Universal Speech Disorder Recognition: Towards a Foundation Model for Cross-Pathology Generalisation

Giulia Sanguedolce[1,2,3], **Dragos C. Gruia**[3,4], **Sophie Brook**[3,4],
Patrick A. Naylor[2], **Fatemeh Geranmayeh**[3,4]
[1]Department of Computing, Imperial College London, UK
[2]Department of Electrical and Electronic Engineering, Imperial College London, UK
[3]Department of Brain Sciences, Imperial College London, UK
[4]Imperial College Healthcare NHS Trust, London, UK
{g.sanguedolce22, sophie-mei.brook22, dragos-cristian.gruia19
p.naylor, fatemeh.geranmayeh00}@imperial.ac.uk

Abstract

Although Automatic Speech Recognition (ASR) systems hold great potential for diagnosing and monitoring speech disorders, their clinical utility has been limited due to the scarcity of pathological speech data capable of capturing the heterogeneity of speech disorders and the high-dimensional output space. This work presents a significant advancement by fine-tuning the Whisper foundation model on our novel in-house SONIVA dataset, which consists of ≈600 stroke patients with diverse speech impairments. We address the critical challenge of limited training data in healthcare ASR demonstrating that our stroke-specific model generalises effectively across diverse speech pathologies. Our model outperforms both the off-the-shelf Whisper and traditional disorder-specific ASR models, demonstrating improved recognition accuracy on SONIVA unseen patients, as well as AphasiaBank and DementiaBank. This extends beyond stroke-related language impairments, improving performance also in other neurological disorders. This cross-pathology approach was achieved through training on a single disorder with a heterogeneous impairment profile, representing a major step towards more adaptable disordered speech recognition in healthcare. Our study highlights how adapting foundation models for clinical tasks can advance universal speech disorder recognition despite data scarcity, with broad implications for diagnosis, monitoring, and patient care, potentially enabling more accessible and personalised speech therapy.

1 Introduction

Recent advancements in large-scale pre-trained models have hastened significant progress in healthcare research, transcending traditional machine learning approaches through their transfer learning ability and multi-task generalisation. Within the field of neurology, by integrating diverse inputs such as acoustic signals, linguistic structures, and clinical metadata, these models can infer unique and scalable representations of patient health directly from speech. Given the heterogeneity of speech errors, Medical Foundation Models (MFM), reaching remarkable in various healthcare domain, represent an optimal opportunity for capturing the intricate temporal and spectral features of pathological speech while concurrently modeling higher-level linguistic and cognitive impairments. However, effective application of MFM in clinical speech settings requires fine-tuning on specific datasets, as the inherent variability and complexity of speech disorders with a given neurological damage can limit performance without tailored adjustments. Thus, while MFM demonstrate significant potential, targeted fine-tuning is essential to fully leverage their capabilities in addressing the unique challenges underlying neurological conditions of different nature.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

SONIVA, our in-house built database, addresses the critical gap in available datasets for fine-tuning foundation models like Whisper. Post-stroke speech databases present an optimal condition for this purpose due to the significant speech impairment heterogeneity that results from a wide range of stroke lesion topologies. This diversity leads to a range of speech symptoms that frequently overlap with those found in other neurological disorders, such as Parkinson's disease (e.g. reduced phonation), cerebral palsy, multiple sclerosis, amyotrophic lateral sclerosis (dysarthric speech), or dementia (e.g. reduced speech fluency and lexical complexity). The primary objective of this paper is to investigate the efficacy of fine-tuning Whisper for post-stroke speech recognition. The second aim is to evaluate its generalisability using three distinct datasets: AphasiaBank [1], TORGO [2] and DementiaBank [3]. AphasiaBank is an international post-stroke aphasia speech bank, which introduces variability in terms of dialects and test protocols. TORGO and DementiaBank expand the range of speech disorders in our study: TORGO represents speech impairments caused by cerebral palsy and amyotrophic lateral sclerosis (ALS), while DementiaBank captures the speech changes associated with cognitive decline in dementia.

For this task Whisper model was selected due to its demonstrated ability to achieve near-human-level accuracy on healthy speech even in low-resource data scenarios, as shown in [4] when compared with the latest ASR models. Additionally, Whisper can be deployed locally, allowing all processing to occur on secure, on-site systems. This minimizes the need to transmit patient sensitive data over networks or to cloud services, ensuring compliance with strict privacy regulations. Local deployment also enables real-time processing without relying on internet connectivity, which is crucial in clinical settings where consistent access to high-speed internet may not be available.

After fine-tuning the model we demonstrate that the additional training not only enhances its performance for stroke-related speech impairments, but also enables the model to leverage this variability to generalize to speech disorders in pathologies beyond stroke. This showcases the potential clinical applicability of our fine-tuned model, as it holds implications for improving diagnostic processes, therapy planning, and overall patient care across a spectrum of neurological conditions. Such work represents a pioneering step towards developing a universal speech recognition system capable of handling diverse communication disorders.

2 Background

2.1 Speech and neurological disorders

Stroke can lead to a wide range of communication disorders, affecting various aspects of speech and language processing depending on the location and extent of brain damage. These impairments can manifest in varying degrees of severity, affecting different aspects of the communication process and often coexisting in combination [5]. Aphasia, a common consequence of stroke, disrupts language processing, causing difficulties in speech production, comprehension, reading, or writing. Patients with fluent aphasia produce connected speech that sounds natural in rhythm and intonation, but may experience word-finding problems affecting mostly nouns and picturable action words [6]. They often exhibit phonemic (e.g., "cat" to "pat"), semantic (e.g., "fork" to "knife"), or verbal paraphasias (e.g., "apple" to "gorilla"). Non-fluent aphasia, on the other hand, is characterised by slow and hesitant elocution with episodes of agrammatism, such as the absence or improper use of function words and verbs [7]. The speech is typically fragmented, choppy, and awkwardly articulated [6], often containing neologisms or jargon.

Cognitive-communication disorders may also arise, affecting higher-level language functions and impacting a patient's ability to effectively convey ideas or engage in social interactions. Apraxia of speech, another potential sequela of stroke and some neurodegenerative dementias, affects the motor planning necessary for speech, resulting in inconsistent articulation errors and challenges in sequencing speech sounds correctly. Voice quality can be significantly altered by dysphonia, characterised by changes in pitch, loudness, or overall voice quality due to impairment of the laryngeal muscles [8]. Dysarthria, another motor speech disorder that targets the articulatory muscle control, often accompanies these impairments following stroke, leading to slurred speech, altered rhythm, and reduced intelligibility due to muscular weakness or incoordination [9]. These various disorders frequently coexist creating unique speech patterns for each individual. Furthermore, there is heterogeneity of these impairments between patients, and within the same patient over time [10].

This broad spectrum of post-stroke impairments is key to the creation of a universal MFM. Stroke episodes stems from the variable locations where blood clots can form in the brain, potentially affecting multiple regions responsible for different aspects of speech production and language processing. In contrast, other disorders unlike stroke commonly arise from more localized or focal damage, resulting in a more homogeneous presentation of language deficits.

2.2 Automation of personalised assessment

Given the heterogeneity of language disfluencies, a one-size-fits-all approach to speech assessment and therapy is inadequate. Each patient presents a unique combination of impairments, necessitating highly individualized treatment plans. Traditionally, comprehensive speech and language assessments require frequent face-to-face sessions with trained speech-language pathologists [11]. However, accessing this level of tailored diagnosis and ongoing therapy can be costly for both patients and healthcare providers, further compounded by the logistical challenges posed by additional impairments resulting from stroke [12; 13; 14]. This situation underscores the need for more flexible, scalable tools that can accommodate the wide range of speech patterns observed in stroke survivors. Automatic Speech Recognition (ASR) systems could potentially address these challenges, streamlining the assessment process and allowing more frequent and cost-effective evaluations, potentially remote. Providing a detailed, objective measurement of speech characteristics, these models could potentially support clinical decision-making for speech monitoring.

Despite the significant advancements in speech recognition technology for typical speech, the accuracy achieved for pathological speech remains substantially lower. This discrepancy is primarily attributed to the lack of big datasets of pathological speech impairments that contain the diverse range of potential speech errors. Recent developments in foundation models have shown promising results in addressing this challenge. These large-scale models, trained on vast amounts of diverse data, already demonstrated the ability to generalise across different domains, outperforming traditional machine learning models that were specifically trained on pathological speech [15].

2.3 Medical Foundation Models for language and speech

The advent of foundation models in healthcare has significantly impacted language research, especially with BioBERT [16] and ClinicalBERT [17], released in early 2019 following BERT [18], that marked a significant milestone in healthcare language modelling. The launch of ChatGPT [19] in late 2022 further accelerated this trend, broadening the accessibility and application of foundation models in healthcare settings [20]. The advancements of Large Language Models (LLMs) have surely played a crucial role in enhancing foundation models, particularly in their ability to understand and generate context-rich, domain-specific language, which is essential for accurate language analysis in healthcare settings. Looking at the most recent work, multimodal LLMs, such as GPT-4V [21], offered significant potential in augmenting clinical decision-making and management, though challenges in regulation and validation remain [22]. Me-LLaMA [23], a novel medical LLM family, also demonstrated superior performance across general and medical tasks compared to existing open-source medical LLMs.

In the domain of MFM applied to speech processing, the authors of [24] investigated the impact of layer selection in speech foundation models, specifically wav2vec 2.0 [25], on the prediction of pathological speech features relevant to neurological conditions. The authors found that selecting optimal layers can significantly enhance performance, even though the generalisation to unseen data in their work remained a challenge. An interesting approach was shown in [26], where the authors developed a Perceiver-based sequence classifier for detecting speech abnormalities associated with neurological disorders. Their model leveraged a Universal Speech Model (USM), a type of foundation model pre-trained on an extensive dataset of 12 million hours of audio. The model was evaluated on three critical clinical tasks: vowel prolongation, alternating motion rate, and sequential motion rate, aiming to predict 14 distinct speech attributes. This approach exemplifies the potential of using large-scale pre-training in the speech domain, similar to the successes seen with LLMs in text-based tasks.

Building upon these advancements in speech foundation models, recent research has explored the specific application of Whisper to the domain of speech disorders [27; 28; 29; 15; 30; 31; 32; 33], given its robustness also on low resources databases[34]. Sanguedolce et al. (2023) [15] used

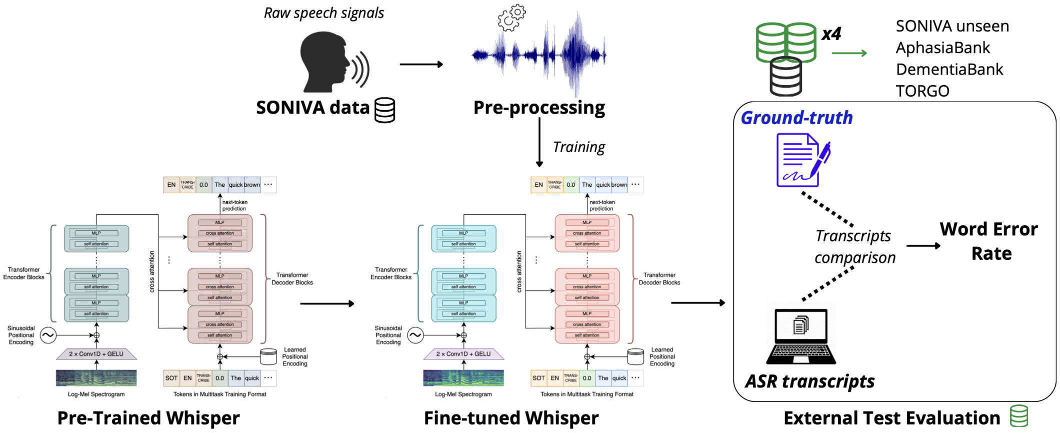

Figure 1: Pipeline of the main experiment on Whisper [4] fine-tuned on SONIVA data only, followed by evaluation on multiple external datasets to assess generalisation.

Whisper on post-stroke speech, reaching a Word Error Rate (WER) of 38.5% with the off-the-shelf model, in line with those achieved by traditional ASR trained specifically on pathological speech. In work [35], the authors leveraged Whisper with AphasiaBank database to differentiate various types of aphasias and distinguish them from healthy controls. They extracted linguistic features such asfluency, coherence, lexical richness, syntax, and pronunciation, reaching an F1 score of 90.6%. Using the Cantonese version of AphasiaBank, earlier studies with traditional ASR have achieved notable advancements [36]. Using Multilayer Time Delay Neural Networks (MT-DNN) with a Bidirectional Long Short-Term Memory (BLSTM) structure, the authors achieved a WER of 18.5% for unimpaired speech and 42.4% for impaired speech. More recent studies have shown significant improvements using attention mechanisms like E-Branchformer and Conformers, reaching an average WER of 26% across various aphasia severity levels [37] when trained on Aphasia Bank. Dysarthria has been also extensively researched, particularly due to the availability of specialised databases such as TORGO [2] or UA-Speech [38]. Nevertheless, such databases contain a limited number of speakers ($N = 8$ and $N = 15$ respectively), hindering the development of robust models. For this reason, [39] implemented a two-stage training data augmentation scheme. This approach significantly improved ASR performance on dysarthric speech, achieving a notable 20.6% WER. Similarly, as a data augmentation approach, the study from [27] demonstrated that using diffusion-based text-to-speech models to generate synthetic speech samples that mimic dysarthric speech characteristics can improve fine-tuned ASR performance, particularly on the Whisper model.

Overall, these studies highlight the need to develop models and techniques that maximise the utility of existing datasets, addressing the challenge of limited availability of pathological speech data. Even though there has been a progression from traditional ASR systems to more advanced models such as Whisper, critical technical challenges still persist in handling the domain shift between typical and pathological speech, especially when comparing the average WER reached on healthy speech (10-13% [4]). To address these issues, our study introduces a fine-tuned Whisper model approach for pathological speech recognition. By leveraging transfer learning on diverse speech disorder datasets, we aim to improve ASR performance across various types of speech impairments, potentially bridging the gap between current technological capabilities and clinical needs in automated universal speech analysis.

3 Methods

3.1 SONIVA database

The primary database used in this study is SONIVA, an in-house comprehensive corpus of post-stroke speech built for clinical and scientific research into training automatic speech recognition systems in the context of disordered speech. Such corpus aims to be a substantial and valuable resource for the wider research community. It includes speech recordings from ≈1000 people with stroke who participated in two longitudinal studies: IC3[1](Imperial Comprehensive Cognitive Assessment in Cerebrovascular Disease [40; 41]), and PLORAS[2] (Predicting Language Outcome and Recovery After Stroke [42]). All participants were consented as per appropriate study ethics approval prior to testing. The speech recordings contain pictures description tasks from the Comprehensive Aphasia Test (CAT [43]) and a beach scene picture stimuli. The dataset is further complemented by detailed orthographic English transcriptions by trained speech pathologists, as well as phonetic transcription using the International Phonetic Alphabet. The transcriptions further undergo quantitative linguistic assessments using CLAN [44]. For this study, the labeled dataset used consists of 794 audio recordings from 578 unique Patients with Stroke (PwS). Some of these individuals underwent repeated testing to capture longitudinal recovery and individual variation in speech patterns. Overall, 15 hours of speech data were included in this study.

Although all the speech tasks were performed in British English, SONIVA also includes some patients with accents from a wide geographical distribution (Figure 3). The majority of individuals in the database have UK-based (76.77%), while the remaining speech samples had European (2.99%), South Asian (3.67%), African (2.85%), Central American (2.58%), Oceanian (0.68%) and North American (0.27%) origin. The accents were determined by considering both the speakers nationalities and the acoustic-phonetic analysis conducted by trained speech therapists. The remaining 10.19% of the speakers had an unknown or mixed accent, not categorised in the above classes.

The distributions of age at test for the different genders are displayed in Appendix A (Figure 4). The speakers were predominantly male (70.79%), with an average age of 61.96 years at testing. In contrast, the average age for female speakers was 58.52 years. This gender unbalance reflects the world-wide higher incidence of stroke among males [45]. Additionally, the age distributions are both left-skewed as expected given the age distribution of stroke cohorts.

3.2 Data pre-processing

A team of two speech therapists and three trained postgraduate students transcribed the audio recordings verbatim, achieving a high level of consistency (73% inter-rater reliability). The transcriptions adhered to the formatting guidelines of the Codes for the Human Analysis of Transcripts (CHAT) [46], and were subsequently processed using the Computerized Language Analysis (CLAN) [44] software. A pre-processing step was applied to remove the special symbols used for linguistic error coding, such as semantic, phonological and dysfluencies-related errors. The transcribers also provided phonetic alphabet representations in case of neologisms or vocalisations, which were later heuristically converted to a sequence of Latin alphabet phonemes without altering the original order [47]. The human transcriptions included codes for false starts and unique symbols for filler words e.g. “er", “erm", and other isolated sounds or interjections typical of dysfluent speech. Previous research found spelling variations in filler words between American and British English, leading to higher WER due to differences in written linguistic conventions. Specifically, British usage includes “er" and “erm", whereas American usage includes “uh" and “um". To address this, we normalised these filler words according to the prevalent American English conventions in our Whisper training dataset.

The files were segmented by utterances, with the transcriptions and corresponding audio files aligned using the manually inserted timestamps by the expert transcriber, as we did in [15; 30; 48]. This allowed the mapping of the start and end of each utterance to the associated audio segment. By segmenting the data into sentences instead of fixed-length chunks, we accounted for the varying utterance length, mitigating the risk of overfitting during the training process. Some audio files included the assessor’s speech, which was also transcribed. For training purposes, these were excluded to allow the model to solely learn patients’ speech only. Whisper is unable to process long-form audio files exceeding 30 seconds. To address this limitation, such files were segmented further together with their aligned transcriptions. Similarly, audio files shorter than 3 seconds were prone to potential issues in the computation of Fourier transform for spectrograms generation [49]. Rather than discarding these shorter files, they were merged together with the aligned transcripts when they belonged to the same speaker, in order to avoid wasting resources during the training process. The files were subsequently converted into *.wav* format, resampled to 16 kHz with a 16-bit

[1] https://www.ic3study.co.uk; study approved by South West - Frenchay Research Ethics Committee [ref. IRAS 299333], and authorized by the UK’s Health Research Authority (HRA).

[2] https://ploras.ucl.ac.uk; study approved by the London Queen Square Research Ethics Committee [ref. IRAS 133939], and authorized by the UK’s Health Research Authority (HRA).

resolution and downmixed from stereo to mono. At the end of the pre-processing the final duration of the database was 13 hours. Alongside Speech Brain, these processing utilised Python packages Pydub [50], FFmpeg [51] and SoX [52].

3.3 AphasiaBank, TORGO and DementiaBank databases

To evaluate the generalisation capabilities of our fine-tuned model, we utilized three well-established databases in the field of disordered speech recognition. AphasiaBank [1] is a large, shared database of multimedia interactions for the study of communication in aphasia. It contains speech samples from 466 individuals with various types of aphasia, primarily resulting from stroke. The database includes a diverse range of structured speech tasks, including picture descriptions, personal narratives, and procedural discourse. It uses the CLAN and CHAT files structure as used in the SONIVA database.

The TORGO database [2], instead, focuses on dysarthric patients with cerebral palsy and amyotrophic lateral sclerosis. It contains aligned text and acoustic data from 8 speakers, with multiple sessions of testing. The speech samples mainly consist of elicited single-word prompts, with the recordings segmented into individual word utterances. DementiaBank [3] is a large corpus of speech samples from individuals with neurodegenerative cognitive disorders, ranging from mild neurocognitive disorders to Alzheimer's disease. It contains spontaneous speech recordings of subjects performing cognitive tasks, such as picture description tasks or recalling stories all made by long-form files, like SONIVA and AphasiaBank. This database provides insights into the language changes associated with cognitive decline, including difficulties with word-finding, stuttering, false start, decreased syntactic complexity and reduced informational content.

These databases were selected to complement our stroke dataset in the evaluation step. AphasiaBank, sharing the post-stroke etiology but differing in dialect and assessment protocols, tests Whisper's ability to learn general stroke-specific speech patterns beyond our dataset's features. TORGO, representing motor speech disorders from non-stroke neurological conditions, challenges the model with divergent speech patterns but also with single-word audio files, requiring Whisper to adapt its processing strategy between phonetic-focused and language model-based approaches. DementiaBank introduces speech affected by cognitive decline, presenting distinct challenges from aphasia and motor speech disorders. This selection allows us to assess the fine-tuned model's robustness across varied disordered speech, differing in etiology (stroke, neurodegenerative diseases, speech motor disorders), symptomatology (language processing deficits, motor speech impairments, cognitive decline), assessment contexts, and processing demands.

3.4 Model architecture

Whisper showcases the power of foundation models in speech recognition, leveraging a massive pre-training dataset and a versatile architecture. It is an encoder-decoder transformer model designed for multi-task speech processing. The encoder processes 80-channel log-Mel spectrograms using two convolutional layers and sinusoidal positional encoding, followed by a stack of transformer blocks for extracting long-range dependencies. The decoder mirrors this structure with its learned positional embeddings [4]. Whisper's key strength lies in its extensive pre-training on a diverse dataset of 680 000 hours of speech, encompassing various acoustic conditions, speakers, and languages. Such features align with the foundation model paradigm, where large-scale pre-training on diverse data enables robust generalisation. The model varies in size, with differences in parameter count and transformer layer depth. The encoder generates fixed-dimensional vectors that scale with model capacity, while maintaining a constant temporal dimension. Whisper employs a byte-level Byte-Pair Encoding (BPE) tokenizer, the same as GPT-2 [4; 53], supporting both monolingual and multilingual models. This tokenisation strategy allows efficient handling of out-of-vocabulary words and potential cross-lingual transfer. Whisper's near-human accuracy on healthy speech, achieved through weak supervision and diverse training data, establishes a solid foundation for fine-tuning on disordered speech tasks. This suggests strong potential for a MFM that can adapt to the atypical acoustic patterns and heterogeneity found in pathological speech.

3.5 Fine-tuning

SONIVA dataset was partitioned based on individual audio files with varying duration, resulting in a split of 70% for training (551 minutes), 18% for the validation (141 minutes) and 12% of unseen test set (94 minutes). Each division took into account a stratified splitting based on the severity of stroke cases, ensuring a balanced representation of speech severity. The unseen test set comprised audio recordings from patients who were entirely absent from the other splits. By maintaining this rigorous division, we established a truly independent evaluation framework, effectively minimising potential biases and enabling a robust assessment of the model's ability to generalise to unseen individuals. For the main experiment, only SONIVA database was used, as shown in the pipeline in Figure 1. However, after observing the results of this main experiment, we conducted an additional experiment using the TORGO dataset (see Table 1). For this follow-up study, we created an additional TORGO unseen test set by randomly selecting 2 out of 8 patients' speech data. The additional evaluation of external databases were also performed, following the SONIVA experiment.

For the fine-tuning, in any training experiment, the Whisper medium-sized model was selected, opting to update all trainable parameters without freezing any layers. This approach allowed for comprehensive adaptation to both deeper high-level language layers and the final feature-specific layers. A batch size of 16 for each device was adopted, with gradient accumulation to maximise GPU efficiency. The cosine learning rate schedule was implemented, initiating the learning rate at 1×10^{-5} and incorporating a warm-up period of 1000 steps. The optimisation process relied on the `AdamW` algorithm, which updated all model parameters based on the cross-entropy loss function. We conducted model evaluations at 1000-step intervals, with the primary objective of reducing the WER on our validation dataset. At each of these intervals, the model would be saved at the checkpoint, reaching the a maximum of 6000 steps. The best-performing model was retained, as determined by the lowest WER score. To optimise computational efficiency, a mixed-precision arithmetic (fp16) was implemented, with gradient check-pointing to effectively manage memory usage. The entire fine-tuning procedure, which lasted approximately 8 hours, was executed using the PyTorch framework [54] in conjunction with the Hugging Face Transformers library [55]. All computations were carried out on a single NVIDIA RTX 6000 GPU.

The WER was selected as evaluation metric, which is derived from the string edit distance. This measure quantifies the minimum number of modifications needed to transform Whisper's output into the corresponding human-generated transcription. To establish a performance benchmark, we first calculated the WER using the original, off-the-shelf Whisper model. We then compared this baseline performance against the WER achieved by our fine-tuned model. To determine the statistical significance of our results, we employed the non-parametric Wilcoxon rank-sum test. We chose this test due to its robustness when dealing with data that may not follow a normal distribution and the paired observations between *non-tuned* and *fine-tuned* distributions for all the databases.

4 Results

The summary of the WER average distribution is shown in Fig. 2. The differences between the *non-tuned* models and models *fine-tuned* on our SONIVA databasewere all found to be statistically significant ($p < 0.001$ for all comparisons). Compared to the other databases, the model tested on TORGO exhibited a much higher range of WER values for both models. In contrast, the DementiaBank and AphasiaBank datasets had WER values in a range comparable to the both the Validation and unseen SONIVA sets.

As shown in Table 1, fine-tuning on SONIVA significantly improved performance, achieving WERs of 21.51% and 21.93% on the validation and test sets respectively, compared to the baseline model's 39.60% and 43.62%. Subsequently, the model's robustness was first evaluated on AphasiaBank, where it achieved a WER of 28.55% compared to the baseline's 45.77%. This relative reduction of 37.62% is particularly noteworthy given the diverse nature of aphasic speech patterns and the differences in recording conditions and task structures between SONIVA and AphasiaBank.

The cross-disorder evaluation on DementiaBank also reported significantly better performances of the fine-tuned model (36% WER) compared to the baseline Whisper model (42% WER). Focusing on the TORGO database, the fine-tuned Whisper model showed a decrease in performance, from 45% (baseline) to 53% (fine-tuned). This counter-intuitive result highlights the challenges of generalising across different speech tasks (continous speech vs single word tasks), and different types of speech disorders. The limited sample size of TORGO ($N = 8$ subjects) may also have contributed to this outcome, not correctly representing dysarthria severity.

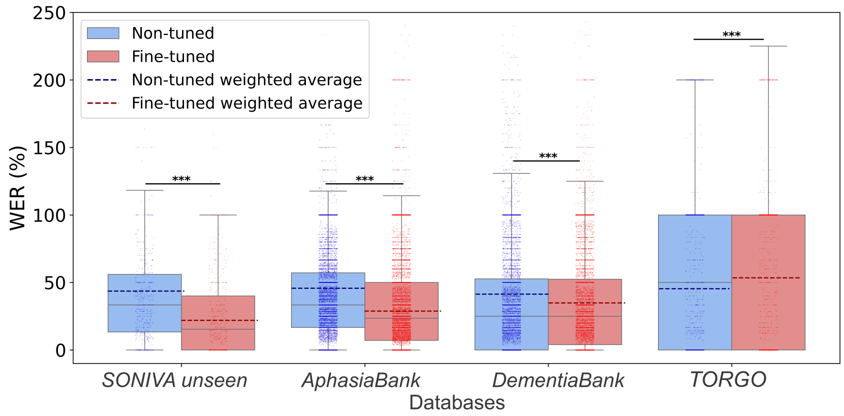

Figure 2: Distributions of *Word Error Rate* (WER) for transcriptions derived from *non-tuned* (red) and SONIVA *fine-tuned* (blue) Whisper, on the four different databases. Each dot correspond to an individual audio file. The dashed lines represent the weighted WER. Wilcoxon rank-sum tests were performed between the two models at a database level. *** : $p < 0.001$.

	Dataset - WER				
	SONIVA - validation	***SONIVA - unseen***	***AphasiaBank***	***DementiaBank***	***TORGO***
Non-tuned Model	39.60%	43.62%	45.77%	41.34%	45.40%
Fine-tuned w/SONIVA	21.51%	21.93%	28.85%	34.89%	53.45%
Fine-tuned w/TORGO	-	42.12%	43.92%	42.48%	26.44%♦

Table 1: Average *Word Error Rate* (WER) using off-the-shelf baseline Whisper (*non-tuned*) and *fine-tuned* on SONIVA or TORGO database. The average is the mean of the WER for the individual audio samples, weighted by the length of the sentence.
♦: obtained from the model fine-tuned on a subset of the TORGO, and tested on unseen speakers.

Given this unexpected performance decline, we conducted an additional experiment by fine-tuning Whisper exclusively on TORGO data, setting aside 2 out of 8 patients as unseen data (as detailed in Section 3.5). This TORGO-specific model showed dramatic improvement in WER on his own TORGO unseen set, decreasing from 45.40% (non-tuned) to 26.44%. This substantial enhancement demonstrates the model's capacity to adapt effectively to the specific speech task when provided with appropriate training data, even on patients not present in the training. Interestingly, this TORGO-fine-tuned model shows a worsened performance on DementiaBank, with WER increasing from 41.34% to 42.48%.

However, the TORGO-fine-tuned model's performance on SONIVA and AphasiaBank datasets remains relatively unchanged compared to the non-tuned model. This pattern indicates that the features learned from predominantly single-word tasks in context of dysarthric speech, does not transfer well to aphasia-related speech impairments captured through spontaneous long-form speech tasks, mirroring previous results on SONIVA fine-tuned model.

5 Discussion

Our Whisper ASR model, fine-tuned on speech from PwS, demonstrated significant generalisation capabilities both within and across speech disorders. The model's robustness within post-stroke speech impairments is evident by consistent WER scores across SONIVA's validation and test sets, indicating reliable out-of-sample prediction. Further validating its performance on aphasic speech, the results on AphasiaBank are particularly noteworthy. Despite differences in accent, test protocol, and speech type (predominantly spontaneous), the model maintained strong performance, suggesting successful adaptation of its attention mechanisms and language modeling components to diverse linguistic contexts. The WER achieved on both unseen SONIVA and AphasiaBank aligns or surpasses state-of-the-art performance on pathological speech recognition (see Sec. 2.3).

Moving to cross-disorder evaluation, the model's performance on DementiaBank revealed promising transfer learning capabilities for neurodegenerative dysfluencies. Our SONIVA-fine-tuned model demonstrated superior performance compared to previous studies in the field. For instance, in analysing speech from patients with fronto-temporal lobar degeneration, [56] reported WERs of 37% for semantic dementia and 61% for progressive nonfluent aphasia, while [57] achieved an average WER of 39.83% on DementiaBank across different speaker subsets. However, when testing our TORGO-fine-tuned model on DementiaBank, performance showed no improvement over the baseline (42.48% WER). This finding reflects the mismatch between TORGO's single-word training domain and DementiaBank's continuous speech data, highlighting the importance of training data structure in cross-disorder generalisation.

The performance degradation observed when testing SONIVA-trained Whisper on TORGO reveals fundamental challenges in adapting foundation models to single-word tasks. Whisper's architecture, optimised for continuous speech, employs several mechanisms specifically designed for sentence-level processing. These include long-range dependencies, positional encoding, and contextual word probabilities—components that become notably less effective when processing single-word utterances, which comprise the majority of TORGO's data. The architectural limitations manifest in multiple ways: positional encoding becomes irrelevant without sequential context, attention mechanisms lack sufficient tokens for meaningful contextual relationships, and the language model's predictive capabilities are underutilized without word dependencies. Additionally, the encoder's convolutional layers, designed for continuous speech patterns, may not optimally capture the isolated phonetic features crucial for single-word recognition. This architectural mismatch explains both the lack of improvement and the performance decline observed after fine-tuning on continuous post-stroke speech. Further, the mismatch between post-stroke and purely dysarthric speech may have stretched the model's ability to generalise effectively.

Indeed, as expected, in our additional experiment with TORGO-exclusive fine-tuning the model showed an improvement only on its own unseen test set, while maintaining the same performance on long-form speech. But importantly, it also did not degrade in performance. This behavior can be attributed, in part, to the limited size of the TORGO dataset, which likely lacked sufficient data to substantially alter the model's pre-trained weights optimised for continuous speech. Consequently, while the model successfully adapted to TORGO's single-word structure, it preserved its baseline capabilities on continuous speech without significant regression. This finding highlights how dataset size influences task-specific adaptation: smaller datasets may enable targeted improvements without compromising the model's general capabilities, effectively constraining the degree of specialisation.

6 Limitations and future work

Our study, while demonstrating the potential of fine-tuned Whisper models for post-stroke speech recognition, reveals several limitations that must be acknowledged. Firstly, we utilised the medium-sized Whisper model, leaving unexplored the potential benefits of its larger variant which, with its increased parameter count, could yield better performance. Additionally, we did not implement speech enhancement techniques during data pre-processing, potentially limiting the model's robustness. This omission may have affected the model's ability to handle background noise and distortions—common challenges in clinical recordings where optimal audio quality is not always guaranteed.

Furthermore, the performance degradation observed on the TORGO dataset when trained on SONIVA data suggests the need for architectural improvements, including selective selective layer freezing during fine-tuning, which may better handle grammatical errors and dysfluencies, or domain adaptation techniques for improved cross-disorder generalisation. Future work can explore hyperparameter optimisation, particularly focusing on the model's decision-making process for token transcription adjusting confidence thresholds. For specific disorders such as dementia, incorporating disorder-specific linguistic features, such as measures of idea density or semantic coherence, into the model architecture presents a possible avenue for enhancing performance. Additionally, to address the main challenges of this work about TORGO's short-form data, we could either merge SONIVA and TORGO training data to handle diverse speech scenarios, or develop two separate specialised models optimised for different utterance lengths.

Yet, it is worth noting that our SONIVA database encompasses both long-form and short-form speech, with patients also performing single-word tasks such as naming, reading, and repetition of 1-2 words. While this short-form data is currently undergoing labeling by our speech therapists, its future integration will enable us to explore the dual-model approach using exclusively post-stroke speech data, incorporating the necessary architectural improvements. This would allow us to address the observed challenges while maintaining the advantages of training exclusively on SONIVA's disorder-specific speech patterns.

Finally, a continuous extensive clinical validation will be crucial to ensure that the model's outputs align with expert human transcriptions and provide meaningful insights for clinical assessment and monitoring. Our interdisciplinary team, including speech-language pathologists and neurologists, provides critical domain expertise guiding our model development and evaluation. This collaboration ensures our technical advancements remain grounded in clinical relevance, potentially accelerating the translation of these ASR systems into practical tools for speech and language disorder management.

7 Conclusion

This study demonstrates the potential of fine-tuning large-scale foundation models for post-stroke speech recognition, achieving significant improvements in transcription accuracy across diverse neurological pathologies affecting speech. While challenges persist in cross-disorder generalisation and specific audio characteristics, our work represents a substantial advance toward robust, clinically applicable universal ASR systems. The enhanced transcription accuracy could improve assessment and monitoring of speech disorders, streamlining clinical workflows and enabling more personalised, data-driven rehabilitation approaches. As we continue to refine and clinically validate these models, MFM in AI-assisted speech analysis is poised to become an invaluable tool in managing speech and language disorders, ultimately enhancing patient care and outcomes.

8 Acknowledgements

The authors would like to thank A. Coghlan and N. Parkinson for assistance with the IC3 study, and J. Friedland, K. Stephenson, G. Gvero, Z. Zaskorska and C. Leong for labelling the speech data. We are deeply grateful to the patients, clinical teams, Prof. Cathy Price, the PLORAS team, the AphasiaBank and DementiaBank for providing additional data, as well as the team from the University of Toronto for the TORGO database. For this work, infrastructure support was provided by the NIHR Imperial Biomedical Research Centre and the NIHR Imperial Clinical Research Facility. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. Funding support was provided to G.S. (UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare:EP/S023283/1]), F.G. (Medical Research Council P79100) and S.B. (EPSRC IAA -PSP415 and MRC IAA -PSP518).

References

[1] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[2] 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.

[3] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, "Dementiabank: Theoretical rationale, protocol, and illustrative analyses," *American Journal of Speech-Language Pathology*, vol. 32, no. 2, pp. 426–438, 2023.

[4] A. Radford, J. W. Kim, and T. e. a. Xu, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[5] E. R. Olafson et al, "Data-driven biomarkers better associate with stroke motor outcomes than theory-based biomarkers," *Brain Commun.*, 2024.

[6] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins, "Automatic word naming recognition for an on-line aphasia treatment system," *Computer Speech & Language*, vol. 27, no. 6, pp. 1235–1248, 2013.

[7] J. S. Damico, N. Müller, and M. J. Ball, *The handbook of language and speech disorders*. Wiley Online Library, 2010.

[8] G. Dyukova, Z. M. Glozman, E. Y. Titova, E. Kriushev, and A. Gamaleya, “Speech disorders in right-hemisphere stroke,” *Neuroscience and Behavioral Physiol.*, vol. 40, pp. 593–602, 2010.

[9] Z. Ghoreyshi, R. Nilipour, N. Bayat, S. S. Nejad, M. Mehrpour, and T. Azimi, “The incidence of aphasia, cognitive deficits, apraxia, dysarthria, and dysphagia in acute post stroke persian speaking adults,” *Indian J. of Otolaryngology and Head & Neck Surgery*, vol. 74, no. Suppl 3, pp. 5685–5695, 2022.

[10] J. D. Stefaniak, F. Geranmayeh, and M. A. Lambon Ralph, “The multidimensional nature of aphasia recovery post-stroke,” *Brain*, vol. 145, no. 4, pp. 1354–1367, 2022.

[11] M. C. Brady and H. e. a. Kelly, “Speech and language therapy for aphasia following stroke,” *Cochrane database of systematic reviews*, no. 6, 2016.

[12] S. S. Mahmoud, R. F. Pallaud, A. Kumar, S. Faisal, Y. Wang, and Q. Fang, “A comparative investigation of automatic speech recognition platforms for aphasia assessment batteries,” *Sensors*, vol. 23, no. 2, p. 857, 2023.

[13] D. Le, K. Licata, and E. M. Provost, “Automatic quantitative analysis of spontaneous aphasic speech,” *Speech Communication*, vol. 100, pp. 1–12, 2018.

[14] R. Palmer and P. e. a. Enderby, “Computer therapy compared with usual care for people with long-standing aphasia poststroke: a pilot randomized controlled trial,” *Stroke*, vol. 43, no. 7, pp. 1904–1911, 2012.

[15] 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 *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 2023, pp. 182–190.

[16] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[17] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, “Publicly available clinical bert embeddings,” *arXiv preprint arXiv:1904.03323*, 2019.

[18] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[19] S. B. Patel and K. Lam, “Chatgpt: the future of discharge summaries?” *The Lancet Digital Health*, vol. 5, no. 3, pp. e107–e108, 2023.

[20] Y. He, F. Huang, X. Jiang, Y. Nie, M. Wang, J. Wang, and H. Chen, “Foundation model for advancing healthcare: Challenges, opportunities, and future directions,” *arXiv preprint arXiv:2404.03264*, 2024.

[21] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.

[22] J. Qiu, W. Yuan, and K. Lam, “The application of multimodal large language models in medicine,” *The Lancet Regional Health–Western Pacific*, vol. 45, 2024.

[23] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth *et al.*, “Me llama: Foundation large language models for medical applications,” *arXiv preprint arXiv:2402.12749*, 2024.

[24] D. A. Wiepert, R. L. Utianski, J. R. Duffy, J. L. Stricker, L. R. Barnard, D. T. Jones, and H. Botha, “Speech foundation models in healthcare: Effect of layer selection on pathological speech feature prediction.”

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[26] H. Soltau, I. Shafran, A. Ottenwess, R. Joseph Jr, R. L. Utianski, L. R. Barnard, J. L. Stricker, D. Wiepert, D. T. Jones, and H. Botha, “Detecting speech abnormalities with a perceiver-based sequence classifier that leverages a universal speech model,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.

[27] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, “Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis,” *arXiv preprint arXiv:2406.08568*, 2024.

[28] Y. Jiang, T. Wang, X. Xie, J. Liu, W. Sun, N. Yan, H. Chen, L. Wang, X. Liu, and F. Tian, "Perceiver-prompt: Flexible speaker adaptation in Whisper for chinese disordered speech recognition," in *Interspeech 2024*, 2024, pp. 2025–2029.

[29] J. Li and W.-Q. Zhang, "Whisper-based transfer learning for alzheimer disease classification: Leveraging speech segments with full transcripts as prompts," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 211–11 215.

[30] G. Sanguedolce, S. Brook, D. C. Gruia, P. A. Naylor, and F. Geranmayeh, "When whisper listens to aphasia: Advancing robust post-stroke speech recognition," in *Interspeech*, 2024.

[31] J. Lee, Y. Choi, T.-J. Song, and M.-W. Koo, "Inappropriate pause detection in dysarthric speech using large-scale speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 486–12 490.

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

[33] P. Best, S. Cuervo, and R. Marxer, "Transfer learning from Whisper for microscopic intelligibility prediction," in *Interspeech 2024*, 2024, pp. 3839–3843.

[34] C. Graham and N. Roll, "Evaluating openai's Whisper asr: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024.

[35] M. Zusag, L. Wagner, and T. Bloder, "Careful Whisper - leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification," in *Interspeech 2023*, 2023, pp. 3013–3017.

[36] Y. Liu and Y. e. a. Qin, "Disordered speech assessment using kullback-leibler divergence features with multi-task acoustic modeling," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 61–65.

[37] J. Tang, W. Chen, X. Chang, S. Watanabe, and B. MacWhinney, "A new benchmark of aphasia speech recognition and detection based on e-branchformer and multi-task learning," *arXiv preprint arXiv:2305.13331*, 2023.

[38] H. Kim, M. H. Johnson, J. Gunderson, A. Perlman, T. Huang, K. Watkin, S. Frame, H. V. Sharma, and X. Zhou, "Uaspeech," 2023. [Online]. Available: https://dx.doi.org/10.21227/f9tc-ab45

[39] C. Bhat, A. Panda, and H. Strik, "Improved asr performance for dysarthric speech using two-stage dataaugmentation." in *Interspeech*, 2022, pp. 46–50.

[40] D.-C. 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, no. 11, p. e076653, 2023.

[41] D. C. Gruia, V. Giunchiglia, A. Coghlan, S. Brook, S. Banerjee, J. Kwan, P. J. Hellyer, A. Hampshire, and F. Geranmayeh, "Online monitoring technology for deep phenotyping of cognitive impairment after stroke," *medRxiv*, 2024.

[42] M. L. Seghier and P. et al., "The ploras database: a data repository for predicting language outcome and recovery after stroke," *Neuroimage*, vol. 124, pp. 1208–1212, 2016.

[43] K. Swinburn, G. Porter, and D. Howard, "Comprehensive aphasia test," *APA PsycTests*, 2004.

[44] G. Conti-Ramsden, "CLAN (Computerized Language Analysis)," *Child Language Teaching and Therapy*, vol. 12, no. 3, pp. 345–349, 1996.

[45] P. Appelros, B. Stegmayr, and A. Terént, "Sex differences in stroke epidemiology: a systematic review," *Stroke*, vol. 40, no. 4, pp. 1082–1090, 2009.

[46] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press, 2014.

[47] M. Perez, Z. Aldeneh, and E. M. Provost, "Aphasic speech recognition using a mixture of speech intelligibility experts," *arXiv preprint arXiv:2008.10788*, 2020.

[48] G. Sanguedolce, J. Guðnason, D. C. Gruia, S. Brook, F. Geranmayeh, and P. A. Naylor, "Voice-source analysis of stroke-induced pathological speech," in *[Manuscript submitted for publication] ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[49] I. G. Torre and M. e. a. Romero, “Improving aphasic speech recognition by using novel semi-supervised learning methods on Aphasiabank for English and Spanish,” *Applied Sciences*, vol. 11, no. 19, p. 8872, 2021.

[50] J. Robert, M. Webbie *et al.*, “Pydub,” 2018. [Online]. Available: http://pydub.com/

[51] S. Tomar, “Converting video formats with ffmpeg,” *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[52] “Sox sound exchange,” accessed: 2023-10-20. [Online]. Available: http://sox.sourceforge.net

[53] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” *arXiv preprint arXiv:1508.07909*, 2015.

[54] A. Paszke and S. e. a. Gross, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[55] T. Wolf and L. e. a. Debut, “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.

[56] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. L. G. Tempini, and J. Ogar, “Learning diagnostic models using speech and language measures,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 4648–4651.

[57] L. Zhou, K. C. Fraser, F. Rudzicz *et al.*, “Speech recognition in alzheimer’s disease and in its assessment.” in *Interspeech*, vol. 2016, 2016, pp. 1948–1952.

A Appendix

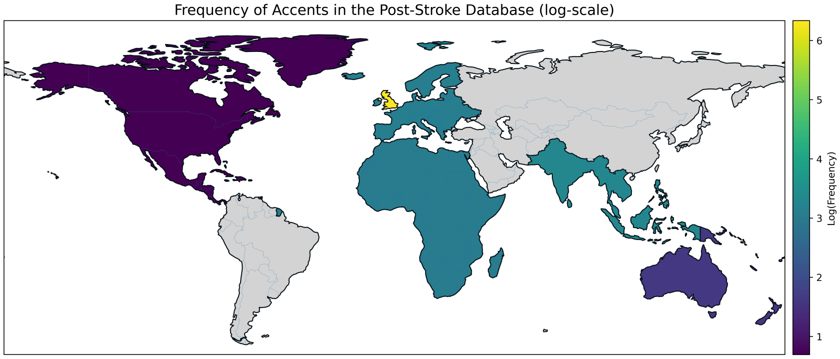
Frequency of Accents in the Post-Stroke Database (log-scale)
Log(Frequency)
1
2
3
4
5
6

Figure 3: Accent Map of the *PwS* database

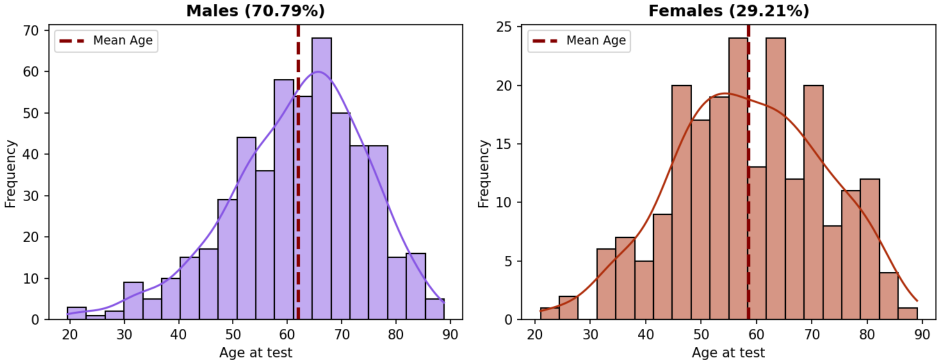

Figure 4: Age Gender Distribution of the *PwS* database

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] ,

   Justification: Yes, the contribution are widely explored in abstract and introduction, as well as the motivation and the final aim.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] ,

   Justification: The limitation are briefly reported already in the results, but then these are more explored in the specific paragraph “Limitation and future work"

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: no theoretical results are reported

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Described in Methods

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Patients data privacy does not allow yet to share the data from the PwS database.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Reported in Methods

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Reported in Methods, Results and figures captions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Described in section 3.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms to the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper discusses solely the potential improvement in healthcare given the automation of assessment and therapy for speech disorders.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks present.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and owners were properly credited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No assets released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Information not fully disclosed yet due to anonymity, but part of the task specifics are included in the Methods section.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: All the risks were disclosed to the subjects during data collection, but not reported here, and approved by the IRB, which is not disclosed due to anonymity. In the camera-ready version, this information will be provided.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.