd Ringelstein, Karl Georg Haeusler, et al. 2017. Intensive speech and language therapy in patients with chronic aphasia after stroke: a randomised, open-label, blinded-endpoint, controlled trial in a health-care setting. *The Lancet*, 389(10078):1528–1538.

Katie Chadd, Kathryn Moyse, and Pam Enderby. 2021. Impact of covid-19 on the speech and language therapy profession and their patients. *Frontiers in Neurology*, 12:629190.

Jack S Damico, Nicole Müller, and Martin John Ball. 2010. *The handbook of language and speech disorders*. Wiley Online Library.

Valery L Feigin, Michael Brainin, Bo Norrving, Sheila Martins, Ralph L Sacco, Werner Hacke, Marc Fisher, Jeyaraj Pandian, and Patrice Lindsay. 2022. World stroke organization (wso): global stroke fact sheet 2022. *International Journal of Stroke*, 17(1):18–29.

Garrett M Fitzmaurice, Nan M Laird, and James H Ware. 2012. *Applied longitudinal analysis*. John Wiley & Sons.

Kathleen C Fraser, Frank Rudzicz, and Elizabeth Rochon. 2013. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Interspeech*, pages 2177–2181.

Fatemeh Geranmayeh, Robert Leech, and Richard JS Wise. 2016. Network dysfunction predicts speech production after left hemisphere stroke. *Neurology*, 86(14):1296–1305.

Fatemeh Geranmayeh, Richard JS Wise, Amrish Mehta, and Robert Leech. 2014. Overlapping networks engaged during spoken language production and its cognitive control. *Journal of Neuroscience*, 34(26):8728–8740.

Dragos Gruia, Sabia Combrie, and Fatemeh Geranmayeh. 2022. Novel unsupervised comprehensive tool for monitoring vascular cognitive impairment following stroke. In *Alzheimer's Association International Conference*. ALZ.

Norezmi Jamal, Shahnoor Shanta, Farhanahani Mahmud, and MNAH Sha'abani. 2017. Automatic speech recognition (asr) based approach for speech therapy of aphasic patients: A review. In *AIP Conference Proceedings*, volume 1883, page 020028. AIP Publishing LLC.

Zengrui Jin, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shujie Hu, Jiajun Deng, Guinan Li, and Xunying Liu. 2022. Adversarial data augmentation using vae-gan for disordered speech recognition. *arXiv preprint arXiv:2211.01646*.

Helen Kelly, Marian C Brady, and Pam Enderby. 2010. Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*, (5).

Christian Kohlschein, Maximilian Schmitt, Björn Schüller, Sabina Jeschke, and Cornelius J Werner. 2017. A machine learning based system for the automatic evaluation of aphasia speech. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE.

Ronald M Lazar and Amelia K Boehme. 2017. Aphasia as a predictor of stroke outcome. *Current neurology and neuroscience reports*, 17(11):1–5.

Duc Le, Keli Licata, Carol Persad, and Emily Mower Provost. 2016. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM transactions on audio, speech, and language processing*, 24(11):2187–2199.

Duc Le, Keli Licata, and Emily Mower Provost. 2017. Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298.

Duc Le, Keli Licata, and Emily Mower Provost. 2018. Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, 100:1–12.

Duc Le and Emily Mower Provost. 2016. Improving automatic recognition of aphasic speech with aphasiabank. In *Interspeech*, pages 2681–2685.

Yuanyuan Liu, Ying Qin, Siyuan Feng, Tan Lee, and PC Ching. 2018. Disordered speech assessment using kullback-leibler divergence features with multi-task acoustic modeling. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 61–65. IEEE.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Seedahmed S Mahmoud, Raphael F Pallaud, Akshay Kumar, Serri Faisal, Yin Wang, and Qiang Fang. 2023. A comparative investigation of automatic speech recognition platforms for aphasia assessment batteries. *Sensors*, 23(2):857.

Website National Aphasia Association. 2016. Aphasia statistics - national survey on aphasia awareness. Accessed: 2022-12-2.

Rebecca Palmer, Pam Enderby, Cindy Cooper, Nick Latimer, Steven Julious, Gail Paterson, Munyaradzi Dimairo, Simon Dixon, Jane Mortley, Rose Hilton, et al. 2012. Computer therapy compared with usual care for people with long-standing aphasia poststroke: a pilot randomized controlled trial. *Stroke*, 43(7):1904–1911.

Rebecca Palmer, Helen Witts, and Timothy Chater. 2018. What speech and language therapy do community dwelling stroke survivors with aphasia receive in the uk? *PloS one*, 13(7):e0200096.

Bart Peintner, William Jarrold, Dimitra Vergyri, Colleen Richey, Maria Luisa Gorno Tempini, and Jennifer Ogar. 2008. Learning diagnostic models using speech and language measures. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4648–4651. IEEE.

Ying Qin, Tan Lee, Anthony Pak Hin Kong, and Sam Po Law. 2016. Towards automatic assessment of aphasia speech using automatic speech recognition techniques. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–4. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Randall R Robey. 1998. A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language, and Hearing Research*, 41(1):172–187.

Mohamed L Seghier, Elnas Patel, Susan Prejawa, Sue Ramsden, Andre Selmer, Louise Lim, Rachel Browne, Johanna Rae, Zula Haigh, Deborah Ezekiel, et al. 2016. The ploras database: a data repository for predicting language outcome and recovery after stroke. *Neuroimage*, 124:1208–1212.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Simona Spaccavento, Angela Craca, Marina Del Prete, Rosanna Falcone, Antonia Colucci, Angela Di Palma, and Anna Loverre. 2014. Quality of life measurement and outcome in aphasia. *Neuropsychiatric disease and treatment*, 10:27.

James D Stefaniak, Fatemeh Geranmayeh, and Matthew A Lambon Ralph. 2022. The multidimensional nature of aphasia recovery post-stroke. *Brain*, 145(4):1354–1367.

Website Stroke Association. No date. Aphasia and its effects. Accessed: 2022-10-2.

Kate Swinburn, Gillian Porter, and David Howard. 2004. Comprehensive aphasia test.

Iván G Torre, Mónica Romero, and Aitor Álvarez. 2021. Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasiabank for english and spanish. *Applied Sciences*, 11(19):8872.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.