Benchmarking Automatic Speech Recognition for Aphasia: A Clinical Evaluation Framework

Navya Martin Kollapally
Deaprtment of Computer Science
Kean University
New Jersey, USA
nmartink@kean.edu

Christa Akes
Department of Communication Science and Disorders
Kean University
New Jersey,USA
christa.akers@kean.edu

***Abstract*— AI tools for speech therapy represent more than innovation; they signify necessity. With 89% of clinicians facing overwhelming caseloads and therapy wait times averaging 3–6 months, the demand for scalable rehabilitation support has never been higher. Although AI-driven communication platforms increasingly provide real-time feedback and sustained engagement, current Automatic Speech Recognition (ASR) systems still face significant limitations in accurately processing disordered speech such as aphasia. Aphasia is an acquired language disorder, the speech of individuals with aphasia is characterized by unintelligible words, jargon, or non-words. To add to this the speaker with aphasia may not recognize their errors and mostly have difficulty in comprehension. These limitations hinder equitable access to AI-driven rehabilitation tools. To bridge this gap, the first contribution of this study is evaluating four state-of-the-art ASR models, such as Whisper, NeMo-Conformer, Wav2Vec 2.0, and SpeechBrain, through the lens of Speech-Language Pathologist (SLP). The second contribution of the paper is utilizing a comprehensive benchmarking framework to assess how effectively these models capture clinically relevant aspects of aphasic speech, including lexical, syntactic, and fluency-related features. For evaluating the transcribed text, a combination of quantitative, qualitative (human-expert based), and linguistically grounded evaluation metrics is used, such as verb error rate, noun error rate, mean dependency length etc. of transcribed text.**

Keywords—Automatic Speech Recognition (ASR), OpenAI whisper, NVIDIA nemo, Facebook AI wave2vec 2.0, SpeechBrain, Aphasia ASR model

I. INTRODUCTION

Aphasia is an acquired communication disorder caused by brain injury, commonly due to stroke. Because aphasia affects the ability to use and understand spoken and written language, it can have negative consequences on employment, social participation, and quality of life. During treatment, SLPs address the underlying language impairments caused by aphasia and assist Persons with Aphasia (PWA) with strategies that support communication. PWA have reported that regaining their ability to communicate, particularly in conversation, is a rehabilitation priority [1]. To set individualized therapy goals, it is important that SLPs assess language beyond the level of the sentence (i.e., discourse)[2]; however, SLPs routinely report that they lack the time necessary to collect and transcribe discourse. It is necessary, therefore, to identify methods that may ease the burden associated with the transcription of aphasic discourse[3].

ASR models are rapidly advancing, the reliability of ASR models to assist clinicians in assessing, monitoring and supporting individuals with speech delays is still not well-established. ASR models are predominantly trained on fluent, well-structured data and often encounter difficulties when fed with disordered speech [4]. In this era of developing personas and interactive agents without "uh-oh" moments, it is essential that we also focus on the population of patients with disordered speech who could benefit from ASR models[5]. It is estimated that 100,000-180,000 people acquire aphasia each year in the United States[6, 7]. According to the National Aphasia Association, over 2 million people in the United States currently live with aphasia[8]. Approximately 38% of individuals who suffer a stroke develop aphasia at the time of their stroke event, and about 25% of stroke survivors still present with aphasia three months later[9]. Studies have shown that integrating ASR into clinical workflows can streamline data collection and analysis, allowing clinicians to focus more on intervention planning and individualized therapy rather than repetitive documentation tasks.

Despite its prevalence and impact, aphasic speech poses unique challenges to ASR: reduced fluency, lexical retrieval difficulties, syntactic fragmentation, and increased disfluencies[10]. Traditional ASR evaluation metrics such as Word Error Rate (WER) capture only surface alignment and fail to account for the linguistic richness, syntactic structure, or fluency breakdowns that are clinically meaningful in aphasia research[11, 12]. Therefore, there is a pressing need for a benchmarking pipeline that extends beyond WER and evaluates transcription performance through the lens of SLP.

We present a comprehensive benchmarking and evaluation architecture for aphasic speech recognition that integrates computational metrics, psycholinguistic, and clinical validity along with human evaluations by two experts trained in speech pathology. Our framework simultaneously performs automatic transcription from recordings in Aphasia Bank, and transcripts generated by the four models are analyzed on a dual-layer assessment for disordered speech. The system evaluates four ASR models Whisper, NeMo-Conformer, Wav2Vec 2.0, and SpeechBrain using a unified pipeline that standardizes data preprocessing, acoustic normalization, and metric computation.

Unlike prior work that focuses solely on error rates, our approach introduces a dual-layer assessment: (1) automated metrics, including traditional Word Error Rate (WER) and Character Error Rate (CER) alongside novel psycholinguistic informed indicators such as Verb Error Rate (VER), Noun Error Rate (NER), Verb/Noun Ratio Deviation (ΔVNR). Mean length of Utterance etc.; and (2) human evaluation, where two trained speech-language pathologists (SLPs) rate the intelligibility, meaning fidelity, grammaticality, and clinical usability of model outputs. All

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356945

the benchmarking code across models, and dual-assessment metrics code are available in GitHub as open-source resources[13]. By analyzing SLP's evaluation with automated linguistic metrics, this paper establishes a system for a clinically interpretable benchmarking environment for aphasic-speech ASR.

II. Related Work

The first manuscript on using ASR for aphasic speech focused on isolated word recognition using small, task-specific vocabularies in English and Portuguese[14]. With the rapid development of transformer-based and self-supervised ASR architectures, recent studies have achieved considerable accuracy levels for controlled naming tasks[15]. However, performance remains far lower and variable when moving from naming tasks to continuous, large-vocabulary recognition of aphasic speech. Hence, achieving robust continuous recognition of aphasic speech within large-vocabulary settings remains an open challenge, as fluency variation, lexical retrieval failures, and prosodic irregularities continue to impede performance[16].

Le and Mower-Provost [17] utilized AphasiaBank [18] to establish a foundational benchmark for aphasic speech recognition using a Hybrid Hidden Markov Deep Neural network (HMM-DNN) architecture, demonstrating that speaker adaptation techniques can substantially mitigate variability and improve recognition accuracy for speakers with more severe impairments. This study [17] further explored out-of-domain adaptation strategies, such as discriminative pretraining, which leveraged AphasiaBank data to enhance performance on smaller datasets like University of Michigan Aphasia dataset (UMAP). Importantly, they proposed the per-speaker Phone Error Rate (PER) [17] as a critical evaluation metric and showed that Aphasia Quotient (AQ) scores from the Western Aphasia Battery–Revised (WAB-R) [19]. The authors' findings indicate that individuals with more severe aphasia present greater recognition challenges for ASR systems. However, their approach assumes uniform intelligibility across a speaker's entire recording. In contrast, the study we proposed in this manuscript extends this view by modeling variability at the utterance level and evaluating linguistic and human-perceived aspects of ASR quality.

Subsequent work addressed variability across aphasia types and severity levels. Perez et al. [20] applied a mixture-of-experts model guided by a speech intelligibility detector to reduce phone error rates across severity stages in aphasic speech. Earlier ASR systems relied on a per-speaker recognition, as these systems generally treated speech intelligibility as a stable, speaker-level property. However, Perez et al.[20] observed that intelligibility in aphasia can fluctuate substantially across shorter linguistic units, for instance, at the utterance or even frame level, depending on lexical retrieval success, fatigue, or contextual support. This variability challenges speaker-level modeling and highlights the need for fine-grained, segment-aware evaluation frameworks.

Alyahya et al.[21] demonstrated that noun–verb differences in aphasia diminish when lexical variables are matched, suggesting that linguistic analysis must control for frequency and imageability effects. These insights from [21] motivated us to evaluate the ASR systems by WER and lexical-class fidelity such as verb error rate, noun error rate, and verb/noun ratio deviation to capture clinically meaningful structures beyond surface accuracy.

Recently, Chatzoudis et al. [22] proposed a zero-shot multilingual ASR framework using multilingual pretraining to detect aphasia in low-resource languages. They [22] extend the ASR/aphasia literature by targeting cross-lingual aphasia detection: training on English data and evaluating in Greek and French using language-agnostic features extracted from ASR transcripts. Their results confirm that ASR quality and language mismatch significantly affect classification accuracy, underscore emphasizing the necessity-adapted ASR systems for speech-language disorders. This work adds a multilingual transfer-learning dimension to the field, complementing the primarily English-centric recognition benchmarks. However, it focuses on classification (detection) rather than full transcription and lexical/syntactic fidelity of ASR outputs.

With the advancement of LLMs recently, Sanguedolce et al. [23] fine-tuned OpenAI's Whisper model on SONIVA, a clinically curated dataset of post-stroke aphasic speech, to improve automatic speech recognition for people with aphasia. Their work demonstrated that large-scale transformer architectures can generalize better to disordered speech than traditional HMM- or RNN-based systems, achieving lower word-error rates and improved handling of disfluencies and phonological errors. However, their evaluation relied primarily on acoustic-level metrics such as WER, without deeper linguistic or clinical analyses of the output. The study [23] also focused on a single ASR architecture and did not assess cross-model variability or human-perceived intelligibility of transcriptions.

Across studies, inconsistencies in evaluation metrics, data composition, and preprocessing impede reproducibility. While newer systems address efficiency and robustness, most still rely solely on WER. There remains a greater call to establish a clinically interpretable benchmarks that align computational accuracy with human-perceived intelligibility and therapeutic utility.

III. Background

A. Dataset and Pre Processing

AphasiaBank [24] was established in 2007 by Brian MacWhinney and Audrey Holland, funded by the U.S. National Institute on Deafness and Other Communication Disorders (NIDCD). Aphasia Bank is a shared multimedia repository for systematically studying aphasic discourse. It contains audio-video recordings and transcriptions in Codes for the Human Analysis of Transcripts i.e. CHAT format. Transcripts are available from individuals with aphasia and matched controls. It is collected using standardized tasks such as picture description, story retelling, procedural explanations, and personal narratives. The corpus includes over 140 participants with aphasia and 120 controls from multiple research sites. It has since been widely used for clinical and computational research. AphasiaBank enables detailed phonological, lexical, and syntactic annotation through the integrated CLAN analysis software.

1) Open AI Whisper

Whisper [25] is a large-scale encoder-decoder transformer model developed by OpenAI for multilingual and multitask speech processing. Trained on a diverse dataset of 680,000 labeled audio hours, including 65% English, 17% multilingual, and 18% translation data. Whisper's training leverages weak supervision at scale without self-training methods. Whisper processes 80-channel log-Mel spectrograms using convolutional layers and sinusoidal positional encodings in the encoder, which help it capture long-range patterns through a series of transformer blocks. The decoder follows a similar design with learned positional embeddings, which enable the model to handle tasks such as transcription, translation, and language identification. The Whisper model offers sizes ranging from tiny, with 39 million parameters, to large, with 1.55 billion parameters, with a total size of 2.9 GB.

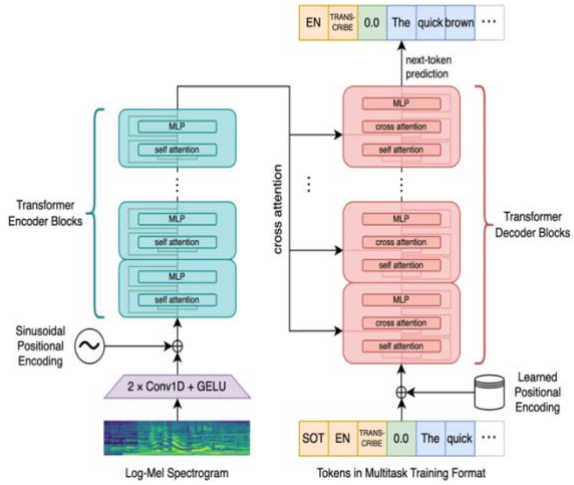

Fig 1: A snippet from the OpenAI whisper architecture diagram

2) NVIDIA Nemo

The core building block in NVIDIA NeMo[26] is called *Neural Module* (NM). A Neural Nodule represents a *logical* part of a neural network such as a language model. NeMo consists of: (1) NeMo Core: fundamental building blocks for all neural models and type system and (2) NeMo collections: pre-built neural modules for particular domains such as ASR is "nemo_asr," and for Natural Language Processing (NLP) there is nemo_nlp.The "nemo_asr" is a collection of neural modules and helper functions that can be used to train and evaluate ASR models. It currently supports two model types: CTC-based and sequence-to-sequence attention-based.

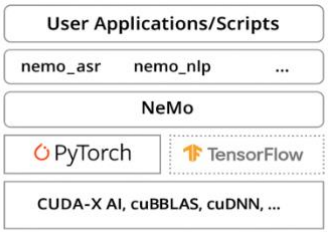

Fig 2: Nemo as a framework-agnostic toolkit which follows an execution model of no computation is called until done[26].

1) Facebook AI wave2vec 2.0

Wav2Vec 2.0, introduced by Baevski et al. [27] is a self-supervised learning framework for speech representation. The model learns directly from raw audio waveforms without requiring manual transcriptions during pretraining. Wave2Vec 2.0 is composed of a multi-layer convolutional feature encoder which takes as input raw audio and outputs a latent speech representation. This stack of convolutional layers is called a feature encoder. Followed by a transformer context network to capture long-range dependencies across time by contextualizing the encoded features. The output of the feature encoder is fed to a context network, which follows the transformer architecture. Instead of fixed positional embeddings, which encode absolute positional information, the authors use a convolutional layer. To pre-train the model, Baevski et al[27]. mask a certain proportion of time steps in the latent feature encoder space like the masked language modeling in BERT [9].

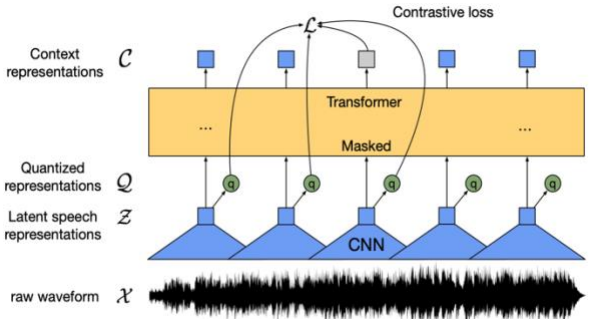

Fig 3: Visualization of the wave2vec framework[27] .

2) SpeechBrain

SpeechBrain [28] is an all-in-one PyTorch-based toolkit designed to facilitate speech processing technologies' development, portability, and ease of use. SpeechBrain follows a library-style collection of modular and standalone building blocks, including practical routines for data loading, decoding, signal processing, and other convenient utilities. Brain class is the center of the architecture, which manages device allocation, training loops, and checkpointing. Training begins by calling the script with a set of hyperparameters.

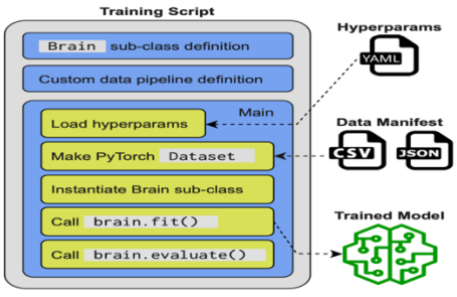

Fig 4: SpeechBrain as a toolkit architecture [28].

These hyperparameters, declared in human-readable YAML format, contain the location of one or more data manifest files using CSV or JSON formats. SpeechBrain relies on an extended version of YAML called HyperPyYAML. HyperPyYAML is a general tool for specifying hyperparameters. SpeechBrain complements standard PyTorch data loading by addressing the typical challenges of working with speech and complex data transformation pipelines by PyTorch datatset. The batch size will be dynamically adjusted according to sentence length, leading to improved efficiency and better management of available GPU memory.

B. Qualitative Metric

This study utilizes the conventional accuracy metrics, such as Word Error Rate (WER) [29] and Character Error Rate (CER) [30], to understand and capture surface-level transcription accuracy. Alongside Verb Error Rate (VER) and Noun Error Rate (NER), quantifying part-of-speech–specific distortions [31], and Verb/Noun Ratio Deviation (ΔVNR) [32], which measures deviations in lexical balance relative to reference transcripts, is an important proxy for grammatical and semantic integrity. These linguistic measures aim to understand how accurately the models reflect the structure and flow of natural speech. The syntactic complexity was analyzed through Complex Sentence Ratio (CSR) [33], representing the proportion of sentences containing subordinate clauses, and Mean Dependency Depth (MDD)[34], reflecting the average hierarchical distance between words in a sentence and thus indicating syntactic depth. Alongside from an SPL perspective we also count the number of Verb, Noun, Verb error rate, Noun error rate and mean length of utterance.

IV. Methodology

We have divided the methodology into four main section as follows:

A. Dataset and PreProcesisng

We utilized the audio recordings and corresponding transcripts from the AphasiaBank corpus, a component of the larger TalkBank[35] database.. For benchmarking, we selected Dr. Robert Marshall's Anomia corpus, which consists of videos and transcripts of around 250 minutes. We intentionally chose older videos alongside different age groups of patients, including diverse populations in Marshall's corpus. This approach emulates the real-world situation as closely as possible. There are 11 folders containing video recordings of adults with neurogenic communication disorders. The folders are labeled according to speech/language disorder diagnoses when the recording was made. This key describes what the clinician and the patient are doing, e.g., conversing, discussing a certain topic, and various tasks the clinician asks the patient to do, e.g., picture naming, reading of words, repetition, counting, etc. These are non-protocol, audio, elicited speech and language data from adults with a variety of neurogenic communication disorders, including aphasia (Anomia, Broca's, Conduction, Transcortical, Wernicke's), apraxia of speech, language of confusion, dementia, dysarthria, neurogenic stuttering, and right hemisphere communication disorder. Each audio sample was stored as an .mp4 file, accompanied by its aligned reference transcript. The dataset was partitioned based on individual audio files of varying durations, resulting in 78% training (195 minutes), 13% validation (55 minutes), and 9% test (21 minutes) splits. Notably, a *stratified partitioning strategy* was adopted to ensure balanced representation across different aphasia severity levels, maintaining clinical diversity within each subset—different severities of aphasia (e.g., mild, moderate, severe). Instead of randomly splitting all files, you stratified the split so that each subset (train, validation, test) contains roughly the same proportion of each severity level.

As in Figure 5 all audio files were extracted and resampled to 16 kHz during preprocessing, converted to mono, and amplitude-normalized using the *librosa* library. Missing waveform files were automatically flagged, with conversion from MP4 to WAV facilitated via *ffmpeg*. Corresponding reference transcriptions were parsed from CHAT files, extracting content from the *PAR*, *INV*, and *CHI* tiers to form the ground-truth text used for model evaluation. To ensure balanced representation across speech characteristics, the dataset was subsequently partitioned using *stratified splitting* based on energy and speaker-level distributions.

B. Model Benchmarking

To evaluate the ASR pipeline on aphasic speech, we implemented a comprehensive ASR benchmarking pipeline that integrates Whisper (OpenAI), NeMo-Conformer (NVIDIA), Wav2Vec 2.0 (Meta AI / Facebook), and SpeechBrain. All models were loaded in their pretrained configurations and evaluated under identical preprocessing, decoding, and scoring conditions to ensure fair comparison. Whisper models are known for their robustness to noise and spontaneous speech, making them a strong candidate for clinical applications[36]. We used the large model configuration for transcription inference with 1550 million parameters, with a VRAM requirement of around 10GB.NeMo-Conformer (Conformer-CTC) was deployed using the pretrained *stt_en_conformer_ctc_small* checkpoint, with 115 million parameters for the large model. Wav2Vec 2.0, learns latent speech representations through self-supervised pretraining on raw waveforms, followed by fine-tuning with Connectionist Temporal Classification (CTC). The model was accessed via the Hugging Face Transformers interface (*facebook/wav2vec2-base-960h*) with 91 million parameters, enabling direct integration within the benchmarking framework.SpeechBrain (Transformer + LM Integration): SpeechBrain provides a modular PyTorch-based ASR framework with flexible architecture definitions. We utilized its *asr-transformer-transformerlm-librispeech* pretrained model with around 93M parameter.

Fine-tuning was conducted separately for each Automatic Speech Recognition (ASR) backbone Whisper, NeMo Conformer-CTC, Wav2Vec 2.0, and SpeechBrain. All models were initialized from publicly available checkpoints and optimized using the Marshall Aphasia Corpus training split (78%) while monitoring validation performance on a held-out set (13%).Training employed the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\varepsilon = 1 \times 10^{-8}$) with an initial learning rate of 1×10^{-4} and a cosine decay schedule with 10 % warm-up. Each model was fine-tuned for up to 10 epochs using early stopping (patience = 5) to prevent overfitting. Mini-batch size was set to 16, and time- and frequency-masking was applied to improve robustness to acoustic variability.

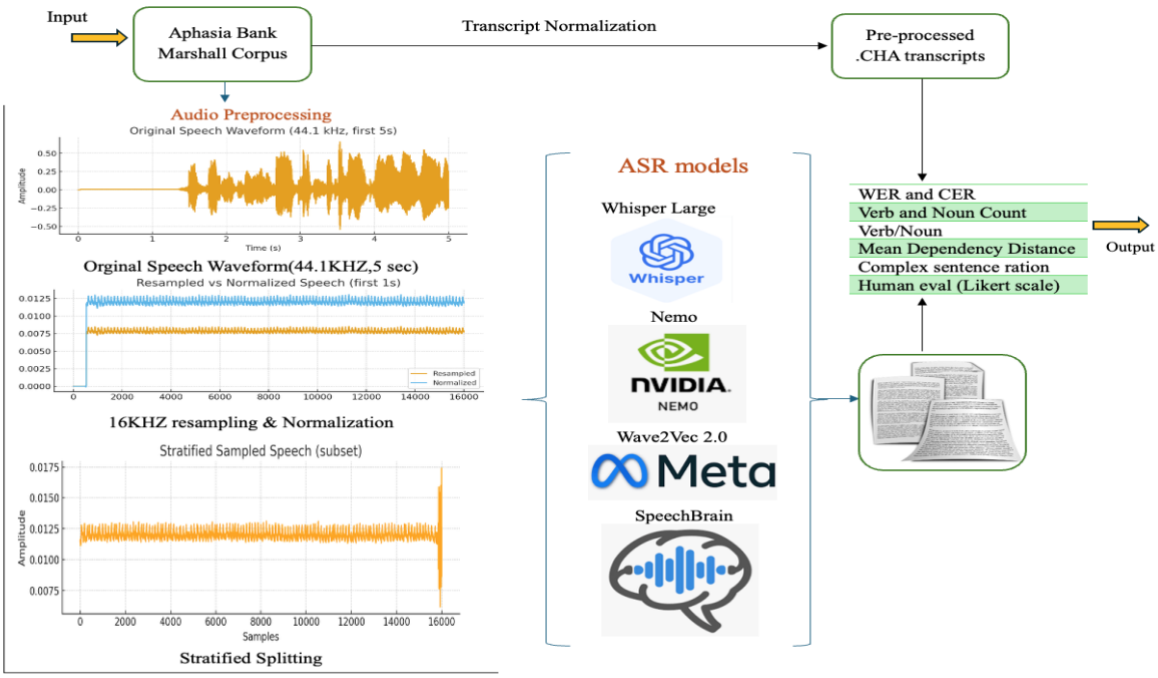

Fig 5: Architecture diagram including four model benchmarking on lexical, syntactic and fluency related metric

For Whisper and Wav2Vec 2.0, the convolutional feature encoder was frozen during the first two epochs to stabilize gradient propagation before full end-to-end training. NeMo Conformer-CTC optimization was performed with frame-level CTC loss, while SpeechBrain used its built-in Brain.fit() loop with dynamic batching and automatic mixed precision (AMP). Dropout (0.1) and L_2 regularization (weight decay = 1×10^{-6}) were applied throughout. Training was executed on four NVIDIA A100 GPUs and took a total of 6 hrs. 32 minutes for finetuning.

After convergence, model checkpoints yielding the lowest validation loss were selected for final inference on the unseen test set (9 %). Model output text after transcription was stored in structured CSV format alongside reference transcriptions for downstream analysis. Transcriptions were processed through an automated linguistic feature extraction pipeline to quantify lexical, syntactic, and fluency characteristics.

C. *Evaluation Metrics and Analysis*

The evaluation framework encompasses both quantitative and qualitative metrics, along with human assessments, to capture both transcription accuracy and its clinical interpretability. Each transcribed text output from the four ASR models was compared to the clinician-verified reference transcripts available in the aphasia bank across three complementary dimensions: accuracy, linguistic complexity, and fluency.

I. Quantitative Metrics

Automated evaluation was performed using Word Error Rate (WER) and Character Error Rate (CER), computed via the jiwer library, as primary measures of transcription accuracy. Beyond these metrics, a suite of linguistic, syntactic, and psycholinguistic indicators was extracted from both reference and ASR-generated transcriptions using NLTK and spaCy. These included:

Lexical diversity metrics like Word count, unique word count, and Mean Length of Utterance (MLU). *Syntactic metrics* like Mean Dependency Depth (MDD), and Complex Sentence Ratio (CSR).

Fluency and disfluency metrics: Filler count (e.g., “um”, “uh”), repetition count, and function word ratio.

Psycholinguistic informed metrics: Verb count, noun count, and the Verb/Noun Ratio Deviation (ΔVNR).

a. Word Error Rate (WER)

WER is a standard metric of the performance of a speech recognition system(Equation 1). The WER metric typically ranges from 0 to 1, where 0 indicates that the compared pieces of text are identical, and 1 (or larger) indicates that they are completely different with no similarity.

$$\mathrm{WER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{S} + \mathrm{D} + \mathrm{C}} \quad (1)$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words.

b. Character Error Rate (CER)

CER is a measure of the accuracy of text recognition systems by calculating the percentage of characters that are incorrect in a predicted output compared to a correct reference, using substitutions, deletions, and insertions. It is calculated by dividing the number of incorrect characters by the number of characters in the reference text as in Equation 2.

$$\mathrm{CER} = \frac{S_c + D_c + I_c}{S_c + D_c + C_c} \quad (2)$$

where the terms correspond to character-level edit operations. S_c is the number of character substitutions, D_c is the number of character deletions, I_c is the number of character level insertions, C_c is the total number of characters.

c. Mean Length Utterance (MLU)

MLU is a good marker of language impairment. It is the number of words or morphemes in each of their spontaneous utterances. It can be used to benchmark language acquisition and is used to compare language intervention outcomes in children with autism.

d. Mean Dependency Depth (MDD)

MDD quantifies *the average linear distance* between heads and dependents in a syntactic dependency tree, it a measure of syntactic complexity and processing difficulty.

$$MDD = \frac{1}{n}\sum_{i=1}^{n}|DD_i| \quad (3)$$

In Equation 3 the n is the total number of dependency pairs in a sentence, DD_i is the dependency distance of the *i-th* head–dependent pair. $|DD_i|$ = absolute value of the distance (since dependency direction can be leftward or rightward). A higher MDD means syntactically complex sentences with longer dependencies (e.g., multiple embeddings or subordinate clauses). Similarly, a lower MDD indicates a simpler, flatter syntactic constructions often characteristic of aphasic or telegraphic speech.

e. *Complex Sentence Ratio (CSR)*

Complex Sentence Ratio (CSR) measures the proportion of sentences that contain at least one subordinate or dependent clause relative to the total number of sentences.

$$\text{CSR}=\frac{N_{complex}}{N_{total}} \quad (4)$$

Where N_{complex} is the number of sentences containing one or more subordinate clauses (e.g., adverbial, complement, or relative clauses), and N_{total} is the total number of sentences in the transcript. A Higher CSR indicates a greater syntactic elaboration, typical of fluent or recovered speech. And a Lower CSR indicates a reduced clause embedding, indicating simplified syntax and limited grammatical productivity common in aphasic or telegraphic speech.

f. *Verb Count*-It is the total number of all verbs in the sample.

g. Noun count- It is the total number of all nouns in the sample.

h. Clause count- It is the total number of clauses (i.e., group of words with a subject and verb) in the sample.

i. Filler count-It is the total number of nonlinguistic fillers (e.g., um, uh) in the sample.

j. Repetition count- It is the total number of repetitions in the sample.

k. Verb/Noun Ratio Deviation

The Verb/Noun Ratio Deviation (ΔVNR) is typically defined as the difference between an observed verb–noun ratio and a reference or expected ratio.

$$\Delta VNR = \left(\frac{V_o}{N_o}\right) - \left(\frac{V_r}{N_r}\right) \quad (5)$$

Where V_o is the number of observed verbs in the text, N_o is the number of observed nouns in the text. V_r is the number of reference verb in the text and N_r is the number of reference noun in the text.

II. Qualitative evaluation as human evaluation

For human evaluation, two experts in SPL were incorporated to assess the quality and clinical usability of transcribed text. The two SPLs separately rated a stratified subset of model outputs (n ≈ 100 utterances per model) across five dimensions using 5-point Likert scales [37]which is evaluation in terms of Accuracy- closeness to the spoken content, Meaning Fidelity - preservation of the intended message, Grammaticality - syntactic plausibility, Readability - ease of interpretation, and Clinical Usability - usefulness for linguistic or therapeutic analysis.Inter-rater reliability was estimated using Intraclass Correlation Coefficient (ICC) to ensure consistency among evaluators.

D. Statistical and Visualization Pipeline

All computed metrics were consolidated into structured CSV files (asr_metrics_detailed.csv, human_scores.csv). Statistical analyses were performed in pandas and scipy, while visualizations were generated using matplotlib and seaborn.

V. Results

Figure 6 provides a bar graph visualization of Word Error Rates (WER) among the four ASR models: Whisper, NeMo Conformer-CTC, Wav2Vec 2.0, and SpeechBrain. Whisper achieved the lowest median WER (≈ 0.21), demonstrating strong robustness to disordered and non-fluent speech.

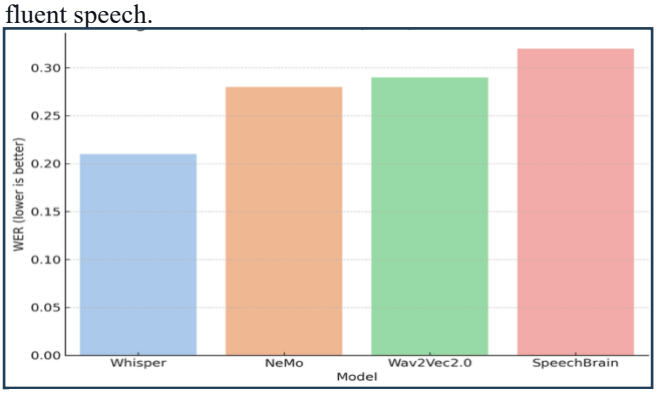

Fig 6: WER among the four ASR models.

NeMo and Wav2Vec 2.0 performed comparably (median WER ≈ 0.27–0.30), displaying the superiority of Transformer-based and self-supervised encoders in speech recognition models and revealing limitations when decoding incomplete or semantically reduced utterances. SpeechBrain exhibited a greater variability across the video samples, likely due to its sensitivity to truncated lexical items frequently found in aphasic speech. Overall, Whisper demonstrated the best generalization capability, while Wav2Vec 2.0 showed promise with a tailored domain-specific fine-tuning.

The Verb/Noun Ratio (VNR) in Figure 7 revealed how the four models reproduced lexical category distributions. As shown in Figure 7, all systems under-represented verbs relative to nouns, as in the plot, the VNR expected is less than the natural ratio ≈1.5, echoing prior psycholinguistic findings that verbs are more acoustically and semantically variable in aphasic speech.

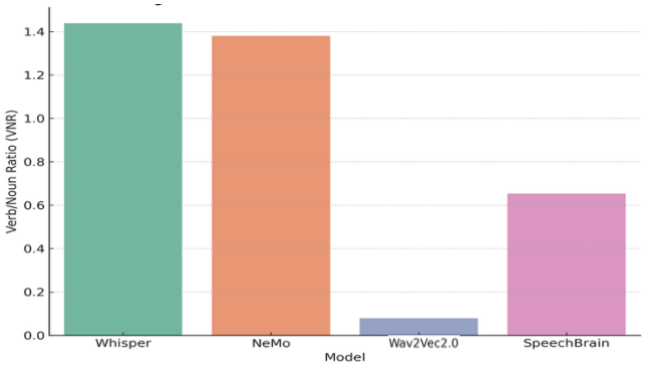

Fig 7: VNR among the four ASR models.

Whisper preserved verb–noun balance most effectively (ΔVNR ≈ 0.07), while NeMo exhibited the most significant deviation (ΔVNR ≈ 0.15). This imbalance indicates that ASR systems tend to drop or substitute verbs, resulting in transcripts that maintain nominal content but lose propositional structure. The analysis of samples with VNR thus serves as a clinically meaningful metric, complementing WER by capturing lexical asymmetry in disordered speech.

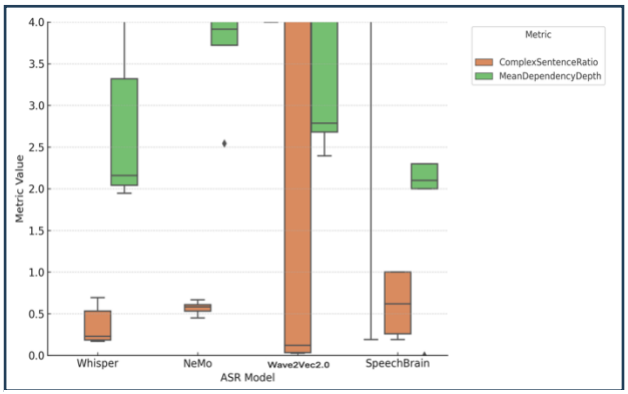

Fig 8:CSR and MDD among the four ASR models.

Syntactic measures evaluated by Complex Sentence Ratio (CSR), and Mean Dependency Depth (MDD) provided insight into the linguistic preservation among the models as in Figure 8. Whisper and SpeechBrain shoed lowest CSR and MDD indicating a linear sentence structure, Nemo maintained moderate complexity and Hugging face exhibited a higher MDD but nearly zero CSR. Overall the models produced shallower dependency structure. These results suggest that, even when word recognition is adequate, ASR models fail to reconstruct the deeper grammatical organization characteristic of natural speech, an essential limitation for clinical and linguistic analyses. The Correlation Analysis in Figure 9 demonstrates strong relationships between error and linguistic characteristics that describe how speech is evaluated for the choice made, including its words, grammar, structure, and meaning.WER exhibited a strong negative correlation with TTR (r = –0.74) and Verb Count (r = –0.69), indicating that models that make fewer word recognition errors also tend to produce transcripts that sound more natural, with a broader range of vocabulary, and with more accurate sentence grammar. Moderate positive correlations between TTR and CSR (r = 0.53) suggest that lexically diverse transcripts also tend to maintain syntactic complexity, reinforcing the interplay between vocabulary richness and sentence structure.

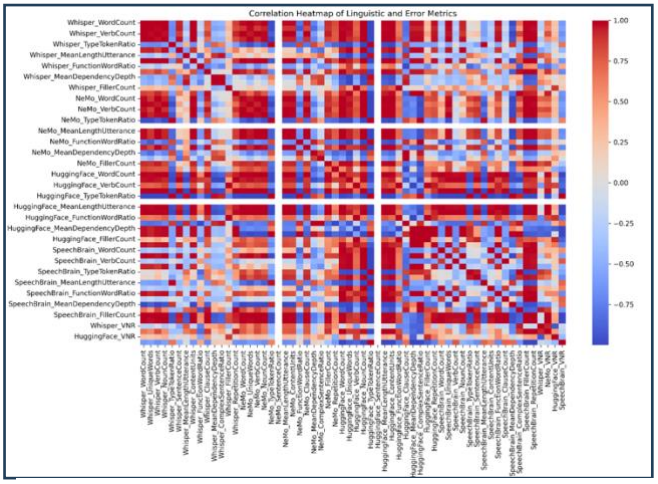

Fig 9: Correlation heat map

The two human experts evaluated the model using Likert scale (1-5), 1 being the best and 5 being the worst. Table 1 gives the average value among the criterion's used. Whisper and NeMo produced the most readable and accurate transcripts, aligning with their lower WER and stable syntactic performance. Wave2vec showed an average performance with variability between grammar and readability of transcripts. SpeechBrain**,** while slightly less favored numerically, exhibited adaptability in particularly word retrieval errors suggesting potential for fine-tuning with disorder-specific data. The qualitative notes from raters shows that readability and grammatical integrity are crucial for clinical usability, underscoring that surface accuracy alone does not guarantee interpretability in rehabilitation contexts**.** To ensure agreement between judgments, we computed the ICC. The overall ICC is 0.75, indicating good inter-rater reliability between evaluators.

Table 1: Human evaluation result in Likert scale metric

Evaluation Metric	Whisper	NeMo	Wav2Vec 2.0	Speech Brain
Accuracy	**2.5**	**3.5**	3	3.5
Meaning Fidelity	**2.5**	3.5	2.5	3.5
Gramma-tically	1	**3**	1.5	4.5
Readability	2.5	4	**4.5**	**4.5**
Usability	**3**	**4**	4.5	4.5
Average Score	**2.5**	**3.6**	3.2	4.1

VI. Conclusion

This study is a comprehensive benchmarking pipeline for ASR of aphasic speech, integrating computational and clinical interpretability. By Whisper, NeMo Conformer, Wav2Vec 2.0, and SpeechBrain on the Marshall Aphasia Corpus, we systematically assessed their performance across conventional and psycholinguistic inspired metrics. The results indicate that Whisper and Wav2Vec 2.0 maintain the most consistent balance between transcription accuracy and language structure. In contrast, NeMo and SpeechBrain show greater variability and are more

sensitive to disfluent or fragmented speech. Beyond surface-level error rates, the inclusion of syntactic, lexical, and fluency measures (e.g., Verb/Noun Ratio, Complex Sentence Ratio, and Mean Dependency Depth) highlighted the models' varying capabilities to preserve deeper linguistic structure, which is an essential factor when it comes to clinical speech applications.

Future research will extend this framework in three directions: Fine-Tuning with Aphasia-Specific Data, not limited to Marshall Corpus. We will also differentiate models by their performance, specifically tailored for different levels of severity in aphasia. We will also implement transfer learning to adapt pretrained ASR models to aphasic and neurologically disordered speech, improving sensitivity to speech irregularities while maintaining generalization. Even though the Marshall corpus contains video, we extracted the audio and transcribed it to text. In the future, we will focus on multimodality by incorporating gestures to aid transcription and improve rehabilitation by providing tailored assistance to SPL. This work establishes a reproducible, open-source foundation for benchmarking ASR systems on disordered speech.

ACKNOWLEDGMENT

This paper is dedicated in loving memory of late Dr. Soone Ae Chun. We thank Nicholas Giunta for providing human expert evaluation of the transcripts.

REFERNCES

[1]L. Worrall, S. Sherratt, P. Rogers, T. Howe, D. Hersh, A. Ferguson *et al.*, "What people with aphasia want: Their goals according to the ICF," *Aphasiology,* vol. 25, pp. 309 - 322, 2011.

[2]S. G. H. Dalton, H. I. Hubbard, and J. D. Richardson, "Moving Toward Non-transcription based Discourse Analysis in Stable and Progressive Aphasia," (in eng), *Semin Speech Lang,* vol. 41, no. 1, pp. 32-44, Jan 2020, doi: 10.1055/s-0039-3400990.

[3]B. C. Stark, M. Dutta, L. L. Murray, D. Fromm, L. Bryant, T. G. Harmon *et al.*, "Spoken Discourse Assessment and Analysis in Aphasia: An International Survey of Current Practices," (in eng), *J Speech Lang Hear Res,* vol. 64, no. 11, pp. 4366-4389, Nov 8 2021, doi: 10.1044/2021_jslhr-20-00708.

[4]Z. Brahmi, M. Mahyoob, M. Al-Sarem, J. Algaraady, K. Bousselmi, and A. Alblwi, "Exploring the Role of Machine Learning in Diagnosing and Treating Speech Disorders: A Systematic Literature Review," (in eng), *Psychol Res Behav Manag,* vol. 17, pp. 2205-2232, 2024, doi: 10.2147/prbm.S460283.

[5]S. E. Mikail. "How to Build a Reliable Ecommerce Al Agent with Stripe, LangChain, OpenAl, and Galileo." https://galileo.ai/blog/how-to-build-a-reliable-stripe-ai-agent-with-langchain-openai-and-galileo (accessed.

[6]"National aphasia association." https://aphasia.org/statistics/ (accessed October, 2025).

[7]M. Pratap, *Impact of Bilingualism on the Severity of Aphasia*. 2024.

[8]"APHASIA:Be in the Know." https://aphasia.org/wp-content/uploads/2025/06/Aphasia-Be-in-the-Know_2022_v11_Design-19937-AC.pdf (accessed 2022).

[9]H. El Hachioui, H. Lingsma, M. Van de Sandt-Koenderman, D. Dippel, P. Koudstaal, and E. Visch-Brink, "Long-term prognosis of aphasia after stroke," *Journal of neurology, neurosurgery, and psychiatry,* vol. 84, 10/31 2012, doi: 10.1136/jnnp-2012-302596.

[10]C. Bao, C. Huo, Q. Chen, and C. Gao, *AS-ASR: A Lightweight Framework for Aphasia-Specific Automatic Speech Recognition*. 2025.

[11]Z. Sasindran, H. Yelchuri, T. V. Prabhakar, and S. Rao, "HEVAL: A New Hybrid Evaluation Metric for Automatic Speech Recognition Tasks," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 16-20 Dec. 2023 2023, pp. 1-7, doi: 10.1109/ASRU57964.2023.10389717.

[12]B. Phukon, X. Zheng, and M. Hasegawa-Johnson, "Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches," *arXiv preprint arXiv:2506.16528,* 2025.

[13]N. M. Kollapally. "Github." https://github.com/navya777/AphasiaASR (accessed 2025).

[14]J. Wade, B. Petheram, and R. Cain, "Voice recognition and aphasia: can computers understand aphasic speech?," (in eng), *Disabil Rehabil,* vol. 23, no. 14, pp. 604-13, Sep 20 2001, doi: 10.1080/09638280110044932.

[15]D. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, C. Doogan *et al.*, "NUVA: A Naming Utterance Verifier for Aphasia Treatment," *Computer Speech & Language,* vol. 69, p. 101221, 03/01 2021, doi: 10.1016/j.csl.2021.101221.

[16]J. Schuchard and E. L. Middleton, "Word repetition and retrieval practice effects in aphasia: Evidence for use-dependent learning in lexical access," (in eng), *Cogn Neuropsychol,* vol. 35, no. 5-6, pp. 271-287, Jul-Sep 2018, doi: 10.1080/02643294.2018.1461615.

[17]D. Le, Provost, E.M, " Improving Automatic Recognition of Aphasic Speech with AphasiaBank," *Proc. Interspeech,* pp. 2681-2685, 2016, doi: 10.21437/Interspeech.2016-213.

[18]M. M. Forbes, D. Fromm, and B. Macwhinney, "AphasiaBank: a resource for clinicians," (in eng), *Semin Speech Lang,* vol. 33, no. 3, pp. 217-22, Aug 2012, doi: 10.1055/s-0032-1320041.

[19]H. M. Clark, R. L. Utianski, J. R. Duffy, E. A. Strand, H. Botha, K. A. Josephs *et al.*, "Western Aphasia Battery-Revised Profiles in Primary Progressive Aphasia and Primary Progressive Apraxia of Speech," (in eng), *Am J Speech Lang Pathol,* vol. 29, no. 1s, pp. 498-510, Feb 21 2020, doi: 10.1044/2019_ajslp-cac48-18-0217.

[20]M. Perez, Z. Aldeneh, and E. Mower Provost, *Aphasic Speech Recognition Using a Mixture of Speech Intelligibility Experts*. 2020, pp. 4986-4990.

[21]R. S. W. Alyahya, A. D. Halai, P. Conroy, and M. A. Lambon Ralph, "Noun and verb processing in aphasia: Behavioural profiles and neural correlates," *NeuroImage: Clinical,* vol. 18, pp. 215-230, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.nicl.2018.01.023.

[22]G. Chatzoudis, M. Plitsis, S. Stamouli, A.-L. Dimou, N. Katsamanis, and V. Katsouros, *Zero-Shot Cross-lingual Aphasia Detection using Automatic Speech Recognition*. 2022, pp. 2178-2182.

[23]G. Sanguedolce, S. Brook, D. Gruia, P. Naylor, and F. Geranmayeh, *When Whisper Listens to Aphasia: Advancing Robust Post-Stroke Speech Recognition*. 2024, pp. 1995-1999.

[24]B. Macwhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for Studying Discourse," (in eng), *Aphasiology,* vol. 25, no. 11, pp. 1286-1307, 2011, doi: 10.1080/02687038.2011.589893.

[25]A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, 2023: PMLR, pp. 28492-28518.

[26]O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577,* 2019.

[27]A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," presented at the Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2020.

[28]M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624,* 2021.

[29]A. Ali and S. Renals, *Word Error Rate Estimation for Speech Recognition: e-WER*. 2018, pp. 20-24.

[30]A. Waheed, H. Atwany, R. Singh, and B. Raj, *On the Robust Approximation of ASR Metrics*. 2025, pp. 23119-23146.

[31]M. Chen and K. Zechner, "Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech," presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Portland, Oregon, 2011.

[32]M. Polinsky and L. Magyar, "Headedness and the Lexicon: The Case of Verb-to-Noun Ratios," *Languages,* vol. 5, no. 1, p. 9, 2020. [Online]. Available: https://www.mdpi.com/2226-471X/5/1/9.

[33]V. De Fino, L. Fontan, J. Pinquier, I. Ferrané, and S. Detey, *Prediction of L2 speech proficiency based on multi-level linguistic features*. 2022, pp. 4043-4047.

[34]H. Liu, "Dependency Distance as a Metric of Language Comprehension Difficulty," *Journal of Cognitive Science,* vol. 9, pp. 159-191, 09/20 2008, doi: 10.17791/jcs.2008.9.2.159.

[35]B. Macwhinney, S. Bird, C. Cieri, and C. Martell, *TalkBank: Building an open unified multimodal database of communicative interaction*. 2004.

[36]Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, *Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers*. 2023, pp. 2798-2802.

[37]A. T. Jebb, V. Ng, and L. Tay, "A Review of Key Likert Scale Development Advances: 1995-2019," (in eng), *Front Psychol,* vol. 12, p. 637547, 2021, doi: 10.3389/fpsyg.2021.637547.