Application of Whisper in Clinical Practice: the Post-Stroke Speech Assessment during a Naming Task

Milena Davudova[1]* Ziyuan Cai[1]* Valentina Giunchiglia[1,2] Dragos C. Gruia[2]
Giulia Sanguedolce[2] Adam Hampshire[1]† Fatemeh Geranmayeh[2]†
[1]Department of Neuroimaging, King's College London, UK [2]Department of Brain Sciences, Imperial College London, UK
{milena.davudova, ziyuan.1.cai, adam.hampshire}@kcl.ac.uk
{v.giunchiglia20, dragos-cristian.gruia19, g.sanguedolce22, fatemeh.geranmayeh00}@imperial.ac.uk

Abstract

Detailed assessment of language impairment following stroke remains a cognitively complex and clinician-intensive task, limiting timely and scalable diagnosis. Automatic Speech Recognition (ASR) foundation models offer a promising pathway to augment human evaluation, but their effectiveness in the context of speech and language impairment remains uncertain. In this study, we evaluate whether Whisper, a state-of-the-art ASR foundation model, can be applied to transcribe and analyze speech from patients with stroke during a picture-naming task. We assess both verbatim transcription accuracy and the model's ability to support downstream prediction of language function, which has major implications for outcomes after stroke. Our results show that the baseline Whisper model performs poorly on single-word speech utterances. Nevertheless, fine-tuning Whisper significantly improves transcription accuracy (reducing Word Error Rate by 87.72% in healthy speech and 71.22% in speech from patients). Further, learned representations from the model enable accurate prediction of speech quality (average F1 Macro of 0.74 for healthy, 0.75 for patients). However, evaluations on an unseen (TORGO) dataset reveal limited generalizability, highlighting the inability of Whisper to perform zero-shot transcription of single-word utterances on out-of-domain clinical speech and emphasizing the need to adapt models to specific clinical populations. While challenges remain in cross-domain generalization, these findings highlight the potential of foundation models, when appropriately fine-tuned, to advance automated speech assessment and rehabilitation for stroke-related impairments.

1 Introduction

Stroke is one of the leading causes of adult death and disability worldwide, with global incidence of 12.2 million per year [1]. One of the most debilitating and common consequences after stroke is aphasia, an acquired disorder of speech and language production and comprehension, with over 30% prevalence in patients with stroke [2–4]. Due to high heterogeneity of symptoms, diagnosis of aphasia requires an in-depth assessment to uncover specific impairments and tailor speech therapy to individual patient's needs. Computerized assessments provide a cost-effective and accessible platform for post-stroke speech evaluation [5], but they often depend on manual transcription and scoring by trained clinicians, which is time-consuming, prone to inconsistencies, and difficult to scale.

Advancements in automatic speech recognition (ASR) methods can fill this gap by automating the speech evaluation pipelines. The extensive linguistic knowledge embedded in foundation ASR

*Equal contribution.

†Co-senior and co-corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

models, acquired through pretraining on large and diverse datasets, has the potential to be leveraged in a zero and few shot learning context to automate post-stroke speech evaluation pipelines [6, 7]. However, whilst the latest ASR models perform exceedingly well on healthy speech data [8], several limitations have hindered the translation of ASR to pathological speech analysis. Firstly, current ASR models have been trained on healthy speech, limiting their generalizability to pathological speech. Secondly, the potential variability in speech impairments in aphasia necessitates large clinical speech databases for effective training of ASR models, which are currently relatively limited.

In this study, we aimed to assess the performance of Whisper [8], a state-of-the-art transformer-based ASR model pretrained on 680,000 hours of multilingual data, on a stroke-specific speech database derived from a commonly used Naming task. We fine-tuned Whisper on speech from age-matched healthy older adults and patients with stroke and compared the verbatim transcription performance of the fine-tuned and the baseline Whisper models. Further, we evaluated whether Whisper-derived representations could be effectively applied to clinically-relevant downstream tasks, specifically speech impairment severity prediction. Lastly, we assessed the quality of these predictions by conducting a divergent and convergent validity analysis with patients' known clinical features.

2 Related Work

Current approaches for automated speech impairment analysis involve three key steps. First, speech-to-text ASR models are employed to generate speech transcriptions from audio recordings. Then, a feature extraction process is applied to the produced transcriptions and audio data. The extracted lexico-acoustic features are used in downstream analyses, such as classification and regression, to estimate the degree of speech impairment [9]. Different algorithms have been used for each step.

2.1 Recurrent Neural Networks

Previous studies have trained customized ASR models to transcribe aphasic speech, subsequently deriving acoustic and transcription-based features to estimate aphasia severity [10, 9]. Le and colleagues [10] utilized a Bidirectional Long-Short Memory-Recurrent Neural Network (BLSTM-RNN) model trained on acoustic features of aphasic speech derived from the AphasiaBank dataset [11] and frame-level senone and monophone labels, obtained through a Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) system, to produce speech transcriptions. Similarly, Qin and colleagues [9] utilized a Time-Delay Neural Network combined with a Bidirectional Long-Short Term Memory (TDNN-BLSTM) model to transcribe Cantonese aphasic speech. The model was trained on acoustic features derived from healthy speech and HMM-GMM-derived triphone target labels [9]. The ASR transcriptions were used to extract features based on text alone [10] or in combination with audio features [9] for downstream estimation of aphasia severity by regression. While the above studies reached promising accuracy, the complex architectures of employed ASR systems required a multi-step training process, limiting global optimization [12]. Moreover, the models were based on RNN variants, which are less effective than Transformers in modeling long-range dependencies [13].

2.2 Transformers

Transformers have revolutionized the field of ASR by implementing self-attention mechanisms, which capture temporal relationships across the entire input sequence simultaneously, thus overcoming the range limitations of RNN-based architectures and enabling more efficient encoding of contextual information [14, 15]. Multiple ASR transformer-based models are currently available for speech-to-text transcription including Transformer Transducer [16], wav2vec 2.0 [17], SpeechStew [18], Conformer [19] and Whisper [8]. However, while transformer-based models demonstrate exceptional performance on healthy speech ASR [8], they often encounter challenges adapting to aphasic speech recognition due to the high heterogeneity in lexico-acoustic features of post-stroke speech. Recent studies have leveraged pretrained transformer-based ASR models, adapting them to enhance performance on aphasic speech. Within this framework[20], XLSR-53, a model built upon wav2vec 2.0 architecture, was fine-tuned to adapt it to aphasic speech transcription. XLSR-53 [21], pretrained on the extensive dataset which contains 56,000 hours of unlabeled audio, uses convolutional neural network layers to generate latent speech representations [17]. These representations are then passed through a transformer network to encode contextualized speech features [17]. However, wav2vec 2.0 architecture is outperformed by other transformer-based foundation models such as

Whisper [8, 22]. Whisper-derived representations have been successfully utilized in previous studies to classify dysfluencies in stuttering [23] and to estimate dysarthria severity [24]. Further, Whisper and XLSR-53 model transcripts have been successfully used for relevant speech feature extraction, later used for aphasia type classification, reaching an average F1 score of 90.6 [25]. Recent work also investigated hyperparameter tuning and fine-tuning of Whisper to better fit aphasic speech, reaching promising results of 38.5% and 21.93% WER on aphasic speech transcription, respectively [26, 27]. However, the vast majority of previous work has focused on sentential speech, and the applications of Whisper to single-word utterances remain to be studied.

3 Method

3.1 Dataset

Speech data from age-matched controls and patients with stroke were derived from participants undergoing speech and cognitive testing as part of the the IC3 study [5]. Specifically the data was obtained from the Naming task, modified from the 30-item Boston Naming Task [28], where participants were instructed to name a depicted black and white line drawing picture with a single word. The speech recordings from this task, amongst other speech production tasks, have contributed to a larger speech database SONIVA (Speech Recognition Validation in Aphasia) [29]. The data used in this study will henceforth be referred to as SONIVA-Naming. The accuracy of the Naming task performance was evaluated across each trial using four accuracy metrics: semantic content (Semantic), fluency of articulation (Dysfluency), presence of self-corrections in the response (Self-correction) and phonological correctness (Phonology). Each trial was scored by trained raters on a 3-point scale: 0, 1, or 2, where 2 corresponds to the highest accuracy. The verbatim transcriptions of the audio files ranged from single word to sentences, depending on the participants responses. The inter-rater reliability against a qualified speech therapist was high (Mean Intraclass Correlation Coefficient = $0.83\pm$ 0.0046), and all ambiguous cases were resolved by consulting a speech therapist. In total, the SONIVA-Naming healthy dataset comprised 3960 Naming task trials from 132 healthy participants, while the SONIVA-Naming patient dataset consisted of 2609 trials from 87 patients. Demographic characteristics of the sample are described in Appendix A. There was no group difference for age (Mann-Whitney U statistic: $5542.50, p = 0.88$). Patients had lower education (χ^2=26.19, $p < 0.001$) and a higher proportion of non-native English speakers (χ^2=4.02, $p = 0.04$). A noise-free, single-word, baseline synthetic dataset was additionally generated using the Google Text-to-Speech library for Python [30]. Seven English accents were used to generate a total of 210 synthetic audio files across the 30 stimuli words available in the Naming task.

An unseen dataset was used to assess model generalizability for single-word pathological speech transcription. This dataset was derived from the TORGO database [31], which includes speech from individuals with dysarthria, a motor speech disorder common in cerebral palsy and amyotrophic lateral sclerosis. Single-word recordings were used to evaluate model transcription of short-form data. The TORGO dataset comprised 1,240 utterances from 7 participants (4 male, 3 female).

3.2 Model

OpenAI's Whisper [8] was fine-tuned for the verbatim transcription task. For the accuracy classification task, we employed WhisperForAudioClassification, an adaptation that retains the original Whisper encoder but replaces the decoder with a linear classification head, implemented via the Hugging Face Transformers library [32]. Whisper architecture is described in Appendix B.

3.3 Data Pre-processing

All audio data were resampled to 16 kHz. In the main dataset used for model training and evaluation (SONIVA-Naming), recordings without at least one ground truth accuracy metric (Semantic, Dysfluency, Self-correction, Phonology) were excluded when fine-tuning models for the accuracy score prediction task (n = 23 and n = 114 for SONIVA-Naming healthy and patient datasets, respectively). Data were divided into training, validation, and test sets in a 7:1:2 ratio. Splitting was performed at the participant level such that all trials from the same participants were in the same partition. Due to high class imbalance across accuracy scores, the training sets were randomly downsampled to the minority

class in each metric. This resulted in 144, 204, 90, and 21 audio files for the SONIVA-Naming healthy dataset and 198, 729, 108, and 171 trials for the patient dataset, respectively.

3.4 Model Training

Verbatim Transcription Task Whisper was separately fine-tuned on the training set of synthetic, SONIVA-Naming healthy, and SONIVA-Naming patient data, as well as combined datasets of all participant data (SONIVA-Naming healthy and SONIVA-Naming patient) and all available data (synthetic, SONIVA-Naming healthy and SONIVA-Naming patient) (See Appendix C). Small and Medium Whisper models [33, 34] were used to determine the optimal model size. In total, ten fine-tuned models were obtained: fine-tuned on synthetic data (*ft-syn*), on SONIVA-Naming healthy data (*ft-h*), on SONIVA-Naming patient data (*ft-p*), on all participant data (*ft-hp*) and all available data (*ft-all*), each in Small and Medium sizes.

Models were trained on a single NVIDIA RTX 6000 GPU for a maximum of 1000 steps, with a batch size of 16, which took approximately 2-4 hours. A learning rate of 1×10^{-5} was used, with a linear learning rate scheduler incorporating 250 warm-up steps. The model parameters were updated with AdamW optimizer using Cross-Entropy loss function.

Models were evaluated on the validation set (batch size of 8) at 50 step intervals. The best model was chosen based on Word Error Rate (WER) on the validation set and was used for final evaluation on the test sets. The evaluation was completed on four different test sets, namely the synthetic, SONIVA-Naming healthy, and SONIVA-Naming patient as well as the unseen TORGO database.

Accuracy Prediction Task The setup was formulated as a multi-class classification task, where one of three possible accuracy scores (0, 1 or 2) was predicted for each given Naming task trial recording in the SONIVA-Naming healthy and patient datasets (See Appendix C).

The encoder was initialized either with the baseline pretrained encoder weights (Whisper without fine-tuning) or with the encoder weights of the models fine-tuned for verbatim transcription of healthy and patient speech. The *ft-h* and *ft-p* model weights were used when predicting accuracy scores on healthy and patient data, respectively, to capture the most relevant speech representations of each group. This setup resulted in a total of 32 models: models were trained separately for each accuracy metric (Semantic, Dysfluency, Self-correction, Phonology), with two encoder weight configurations (fine-tuned or baseline), each in two sizes (Small and Medium) [33, 34] for each dataset separately (SONIVA-Naming healthy and SONIVA-Naming patient). During training, the encoder weights were frozen and only the linear classification head was trained (i.e., linear probing).

Models were trained on a single NVIDIA RTX 6000 GPU for a maximum of 8000 steps and with a batch size of 16, which took approximately 6-8 hours. A learning rate of 1×10^{-5} was used, with a linear learning rate scheduler incorporating 500 warm-up steps. Models were evaluated on the validation set at 100 step intervals. The model parameters were updated with AdamW optimizer using Cross-Entropy loss function. The best model was chosen based on F1 Macro metric on the validation set and was used for final evaluation on the test set.

3.5 Evaluation metrics

Word Error Rate (WER) The WER was defined based on the string edit distance, which consists of the ratio of necessary string modifications needed to convert the model prediction into the ground truth label divided by the number of total words spoken (1).

$$WER = \frac{S + I + D}{N} \tag{1}$$

where S is the number of substitutions, I of insertions, D of deletions, and N is the total number of words spoken. The score was then multiplied by 100 to obtain a percentage measure.

F1 Macro Performance on the accuracy score prediction task was assessed based on binarized F1 scores. Specifically, the accuracy score 1 was merged with 0 to derive an impaired class (class 0),

while accuracy score of 2 was converted to 1 (class 1) to represent unimpaired responses. The F1 score for the impaired and unimpaired classes were calculated, and then the average between the two was used to obtain the F1 macro, which was used as the main performance metric (2).

$$F1_{Macro} = \frac{\sum_{i=1}^{n} F1_i}{n} \tag{2}$$

where n is the number of classes.

Target Word Detection Accuracy Verbatim transcriptions of the SONIVA-Naming dataset trials often contained words beyond the target Naming task word. A target word detection accuracy was calculated by assessing whether the predicted transcriptions correctly identified the target word only. A prediction was classified as a True Positive if the target word appeared in both the ground truth and prediction and a True Negative if absent in both. The presence of the target word in the ground truth label but not in the prediction indicated a False Negative and the reverse - a False Positive.

3.6 Statistical Analysis

Model size and type analysis The effect of model type (i.e., baseline vs fine-tuned) and size (i.e., Small vs Medium) on WER across all trials was assessed on the test set using the Friedman test, followed by Mann-Whitney U tests with Bonferroni correction for multiple comparisons. The non-parametric tests were used due to the non-normal distributions of the WER within each dataset.

Predictive Validity Analysis To evaluate the clinical validity of the predicted speech accuracy scores, an overall predicted accuracy score was derived for each patient in the SONIVA-Naming patient test set ($n = 18$) by averaging their predicted trial-by-trial scores for each metric. Patients were classified as either impaired or unimpaired on each predicted accuracy metric. Patients with an overall predicted score smaller or equal to 1 were classified as impaired. Otherwise, they were classified as unimpaired. Impairment status was validated against multiple known clinical and demographic factors. Convergent validity was tested against a manually assessed speech fluency metric, stroke history and English as a second language status – hypothesized to affect predicted impairment status. Divergent validity was tested against sex, low density lipoprotein (LDL) cholesterol levels and smoking status, hypothesized to not show any relationship with predicted impairment status.

To complete the analysis, all categorical variables (e.g., sex, smoking status, English as second language, previous stroke history), were one-hot encoded. The distribution of continuous variables (LDL cholesterol, speech fluency) was tested for normality using the Shapiro-Wilk test. Since the assumption of normality was met, continuous variables between groups were compared with a Student's t-test. Categorical variables were compared between groups using Fisher's exact test.

4 Results

4.1 Verbatim Transcription

A significant effect of model size and type on WER was detected when testing on synthetic ($\chi^2(11, N=42) = 313.95$, $p < 0.001$), SONIVA-Naming healthy ($\chi^2(11, N=806) = 5215.60$, $p < 0.01$), and SONIVA-Naming patient datasets ($\chi^2(11, N=524) = 2442.80$, $p < 0.01$). All fine-tuned models significantly outperformed baseline Whisper in transcription accuracy on the synthetic, healthy and patient datasets ($p < 0.01$) (Table 1). Fine-tuning on healthy and patient speech resulted in 87.72% and 71.22% improvement in WER for healthy and patient speech compared to baseline model of the same size. A significant improvement was also observed in case of synthetic data, where the best-performing models were Medium *ft-h*, Medium *ft-p* and Small and Medium *ft-all*, reaching a WER of 0%, compared to a WER of minimum 85.71% in case of Whisper baseline. The best-performing model on the healthy data was the Small *ft-h*, reaching average WER of 8.82%. On the patient dataset, Medium *ft-hp* reached the lowest average WER of 26.35%. However, this Medium model was not

significantly better than its Small counterpart ($p > 0.05$). Additionally, when evaluated on the patient dataset, the difference in performance between *ft-hp* and single-dataset trained models *ft-h* and *ft-p* was insignificant ($p > 0.05$) in both model sizes.

On the unseen TORGO database, a significant effect of model size and type on WER was also detected ($\chi^2(11, N{=}1240) = 850.08$, $p < 0.001$) (Table 1). All fine-tuned models significantly outperformed baseline Whisper in verbatim transcription accuracy ($p < 0.01$). The best performing model was Medium *ft-syn*, which achieved a 22.64% reduction in WER compared to the baseline model in the same size. However, it did not perform significantly better than other fine-tuned models ($p > 0.5$), with the exception of Medium *ft-hp* ($p = 0.045$).

Table 1: Comparison of word error rate (WER) across testing datasets.

Word error rate (%)					
Model	**Size**	**Synthetic**	**SONIVA healthy**	**SONIVA patient**	**TORGO**
Baseline	Small	92.85	96.54	100.54	97.82
	Medium	85.71	90.87	97.57	96.13
ft-syn	Small	45.23	53.64	75.83	77.72
	Medium	21.42	48.06	71.53	**73.49**
ft-h	Small	2.38	**8.82**	28.38	77.39
	Medium	**0**	9.89	27.72	75.65
ft-p	Small	7.14	14.60	28.82	75.01
	Medium	**0**	11.94	26.80	74.62
ft-hp	Small	2.38	12.69	27.79	78.92
	Medium	2.38	12.58	**26.35**	79.67
ft-all	Small	**0**	12.00	27.04	78.10
	Medium	**0**	9.07	29.15	76.58

4.2 Target Word Detection accuracy

Compared to baseline Whisper, fine-tuned models had a higher target word detection accuracy in both healthy and patient speech (Table 2). For healthy speech, the best-performing model was the Small *ft-h*, reaching an accuracy of 0.97. This increase in performance was due to the *ft-h* model making noticeably fewer False Negative mistakes ($n = 27$) compared to baseline Whisper ($n = 496$). For patient speech, the best-performing models were the Small and Medium *ft-hp*, reaching an accuracy of 0.92. This model made fewer False Negative mistakes ($n = 39$) than the Medium baseline model ($n = 286$). However, Small and Medium *ft-hp* models produced 2 False Positive mistakes, compared to 1 and 0 False Positive mistakes in Small and Medium baseline models, respectively. Overall, an improvement of up to 61% and 51% was observed for healthy and patient data, respectively.

4.3 Accuracy Score Prediction

The best-performing models for predicting the Semantic, Dysfluency, and Self-correction accuracy scores during the Naming task on healthy data were the Medium models initialized with *ft-h* weights, achieving F1 Macro scores of 0.7449, 0.8390, and 0.7539, respectively (Table 3). For the Phonology metric, the best-performing model was the Small model also initialized with *ft-h* weights, reaching an F1 Macro of 0.6424. Compared to their respective baselines, an average improvement of 7.24 ± 4.01% was observed across all accuracy metrics. Although all models had moderate to high F1 Macro scores, there was a discrepancy between their performance on correct/unimpaired (class 1) and incorrect/impaired (class 0) trials. The F1 scores for class 1 trials were 0.9814, 0.9564, 0.9860, and 0.9744, whereas the F1 scores for class 0 trials were 0.5085, 0.7215, 0.5217, and 0.3103 for the Semantic, Dysfluency, Phonology, and Self-correction metric models, respectively.

In context of patient speech, the best-performing model for predicting Semantic accuracy scores was the Small model initialised with *ft-p* encoder weights, reaching F1 Macro of 0.7659. For predicting Dysfluency and Self-correction, the best-performing models were the Small models initialized with baseline encoder weights, reaching F1 Macro of 0.9021 and 0.7112, respectively (Table 3). For Phonology, the best model was the Medium model initialized with *ft-p* encoder weights, achieving F1 Macro of 0.6435. The observed improvement of fine-tuned models in Phonology and Semantics was in average 4.425 ± 0.175%. In Dysfluency and Self-correction, the baseline models performed on average 1.45±0.06% better than corresponding models initialized with *ft-p* encoder weights. While the discrepancy between performance on class 1 and class 0 trials was still evident, it was lower than

Table 2: Model target word detection performance on the SONIVA-Naming dataset.

Testing Dataset	Size	Model	Accuracy	True Positive	True Negative	False Positive	False Negative
SONIVA-Naming healthy	Small	Baseline	0.36	256	24	0	496
		ft-syn	0.60	442	24	0	310
		ft-h	**0.97**	725	24	0	27
		ft-p	0.92	693	24	0	59
		ft-hp	0.96	723	24	0	29
		ft-all	0.96	719	24	0	33
	Medium	Baseline	0.40	284	24	0	468
		ft-syn	0.59	432	24	0	320
		ft-h	0.96	724	24	0	28
		ft-p	0.95	716	24	0	36
		ft-hp	0.95	714	24	0	38
		ft-all	0.96	721	24	0	38
SONIVA-Naming patient	Small	Baseline	0.41	126	87	1	310
		ft-syn	0.52	188	87	1	248
		ft-h	0.91	390	85	3	46
		ft-p	0.90	383	86	2	53
		ft-hp	**0.92**	397	86	2	39
		ft-all	0.90	389	85	3	47
	Medium	Baseline	0.45	150	88	0	286
		ft-syn	0.48	165	87	1	271
		ft-h	0.91	393	85	3	43
		ft-p	0.91	389	86	2	47
		ft-hp	**0.92**	397	86	2	39
		ft-all	0.88	375	86	2	61

on healthy data. The F1 scores in class 1 trials were 0.9105, 0.9282, 0.8944 and 0.8604, whereas F1 scores in class 0 were 0.6213, 0.8759, 0.5279 and 0.4265 for Semantic, Dysfluency, Self-correction and Phonology metric models, respectively.

Table 3: F1 Macro for different accuracy metrics and testing datasets.

Testing Dataset	Size	Encoder Weight	Semantic	Dysfluency	Self-correction	Phonology
SONIVA-Naming healthy	Small	Baseline	0.6196	0.7823	0.6124	0.5606
		ft-h	0.6948	0.8375	0.6876	**0.6424**
	Medium	Baseline	0.7094	0.8009	0.6197	0.5334
		ft-h	**0.7449**	**0.8390**	**0.7539**	0.6045
SONIVA-Naming patient	Small	Baseline	0.7234	**0.9021**	**0.7112**	0.5873
		ft-p	**0.7659**	0.8870	0.6973	0.5771
	Medium	Baseline	0.7520	0.9017	0.6775	0.5975
		ft-p	0.7370	0.8891	0.6613	**0.6435**

4.4 Predictive Validity Analysis

Self-correction accuracy scores were excluded from this analysis as only two patients were identified as impaired based on model predictions of this metric, making further statistical analysis impossible.

The predictions of the best-performing accuracy prediction models were used to identify patients as impaired or unimpaired on each metric (Small Whisper initialized with *ft-p* encoder weights, small Whisper initialized with baseline encoder weights, and Medium Whisper initialized with *ft-p* encoder weights for Semantic, Dysfluency and Phonology, respectively).

Patients identified as impaired based on model predictions had significantly lower speech fluency than those identified as unimpaired based on Semantic ($t(16) = -3.61, p = 0.01$), Dysfluency ($t(16) = -2.3880, p = 0.02$) and Phonology ($t(16) = -2.3880, p = 0.02$) metrics. Significantly more impaired patients spoke English as a second language compared to unimpaired patients, based on Dysfluency and Phonology metrics ($p = 0.02$). Additionally, a higher number of patients with previous stroke history were identified as impaired than unimpaired based on Semantic ($p = 0.02$), Dysfluency ($p = 0.04$) and Phonology ($p = 0.04$) metrics. The divergent validity analysis showed no significant differences between impaired and unimpaired patients for sex, LDL cholesterol or smoking status, as expected ($p > 0.05$) (See Appendix E).

5 Discussion

The baseline Whisper model performed substantially worse on the verbatim transcription of single-word SONIVA-Naming healthy speech compared to previously reported results on healthy speech. Previous studies have reported WER as low as 3.40% and 2.90% for the baseline Whisper Small and Medium models, respectively [8]. However, these results were obtained using the LibriSpeech corpus [35], a standard ASR benchmark consisting of continuous speech from audiobook recordings. In contrast, the markedly poorer performance observed in the current study likely reflects the short-form nature of the single-word Naming task-derived speech, which provides limited contextual information for decoding and therefore reduces transcription accuracy.

Although the synthetic speech dataset was free from natural speech variability and background noise, the baseline Whisper models also performed poorly on this dataset (WER = 92.85% and 85.71% for the Small and Medium models, respectively). The observed performance further supports that Whisper does not generalize directly to short-form speech and highlights the necessity of task-specific fine-tuning. Further, once fine-tuned, all models achieved their best performance on the synthetic dataset, as anticipated, since it represented the linguistically and acoustically simplest condition.

All fine-tuned Whisper models improved transcription performance across datasets. Models trained on healthy and patient data further improved the WER compared to those trained on synthetic data, suggesting the importance of natural speech diversity in improving transcription performance.

The fine-tuned model showed adequate performance for post-stroke speech transcription. In patients, the best performing model was the Medium *ft-hp* model achieving a 26.35% WER. However, this result was not significantly different from Medium *ft-p* and *ft-h* models (26.80% WER ; $p > 0.05$ and 27.72% WER ; $p > 0.05$, respectively). Similarly, target detection performance of Medium *ft-hp*, *ft-p* and *ft-h* models was also comparable (0.92, 0.91 and 0.91 accuracy, respectively).

This suggests that fine-tuning on age-matched, task-specific healthy speech may generalize sufficiently to patient data for the same task. This is particularly relevant in clinical applications since collection of healthy speech data is more accessible, and can facilitate a larger dataset for fine-tuning.

Further, comparable performance between *ft-h*, *ft-p* and *ft-hp* models on patient speech suggests that the main performance advantage comes from fine-tuning Whisper on short-form speech produced in the single-word Naming task, irrespective of the data source.

The WER achieved by the *ft-p* model was consistent with previous finding on wav2vec 2.0 and Whisper models fine-tuned on aphasic speech (WER 22.30%–55.50% depending on aphasia severity and 21.93% WER, respectively) [20, 29], and outperformed hyperparameter-tuned Whisper (WER 38.50%) [26]. However, as prior work targeted continuous speech, direct comparison is limited.

Medium models achieved the best performance across most datasets but did not differ significantly in WER from the smaller variants. Given the higher computational cost and training time, Small models may offer a more efficient option for future clinical use, particularly when scaling to larger datasets.

Similar to the SONIVA-Naming test set, fine-tuned Whisper models showed markedly improved transcription on the unseen TORGO dataset compared to the baseline, underscoring the need to adapt pretrained models for out-of-domain clinical data. Nonetheless, the best-performing model (Medium *ft-syn*) still produced a high WER of 73.49%, well above clinical usability. Models fine-tuned on SONIVA patient data also performed poorly on TORGO (Medium WER: 74.62%). This can be attributed to clinical heterogeneity: SONIVA targets stroke-induced aphasia involving both motoric and linguistic processing difficulties, whereas TORGO focuses on dysarthria and motoric impairments only. These results emphasize the importance of tailoring pretrained models to the specific linguistic and pathological features of each condition to achieve robust generalization and clinical utility.

In terms of target detection accuracy on the SONIVA-Naming task data, fine-tuned models showed higher True Positive and lower False Negative scores compared to baseline Whisper, improving clinical applicability. However, on patient data, the fine-tuned models produced more False Positive mistakes, as a result of failing to transcribe phonological errors made by patients (e.g., predicting target word ‘comb’ instead of ‘comg’). As Whisper is designed to predict the next most probable token in the sequence, these discrepancies are expected.

Whisper was also used to complete a stroke-specific downstream task, aimed at estimating the level of speech impairment. In case of healthy data, the models initialized with *ft-h* encoder weights had

the best F1 Macro in accuracy score prediction, suggesting the importance of fine-tuning. However, for patient data, models initialized with *ft-p* encoder weights improved performance on the Semantic and Phonology metrics but not on Dysfluency and Self-correction. Specifically, while the fine-tuned model outperformed previous studies that used Whisper encoder features for classifying dysfluencies in stuttering (F1 Macro of 0.71) [23], its performance was slightly lower than the baseline for Dysfluency. This reduction in performance suggests that while fine-tuning for verbatim transcription captures linguistic features like semantics and phonology, it may not capture more complex patterns in patient speech, such as hesitations and self-corrections, which are less vital for transcription but crucial for Dysfluency and Self-correction predictions, leading to worse performance on these metrics for patient speech. It is, however, worth noting that the absolute difference between the best baseline encoder models and their fine-tuned encoder versions was negligible (1.45±0.06%).

In addition, the performance in the accuracy prediction task was higher on trials with unimpaired scores compared to impaired ones, especially in the healthy dataset. This is expected since in the case of healthy data, the models were exclusively fine-tuned on healthy speech where the number of impaired trials is inevitably limited, leading to smaller sample sizes and overall performance.

Despite this difference between model performance on impaired and unimpaired trials, the patient - level clinical validity analysis confirmed that the predicted accuracy scores were clinically meaningful. A significant difference between patients identified as impaired or unimpaired based on model predictions was observed in speech fluency, English proficiency status, and previous stroke history, as expected [37–39]. The results of the divergent validity analysis showed no differences in features unrelated to speech impairment such as cholesterol levels, sex and smoking status.

6 Limitations and Future Work

We note several limitations and directions for future work. First, patients were not stratified by speech impairment severity when splitting into training, validation, and test sets, which may have led to uneven distributions and limited generalizability across impairment levels. Nonetheless, this should not affect the main conclusions about Whisper, as similar findings were observed in healthy data.

In the accuracy prediction task, performance could be improved by addressing class imbalance through multiple binary classification tasks (instead of a multi-class one) and by using approaches alternative to downsampling. For example, data perturbation, modifying audio features such as pitch, frequency, or tempo, to create more samples has been used in disordered and aphasic speech transcription [42, 43, 41]. Alternatively, in-domain data augmentation, such as incorporating similar tasks from datasets like AphasiaBank [11], could also be applied [40].

Whisper's transcription accuracy could also be evaluated using beam search instead of greedy decoding, to create a more complete assessment of its applicability to stroke-specific data. Greedy decoding selects the most probable token at each step in the output sequence, whereas beam search considers multiple candidate sequences and chooses the final output based on cumulative probability [8, 44].

In order to obtain a broader overview on the application of foundation models to pathological speech in stroke, Whisper should be compared with other transformer-based architectures, such as Seamless M4T v2, a multimodal model capable of speech transcription that has shown competitive results compared to Whisper on ASR tasks [45]. Finally, future work should explore whether token-level loss analyses, encoder probing, or modality-specific pretraining can help disentangle the extent to which current models capture generalizable acoustic markers of clinical speech, particularly in settings where contextual compensation from the decoder is limited, such as single-word utterances.

7 Conclusion

We evaluated the applicability of Whisper, a general foundation model, to a specific clinical case study on a single-word speech task in patients with stroke. Our findings confirmed that additional fine-tuning on single-word data is required to achieve efficient transcription performance that can facilitate clinically useful downstream tasks. Further, comparisons of fine-tuned model performance on in-domain and out-of-domain neurological disorder datasets suggests further need for model adaptation to disorder-specific data. With such fine-tuning we can advance the automated analysis of pathological speech allowing for more efficient diagnosis and monitoring of stroke.

References

[1] V. L. Feigin, M. Brainin, B. Norrving, S. Martins, R. L. Sacco, W. Hacke, M. Fisher, J. Pandian, and P. Lindsay, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, Jan. 2022. doi: 10.1177/17474930211065917.

[2] H. L. Flowers, S. A. Skoretz, F. L. Silver, E. Rochon, J. Fang, C. Flamand-Roze, and R. Martino, "Poststroke Aphasia Frequency, Recovery, and Outcomes: A Systematic Review and Meta-Analysis," *Archives of Physical Medicine and Rehabilitation*, vol. 97, no. 12, pp. 2188–2201.e8, 2016. doi: 10.1016/j.apmr.2016.03.006.

[3] J. Hartman, "Measurement of early spontaneous recovery from aphasia with stroke," *Annals of Neurology*, vol. 9, no. 1, pp. 89–91, 1981. doi: 10.1002/ana.410090119.

[4] D. T. Wade, R. L. Hewer, R. M. David, and P. M. Enderby, "Aphasia after stroke: natural history and associated deficits," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 49, no. 1, pp. 11–16, 1986. doi: 10.1136/jnnp.49.1.11.

[5] D. Gruia, W. Trender, P. Hellyer, S. Banerjee, J. Kwan, H. Zetterberg, A. Hampshire, and F. Geranmayeh, "IC3 protocol: a longitudinal observational study of cognition after stroke using novel digital health technology," *BMJ Open*, vol. 13, p. e076653, Nov. 2023. doi: 10.1136/bmjopen-2023-076653.

[6] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution," arXiv preprint arXiv:2202.10054, 2022. [Online]. Available: https://arxiv.org/abs/2202.10054.

[7] K. W. Church, Z. Chen, and Y. Ma, "Emerging trends: A gentle introduction to fine-tuning," *Natural Language Engineering*, vol. 27, no. 6, pp. 763–778, 2021. doi: 10.1017/S1351324921000322.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *Proc. of the 40th International Conference on Machine Learning (ICML)*, vol. 202, pp. 28492–28518, 2023. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html.

[9] Y. Qin, T. Lee, and A. P. H. Kong, "Automatic assessment of speech impairment in Cantonese-speaking people with aphasia," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 331–345, 2020. doi: 10.1109/JSTSP.2019.2956371.

[10] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Communication*, vol. 100, Apr. 2018. doi: 10.1016/j.specom.2018.04.001.

[11] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for studying discourse," *Aphasiology*, vol. 25, pp. 1286–1307, Nov. 2011. doi: 10.1080/02687038.2011.589893.

[12] S. Wang and G. Li, "Overview of end-to-end speech recognition," *Journal of Physics: Conference Series*, vol. 1187, p. 052068, Apr. 2019. doi: 10.1088/1742-6596/1187/5/052068.

[13] X. Bai, Y. Li, Z. Zhang, M. Xu, B. Chen, W. Luo, D. Wong, and Y. Zhang, "Sentence-state LSTMs for sequence-to-sequence learning," in *Proc. NLPCC 2021: Natural Language Processing and Chinese Computing*, Qingdao, China, Oct. 2021, pp. 104–115. doi: 10.1007/978-3-030-88480-2_9.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[15] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Inf. Fusion*, vol. 99, no. C, pp. 1–55, Nov. 2023. doi: 10.1016/j.inffus.2023.101869.

[16] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss," in *Proc. IEEE ICASSP*, 2020, pp. 7829–7833. doi: 10.1109/ICASSP40776.2020.9053896.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, pp. 12449–12460.

[18] W. Chan, D. S. Park, C. A. Lee, Y. Zhang, Q. V. Le, and M. Norouzi, "SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network," *CoRR*, vol. abs/2104.02133, 2021. [Online]. Available: https://arxiv.org/abs/2104.02133

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," *arXiv preprint*, arXiv:2005.08100, 2020. [Online]. Available: https://arxiv.org/abs/2005.08100

[20] I. G. Torre, M. Romero, and A. Álvarez, "Improving aphasic speech recognition by using novel semi-supervised learning methods on AphasiaBank for English and Spanish," *Appl. Sci.*, vol. 11, no. 19, p. 8872, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/19/8872

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *arXiv preprint arXiv:2006.13979*, Jun. 2020. doi: 10.48550/arXiv.2006.13979.

[22] D. R. Yerramreddy, J. Marasani, P. S. V. Gowtham, H. Guduru, and Anjali, "Speech Recognition Paradigms: A Comparative Evaluation of SpeechBrain, Whisper and Wav2Vec2 Models," in *Proc. of the 2024 IEEE 9th Int. Conf. for Convergence in Technology (I2CT)*, 2024, pp. 1–6. doi: 10.1109/I2CT61223.2024.10544133.

[23] H. Ameer, S. Latif, and R. Latif, "Optimizing multi-stuttered speech classification: Leveraging Whisper's encoder for efficient parameter reduction in automated assessment," unpublished, 2024.

[24] S. Rathod, M. Charola, A. Vora, Y. Jogi, and H. A. Patil, "Whisper features for dysarthric severity-level classification," in *Proc. Interspeech 2023*, 2023, pp. 1523–1527. doi: 10.21437/Interspeech.2023-1891.

[25] L. Wagner, M. Zusag, and T. Bloder, "Careful Whisper—leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification," *arXiv preprint*, Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.01327

[26] G. Sanguedolce, P. A. Naylor, and F. Geranmayeh, "Uncovering the potential for a weakly supervised end-to-end model in recognising speech from patient with post-stroke aphasia," in *Proc. 5th Clin. Natural Lang. Process. Workshop*, Toronto, Canada, Jul. 2023, pp. 182–190. doi: 10.18653/v1/2023.clinicalnlp-1.24.

[27] G. Sanguedolce, D.-C. Gruia, S. Brook, P. Naylor, and F. Geranmayeh, "Universal speech disorder recognition: Towards a foundation model for cross-pathology generalisation," in *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024.

[28] C. Roth, "Boston Diagnostic Aphasia Examination," in *Encyclopedia of Clinical Neuropsychology*, J. S. Kreutzer, J. DeLuca, and B. Caplan, Eds. New York, NY: Springer, 2011, pp. 428–430. doi: 10.1007/978-0-387-79948-3_868.

[29] G. Sanguedolce, C. J. Price, S. Brook, D. C. Gruia, N. V. Parkinson, P. A. Naylor, and F. Geranmayeh, "SONIVA: Speech recognition validation in aphasia," in *medRxiv preprint*, Jun. 2025. doi: 10.1101/2025.06.

[30] P. N. Durette and Contributors, "gTTS: Google Text-to-Speech," 2014–2024, version 2.5.4, MIT License. [Online]. Available: https://gtts.readthedocs.io/

[31] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 523–541, 2012. doi: 10.1007/s10579-011-9187-7.

[32] Hugging Face, "WhisperForAudioClassification — Transformers Documentation," Available: `https://huggingface.co/docs/transformers/model_doc/whisper#transformers.WhisperForAudioClassification`.

[33] OpenAI, "Whisper Small," *Hugging Face*, updated Feb. 29, 2024. [Online]. Available: https://huggingface.co/openai/whisper-small

[34] OpenAI, "Whisper Medium," *Hugging Face*, updated Feb. 29, 2024. [Online]. Available: https://huggingface.co/openai/whisper-medium

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[36] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Aug. 2016, pp. 1715–1725. doi: 10.18653/v1/P16-1162.

[37] A. R. Fonseca, J. and I. P. Martins, "Cognitive functioning in chronic post-stroke aphasia," *Applied Neuropsychology: Adult*, vol. 26, no. 4, pp. 355–364, 2019, doi: 10.1080/23279095.2018.1429442.

[38] T. M. H. Hope, 'Ō. Parker Jones, A. Grogan, J. Crinion, J. Rae, L. Ruffle, A. P. Leff, M. L. Seghier, C. J. Price, and D. W. Green, "Comparing language outcomes in monolingual and bilingual stroke patients," *Brain*, vol. 138, no. 4, pp. 1070–1083, 2015, doi: 10.1093/brain/awv020.

[39] M. S. Sharif, E. B. Goldberg, A. Walker, A. E. Hillis, and E. L. Meier, "The contribution of white matter pathology, hypoperfusion, lesion load, and stroke recurrence to language deficits following acute subcortical left hemisphere stroke," *PLOS ONE*, vol. 17, no. 10, pp. 1–25, 2022, doi: 10.1371/journal.pone.0275664.

[40] R. C. Gale, M. Fleegle, G. Fergadiotis, and S. Bedrick, "The Post-Stroke Speech Transcription (PSST) Challenge," in *Proc. of the RaPID Workshop at the 13th Language Resources and Evaluation Conference*, Marseille, France, Jun. 2022, pp. 41–55. [Online]. Available: https://aclanthology.org/2022.rapid-1.6

[41] J. Yuan, X. Cai, and K. Church, "Data Augmentation for the Post-Stroke Speech Transcription (PSST) Challenge: Sometimes Less Is More," in *Proc. of the RaPID Workshop at the 13th Language Resources and Evaluation Conference*, Marseille, France, Jun. 2022, pp. 71–79. [Online]. Available: https://aclanthology.org/2022.rapid-1.9

[42] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, "Investigation of Data Augmentation Techniques for Disordered Speech Recognition," in *Interspeech 2020*, Oct. 2020. doi: 10.21437/Interspeech.2020-1161.

[43] B. Moëll, J. O'Regan, S. Mehta, A. Kirkland, H. Lameris, J. Gustafsson, and J. Beskow, "Speech data augmentation for improving phoneme transcriptions of aphasic speech using wav2vec 2.0 for the PSST challenge," in *Proc. of the 4th RaPID Workshop*, Marseille, France, Jun. 2022, pp. 62–70.

[44] Y. Chen, V. O. K. Li, K. Cho, and S. R. Bowman, "A Stable and Effective Learning Strategy for Trainable Greedy Decoding," arXiv preprint arXiv:1804.07915, 2018. [Online]. Available: https://arxiv.org/abs/1804.07915

[45] Seamless Communication et al., "Seamless: Multilingual Expressive and Streaming Speech Translation," arXiv preprint arXiv:2312.05187, 2023. [Online]. Available: https://arxiv.org/abs/2312.05187

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[47] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "HuggingFace's Transformers: State-of-the-Art Natural Language Processing," arXiv preprint arXiv:1910.03771, 2019. [Online]. Available: https://arxiv.org/abs/1910.03771

A Sample demographics of the SONIVA-Naming dataset

Table 4: Sample demographics of the SONIVA-Naming dataset.

Demographic Feature	SONIVA-Naming healthy	SONIVA-Naming patients
	N = 132	N = 87
	Mean (Standard Deviation)	
Age	61.6 (10.8)	61.8 (13.6)
Sex***		
Male:Female	52:77	62:25
Missing data	3	0
English Language*		
Non-native	27	30
Native	102	57
Education level***		
School	36	45
Degree	43	37
Post-graduate	37	3
Missing data	16	2

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

B Whisper Architecture

Whisper has an encoder-decoder Transformer architecture [8] with different model size availability, corresponding to the number of parameters (39M-1550M) and encoder-decoder block layers (4-32). Each encoder block of Whisper consists of a self-attention layer and a multilayer perceptron. The model processes audio input in a form of an 80-channel log-Mel spectrogram, passing it through two initial convolutional layers followed by the encoder blocks. The encoder outputs information-dense context vectors, which are a high-level feature representation of the input audio sequence, encompassing acoustic, positional, and contextual information through self-attention mechanisms and positional embedding.

The decoder blocks mirror the architecture of the encoder blocks, incorporating an additional cross-attention layer, which enables the decoder to focus on the outputs of the encoder. This architecture allows the decoder to process the context vectors produced by the encoder and the tokenized input text sequences. Whisper utilizes a Byte-Pair Encoding tokenizer to break down the input text into smaller sub-word units [36]. During training, tokenized text sequences correspond to the audio input's ground truth labels, while during inference, they consist of previously predicted tokens. Additionally, Whisper utilizes a range of special tokens, which signal the start and end of the sequence, task type and language. The decoder's raw output is passed through a final softmax layer to obtain the next token prediction.

To complete the accuracy classification task, WhisperforAudioClassification model configuration [32] was used, which retains the same Whisper encoder architecture, but replaces the decoder with a linear classifier projection head, consisting of two linear layers.

C The general study pipeline

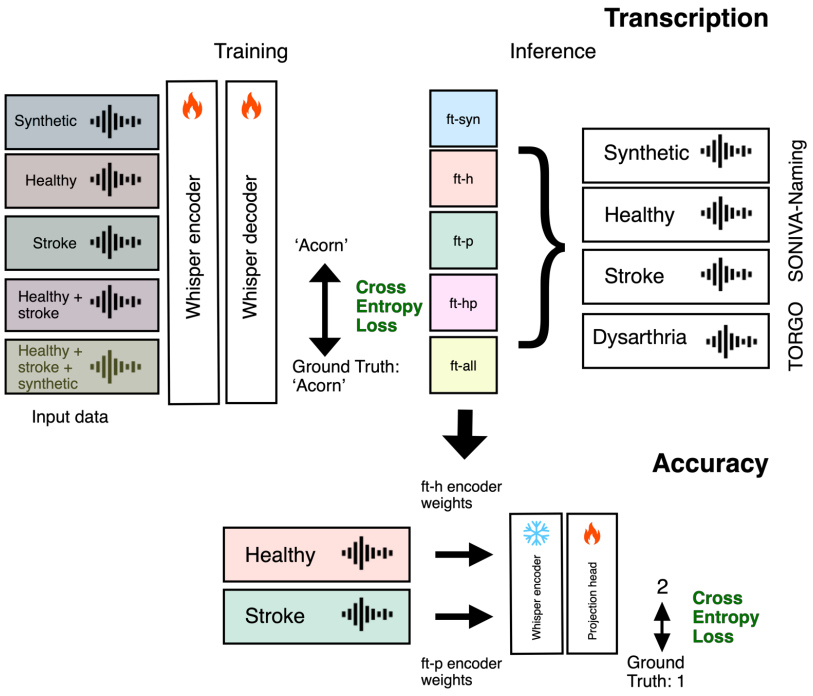

Figure 1: General study pipeline. Overview of the training and inference pipeline for verbatim transcription and accuracy prediction models. Transcription: Training – Whisper models were fine-tuned on speech data from 5 different datasets. A synthetic dataset of audio-transcription word pairs based on stimuli in the Naming task, SONIVA-Naming healthy dataset, collected from healthy age-matched participants performing the Naming task, SONIVA-Naming patient dataset, collected from patients with stroke performing the Naming task, a combined dataset of SONIVA-Naming healthy and patient data, and a combined dataset of all available data (synthetic, SONIVA-Naming healthy and SONIVA-Naming patient). All models were trained using cross-entropy loss to transcribe spoken words (e.g. "acorn") from the Naming task input audio. Inference – Fine-tuned models (*ft-syn*, *ft-h*, *ft-p*, *ft-hp* and *ft-all*) were evaluated on 4 separate datasets – synthetic, SONIVA-Naming healthy, SONIVA-Naming patient, and an unseen testing dataset of dysarthric speech derived from the TORGO database. Accuracy: Training – accuracy prediction models were trained on SONIVA-Naming derived healthy and patient speech in a linear probing framework. The encoder layer of the models was frozen and the classification projection head was trained to predict accuracy scores using cross-entropy loss (e.g "2"). The encoder weights of the models were initialized at baseline or with the weights derived from the models trained for verbatim transcription of healthy or patient speech

D Additional training details

All model training was conducted on a single NVIDIA RTX 6000 GPU. Mixed-precision arithmetic (fp16) and gradient checkpointing was implemented to improve computational efficiency and optimise memory usage. Model training and evaluation were conducted using the PyTorch framework [46] and HuggingFace Transformers library [47].

E Predictive Validity Analysis

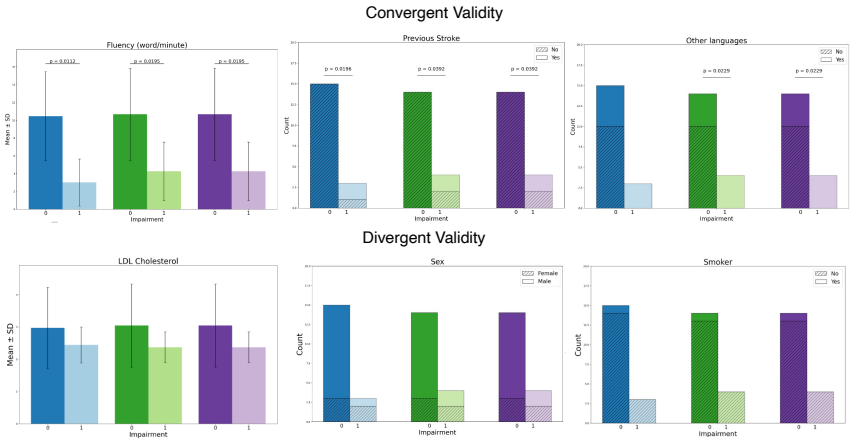

Figure 2: Characteristics of patients identified as impaired (0) or unimpaired (1) from the SONIVA-Naming patient dataset. Predicted impairment status is displayed separately for Semantic (blue), Dysfluency (green), and Phonology (purple) metrics. Speech fluency, previous stroke history, English as second language (other languages), LDL cholesterol level, sex and smoking status were assessed. Error bars indicate standard deviation. Between group significance is denoted with p-values (uncorrected).