

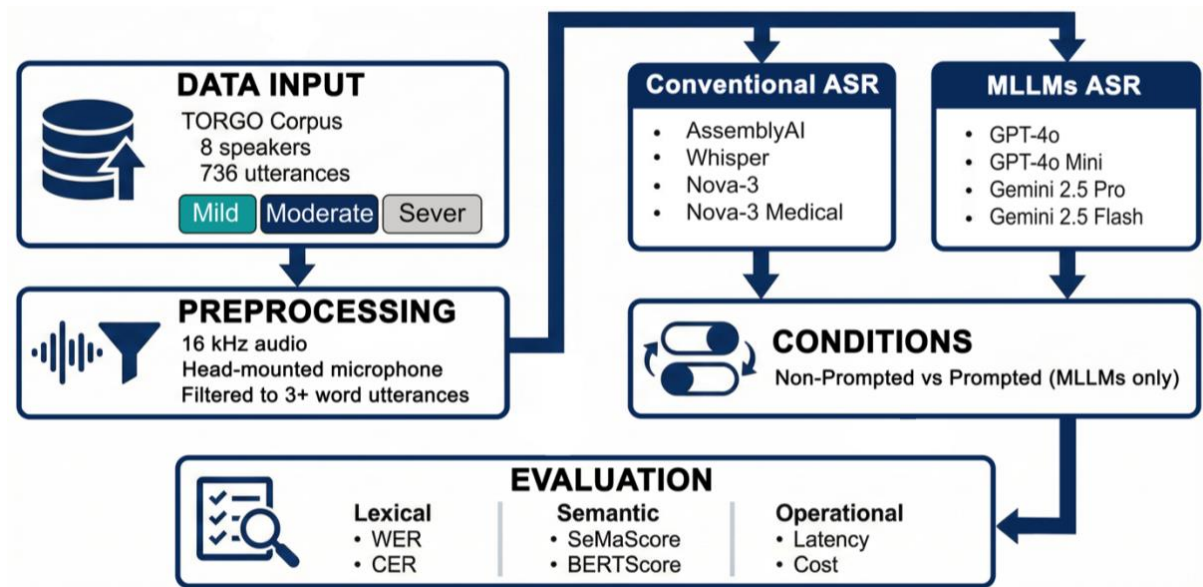
Zero-Shot Recognition of Dysarthric Speech Using Commercial Automatic Speech Recognition and Multimodal Large Language Models

Ali Alsayegh and Tariq Masood

Department of Design, Manufacturing and Engineering Management

University of Strathclyde, Glasgow G1 1XQ United Kingdom

Pre-Print



Highlight:

- Conducts the first systematic zero-shot evaluation of commercial ASR and multimodal large language models on dysarthric speech across severity levels
- Establishes severity-stratified performance baselines for eight commercial speech recognition services using the TORGO dysarthric speech corpus
- Reveals architecture-specific prompting effects where GPT-4o improves by 7.36 percentage points whilst Gemini variants exhibit degradation
- Demonstrates that semantic preservation metrics indicate partial recovery of communicative intent despite elevated lexical error rates
- Provides practical deployment guidance for assistive technology practitioners selecting speech recognition systems for motor speech disorders

Zero-Shot Recognition of Dysarthric Speech Using Commercial Automatic Speech Recognition and Multimodal Large Language Models

Keywords

Speech recognition; Multimodal large language models; Human-machine interaction; Zero-shot learning; Assistive intelligent systems; Dysarthria

Abstract

Voice-based human-machine interaction is a primary modality for accessing intelligent systems, yet individuals with dysarthria face systematic exclusion due to recognition performance gaps. Whilst automatic speech recognition (ASR) achieves word error rates (WER) below 5% on typical speech, performance degrades dramatically for dysarthric speakers. Multimodal large language models (MLLMs) offer potential for leveraging contextual reasoning to compensate for acoustic degradation, yet their zero-shot capabilities remain uncharacterised. This study evaluates eight commercial speech-to-text services on the TORGO dysarthric speech corpus: four conventional ASR systems (AssemblyAI, Whisper large-v3, Deepgram Nova-3, Nova-3 Medical) and four MLLM-based systems (GPT-4o, GPT-4o Mini, Gemini 2.5 Pro, Gemini 2.5 Flash). Evaluation encompasses lexical accuracy, semantic preservation, and cost-latency trade-offs. Results demonstrate severity-dependent degradation: mild dysarthria achieves 3-5% WER approaching typical-speech benchmarks, whilst severe dysarthria exceeds 49% WER across all systems. A verbatim-transcription prompt yields architecture-specific effects: GPT-4o achieves 7.36 percentage point WER reduction with consistent improvement across all tested speakers, whilst Gemini variants exhibit degradation. Semantic metrics indicate that communicative intent remains partially recoverable despite elevated lexical error rates. These findings establish empirical baselines enabling evidence-based technology selection for assistive voice interface deployment.

1 Introduction

The principle of universal access posits that information society technologies should be accessible and usable by all citizens, including individuals with disabilities, without requiring specialised adaptation or a posteriori modification [1]. Voice-based interfaces have proliferated as a primary interaction modality, embedded within smart home devices, mobile assistants, automotive systems, and augmentative and alternative communication (AAC) platforms. For many users with motor impairments, voice control represents not merely a convenience but the most viable, or sole, means of interacting with digital technologies [2]. Yet the recognition accuracy that enables seamless voice interaction for typical speakers systematically fails for individuals with dysarthria, a motor speech disorder characterised by imprecise articulation, atypical prosody, and variable speech rate arising from neurological conditions including cerebral palsy, Parkinson's disease, amyotrophic lateral sclerosis, and stroke sequelae [3]. This recognition disparity creates a paradox wherein those who stand to benefit most from voice-mediated accessibility are precisely those for whom such interfaces remain functionally inaccessible.

The magnitude of this accessibility barrier is substantial. Dysarthria affects an estimated 170 per 100,000 individuals in the general population, with prevalence increasing to 70 to 100% among those with specific neurological conditions [3]. Contemporary automatic speech

recognition (ASR) systems routinely achieve word error rates (WER) below 5% on typical speech [4], yet performance degrades precipitously for dysarthric speakers, with severe cases frequently exceeding 60% WER [5], effectively rendering voice interfaces non-functional for communicating user intent. From the perspective of design for all, this performance disparity represents a significant challenge to inclusive design principles: commercial voice technologies are implicitly designed for a narrow band of typical speech characteristics, thereby excluding users whose speech patterns deviate from this assumed norm [6]. The consequences extend beyond individual inconvenience to systematic digital exclusion, reduced participation in the information society, and diminished autonomy for individuals who may already face communication barriers in their daily lives.

Critically, the deployment context for accessible speech recognition differs fundamentally from research settings. In practice, commercial speech-to-text services are deployed for clinical and assistive applications in zero-shot configurations: collecting speaker-specific adaptation data is frequently impractical, burdensome for users, or technically infeasible, and systems must function out of the box with no prior exposure to the individual's speech patterns [7]. This constraint reflects the reality faced by assistive technology practitioners and AAC developers who must select and configure recognition engines for diverse end-users. Moreover, evaluation of accessibility-relevant performance must extend beyond conventional lexical error metrics. For assistive applications, particularly AAC message generation, voice-controlled environmental systems, and telecare interfaces, semantic preservation may matter more than verbatim accuracy: a transcription that captures user intent despite lexical substitutions serves communicative function better than one that fails entirely [8]. Operational considerations including cost, latency, and integration complexity further determine whether recognition technologies can be viably deployed in resource-constrained assistive contexts [9].

Recent advances in multimodal large language models (MLLMs) have introduced architectures that integrate acoustic processing with powerful language modelling capabilities [10]. These systems may offer advantages for dysarthric speech by leveraging contextual inference to compensate for degraded acoustic evidence, effectively using semantic plausibility to resolve phonetic ambiguity. However, whether such contextual reasoning translates to improved recognition of acoustically degraded speech remains an empirical question. Initial evidence suggests MLLMs can demonstrate robustness to accented and noisy speech [11, 12], though their effectiveness for dysarthric conditions remains largely unexplored. The role of prompting, a technique widely used in text-based applications to shape model behaviour at inference, has not been systematically examined for speech transcription tasks involving atypical input.

To address these gaps, we present a comprehensive zero-shot evaluation of commercial speech recognition services on the TORGO dysarthric speech corpus [13]. Our study makes three primary contributions to the universal access literature. First, we establish severity-stratified baselines for eight commercial services across lexical accuracy, semantic fidelity, and operational efficiency dimensions, providing the empirical evidence base necessary for informed technology selection in assistive contexts. To our knowledge, this represents the first systematic comparison combining conventional ASR and MLLM-based services on dysarthric speech under zero-shot conditions with concurrent lexical, semantic, and operational evaluation. *Second*, we characterise the differential effects of prompting on MLLM-based transcription, revealing architecture-specific responses that have direct implications for deployment configuration. *Third*, we frame our findings within the broader discourse on inclusive voice interface design, articulating how recognition capabilities and limitations translate into concrete accessibility barriers and opportunities for users with dysarthria.

2 Theoretical background and related work

2.1 Universal access and inclusive voice interface design

Universal access, as conceptualised within human-computer interaction research, encompasses the systematic effort to develop information technologies that are accessible, usable, and acceptable to the widest possible range of users across diverse abilities, contexts, and technological platforms [1, 14]. This paradigm extends beyond retrofitting accessibility features onto existing systems, characterised as reactive or accommodating approaches, to advocate for proactive design for all methodologies wherein diversity is treated as a fundamental design parameter from inception [15]. Voice interfaces present particular challenges for universal access: whilst they eliminate barriers associated with visual or manual interaction for some user groups, they simultaneously erect barriers for individuals whose speech characteristics deviate from the typical patterns on which recognition systems are trained.

The accessibility of voice interfaces thus exemplifies what Stephanidis and Savidis [14] term the interaction design challenge of universal access: developing systems that function effectively across the full spectrum of user diversity without requiring users to adapt to system limitations. For dysarthric speakers, current voice technologies fail this challenge: they demand either that users modify their speech to match system expectations (an impossibility for many dysarthric individuals) or that they abandon voice interaction entirely in favour of alternative modalities. This forced modality switching contradicts user-centred design principles and may impose cognitive and physical burdens that compound existing communication challenges [16]. The present study therefore addresses a critical gap in universal access research: characterising the current state of commercial speech recognition accessibility for dysarthric users and identifying pathways toward more inclusive voice interface design.

2.2 Dysarthria and the ASR accessibility gap

Dysarthria encompasses a heterogeneous group of motor speech disorders resulting from disruption to the neuromuscular control of speech production [3]. Acoustic manifestations include imprecise consonant articulation, vowel distortion, abnormal prosody (stress, intonation, and rhythm), hypernasality, and variable speech rate, characteristics that fundamentally violate the acoustic-phonetic assumptions underlying conventional ASR architectures [17]. Studies consistently report substantial recognition performance gaps: whilst state-of-the-art systems achieve WER below 5% on typical speech, dysarthric speech recognition frequently exceeds 30% WER for moderate impairment and 60% WER for severe cases [5, 18, 19]. The heterogeneous aetiologies of dysarthria, spanning spastic, flaccid, ataxic, hypokinetic, hyperkinetic, and mixed presentations, further complicate recognition, as acoustic distortions differ qualitatively across subtypes [20].

Prior research has explored various approaches to bridging this accessibility gap. Speaker-adaptive training using techniques such as Maximum Likelihood Linear Regression (MLLR) has demonstrated relative WER reductions of approximately 37% [21], whilst speaker-dependent systems can achieve further improvements when sufficient adaptation data are available [22]. Self-supervised learning approaches including wav2vec 2.0 and HuBERT, pre-trained on large typical speech corpora, have shown promise for improving performance with limited dysarthric training data [23]. However, these approaches share a common limitation from a universal access perspective: they require speaker-specific data collection and model adaptation, placing burden on end-users and limiting scalability for real-world assistive

deployment where immediate, zero-shot functionality is essential [24]. The present study addresses this limitation by evaluating systems under realistic zero-shot constraints.

2.3 Multimodal language models and context-aware transcription

The emergence of multimodal large language models represents an architectural evolution from purely acoustic-phonetic decoding toward context-aware transcription [10]. By integrating speech processing with powerful language modelling and extensive world knowledge, MLLMs may leverage semantic plausibility to disambiguate acoustically degraded input, potentially compensating for the phonetic confusion that characterises dysarthric speech. However, it remains unclear whether these systems' contextual reasoning capabilities actually improve recognition of acoustically degraded speech, or whether they primarily benefit from linguistic context in typical speech scenarios. Early evidence suggests such models demonstrate enhanced robustness to accented, noisy, and disfluent speech compared to conventional ASR [11, 12], though systematic evaluation on dysarthric speech remains limited.

Beyond speech processing, large language models have demonstrated capabilities in knowledge-intensive retrieval tasks, including extracting structured information from unstructured organisational data [25]. This capacity to bridge degraded or ambiguous input and coherent output is directly relevant to dysarthric speech recognition, where acoustic distortions must be resolved into accurate linguistic representations.

Prompting, the provision of natural language instructions to shape model behaviour at inference, has proven effective for many text-based tasks but remains underexplored for speech transcription. For dysarthric speech specifically, prompts might instruct models to expect atypical acoustic characteristics, prioritise verbatim transcription over grammatical correction, or tolerate phonetic ambiguity. Whether such instructions can improve recognition accuracy, maintain semantic fidelity, or alter the accuracy-latency trade-off remains an empirical question with direct implications for assistive technology deployment.

2.4 Evaluation frameworks for accessible speech technology

Traditional evaluation of ASR performance relies primarily on lexical error metrics, specifically WER and character error rate (CER) [26], that quantify surface-level discrepancies between hypothesis and reference transcriptions. Whilst these metrics capture recognition accuracy in a technical sense, they inadequately represent communicative adequacy for assistive applications [8, 27]. A transcription that substitutes *medication* for *medicine* incurs the same WER penalty as one that substitutes *medication* for *celebration*, despite the former preserving semantic intent whilst the latter corrupts it entirely. For AAC applications, voice-controlled assistive devices, and telecare systems, conveying user intent matters more than verbatim accuracy.

Embedding-based semantic similarity measures including SeMaScore [8] and BERTScore [28] address this limitation by computing similarity in continuous representation space, capturing meaning preservation even when surface forms differ. These metrics were developed and validated primarily on typical speech transcription tasks; their behaviour on dysarthric speech, where both reference utterances and recognition hypotheses may contain atypical lexical patterns, has not been extensively validated. Nevertheless, recent evaluation frameworks for atypical speech recognition advocate multi-dimensional assessment combining lexical accuracy, semantic fidelity, and functional communication measures [27], and we adopt these metrics as reasonable proxies for communicative adequacy pending future user-centred validation. Operational considerations, including cost, latency, and integration complexity,

further determine deployment viability, as high-performing systems that exceed resource constraints cannot serve accessibility needs [9]. The present study adopts this multi-dimensional evaluation approach, providing comprehensive characterisation of system capabilities relevant to universal access requirements.

3 Methodology

This study employed a comparative evaluation design assessing eight commercial speech-to-text services on dysarthric speech under zero-shot conditions. The experimental pipeline (Fig. 1), implemented using the TranscribeSight benchmarking toolkit [29, 30], comprised five stages: corpus preprocessing and speaker stratification, parallel API evaluation across conventional ASR and MLLM-based systems, prompting conditions for MLLMs, multi-dimensional metric computation, and severity-stratified statistical analysis.

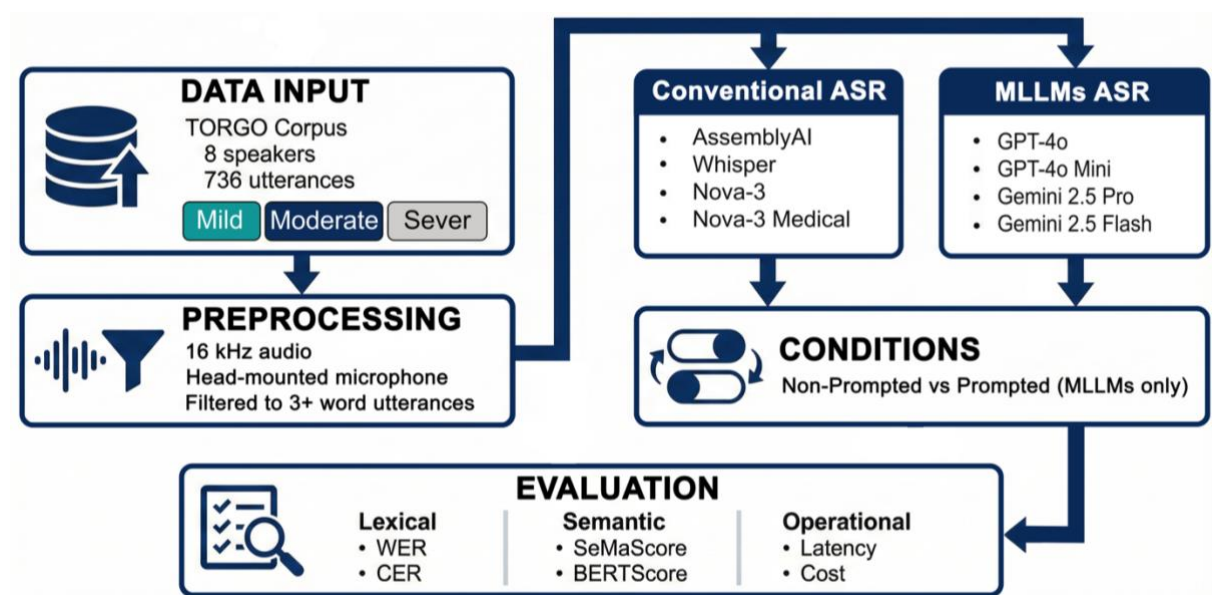


Fig. 1 Experimental pipeline for comparative evaluation of speech-to-text services on dysarthric speech, implemented using the TranscribeSight benchmarking toolkit [29, 30]. The TORGO corpus (8 speakers, 736 utterances) was pre-processed and stratified by severity, then evaluated across eight commercial services: four conventional ASR systems (AssemblyAI, Whisper, Deepgram Nova-3, Nova-3 Medical) and four MLLM-based systems (GPT-4o, GPT-4o Mini, Gemini Pro, Gemini Flash). MLLM systems were evaluated under both non-prompted and prompted conditions. Multi-dimensional metrics (lexical, semantic, operational) were computed per utterance and analysed with severity stratification

3.1 Corpus and participant characteristics

The study employed the TORGO corpus [13], a publicly available database of dysarthric and control speech designed for ASR research. TORGO was selected over alternative dysarthric speech corpora (e.g., UASpeech [31]) because its combination of connected speech utterances and head-mounted microphone recordings closely approximates realistic assistive technology deployment conditions; additionally, TORGO's established use in prior dysarthric ASR research facilitates comparison with existing literature. We prioritised evaluation depth across multiple commercial systems over breadth across corpora; the evaluation pipeline is designed for extension to additional datasets in future work. TORGO contains recordings from eight speakers with dysarthria secondary to cerebral palsy ($n = 5$) or amyotrophic lateral sclerosis ($n = 3$), alongside seven age-matched and gender-matched control speakers. Recordings span

single words, sentences, and paragraph-level utterances captured using both head-mounted and array microphones. For this study, we restricted evaluation to the dysarthric subset and stratified analyses by the corpus-provided severity classifications: mild (speakers F04, M03), moderate (speakers F03, M05), and severe (speakers F01, M01, M02, M04). These severity labels are corpus-provided rather than derived from standardised clinical assessments such as the Frenchay Dysarthria Assessment; consequently, they should be interpreted as relative categories within this dataset rather than clinically validated severity levels.

It is important to note that dysarthria secondary to cerebral palsy and ALS may manifest differently in terms of acoustic characteristics. Cerebral palsy typically produces spastic or mixed dysarthria with relatively stable speech patterns, whilst ALS-related dysarthria often involves progressive deterioration and may include flaccid components. The TORGO corpus does not enable systematic comparison across aetiologies due to sample size constraints, and our findings should be interpreted as characterising recognition performance across this mixed aetiology sample rather than generalising to specific dysarthria subtypes.

Control speaker recordings were excluded from analysis as the primary research question concerns the accessibility gap for dysarthric users rather than general ASR benchmarking. Typical-speech recognition is already well-characterised in existing benchmarks, and including controls would not inform assistive deployment decisions where the critical question is absolute performance on atypical speech rather than relative degradation from a typical baseline. Pipeline validation on typical speech is reported in our prior work [29], where we benchmarked eight ASR services on English LibriSpeech samples using default API configurations within the TranscribeSight framework; the present study focuses on the accessibility gap for dysarthric users. A three-stage inclusion protocol was applied to ensure evaluation focused on phrase-level recognition relevant to assistive applications: (i) utterances with fewer than three words were excluded to focus on connected speech; (ii) items lacking reference transcriptions were removed; and (iii) non-lexical or instructional prompts were discarded. No text normalisation beyond the original TORGO references was applied. The final dataset comprised 736 utterances distributed across severity levels (Table 1), providing sufficient statistical power for severity-stratified analysis whilst maintaining ecological validity for assistive deployment scenarios. The excluded utterances were predominantly single words and short phrases from elicitation tasks; their exclusion may limit generalisability to single-word command recognition scenarios common in some AAC applications, though such command-and-control applications typically employ constrained vocabularies and grammar-based recognition approaches that may exhibit fundamentally different recognition patterns than the open-vocabulary evaluation conducted here. Future research should specifically address short-command recognition using appropriately designed evaluation protocols.

Table 1 Distribution of TORGO speakers and utterances by dysarthria severity

Severity	Speaker	Initial utterances	Final utterances
Mild	F04	431	100
	M03	407	95
Moderate	F03	546	138
	M05	502	124
Severe	F01	117	20
	M01	371	84
	M02	389	91
	M04	391	84
Total		3,149	736

3.2 Systems evaluated

Eight commercially available speech-to-text services were evaluated, representing the current landscape of technologies potentially deployable for assistive applications. *Conventional ASR systems* included AssemblyAI (default model), OpenAI Whisper large-v3 [32] via API, Deepgram Nova-3, and Deepgram Nova-3 Medical (a variant optimised for healthcare terminology). *MLLM-based services* included OpenAI GPT-4o (gpt-4o-audio-preview), OpenAI GPT-4o Mini, Google Gemini 2.5 Pro, and Google Gemini 2.5 Flash. All systems were accessed via their public APIs during a single evaluation window (17 to 18 September 2025) to ensure temporal consistency in model versions.

Audio data were sourced exclusively from the TORGO wav_headMic recordings, which were captured using a head-mounted electret microphone at an original sampling rate of 22.1 kHz, subsequently downsampled to 16 kHz for distribution. This microphone placement provides close-talking audio with reduced ambient noise compared to the alternative array microphone recordings also available in TORGO. Audio preprocessing was minimal to reflect realistic deployment conditions: recordings were converted to 16 kHz mono WAV format where required by API specifications, with no noise reduction, normalisation, or enhancement applied. Some TORGO recordings affected by electromagnetic interference from the articulograph system were excluded from the corpus by its creators; the remaining recordings represent typical quality achievable in controlled recording environments. Critically, all systems were assessed in a zero-shot configuration: no task-specific training, fine-tuning, or speaker adaptation was performed. Conventional ASR systems were invoked with default API parameters. This evaluation design reflects realistic deployment constraints for assistive technology applications where system configuration must generalise across diverse end-users without per-user customisation.

For reproducibility, we report the specific model identifiers and API configurations used: OpenAI Whisper large-v3 (whisper-1 endpoint), GPT-4o (gpt-4o-audio-preview, September 2025 version), GPT-4o Mini (gpt-4o-mini-audio-preview), AssemblyAI (default model via v2 API), Deepgram Nova-3 and Nova-3 Medical (nova-3 and nova-3-medical models), and Google Gemini 2.5 Pro and Flash (gemini-2.5-pro-preview and gemini-2.5-flash-preview). All API calls used provider default parameters; temperature settings were not modified where configurable. This evaluation should be understood as a time-stamped benchmark representing the state of commercial speech recognition practice as of 17-18 September 2025. Results may differ with subsequent model updates, as commercial APIs are subject to periodic revision without version guarantees; practitioners should verify current performance before deployment decisions. The evaluation pipeline was implemented using TranscribeSight [29, 30], an open-source benchmarking toolkit; code and configuration files are available in the linked repository to facilitate replication and longitudinal tracking of API evolution.

3.3 Prompting protocol for MLLM systems

MLLM-based services were evaluated under two conditions: (i) a no-prompt baseline reflecting default system behaviour, and (ii) a prompted condition employing a verbatim-transcription instruction. The prompt was designed based on anecdotal observations during pilot experiments that MLLMs sometimes over-correct or paraphrase dysarthric speech output, potentially introducing errors by 'correcting' accurate but disfluent transcriptions. The instruction was designed to minimise such over-correction, paraphrase, and grammatical normalisation:

'You are tasked with transcribing English dysarthric speech. The speech input may be slurred, slow, or imprecise. Produce a verbatim transcript in English, preserving the spoken output without correction, normalisation, or substitution. In cases where words are unclear, render them as closely as possible to the audible signal rather than inferring or guessing. Output requirements: Provide only the plain transcript text. Do not include labels, headings, explanations, or formatting of any kind.'

This prompt was supplied once per request. No few-shot examples or additional system instructions were provided, and where temperature or decoding parameters were configurable, provider defaults were retained. Single API calls were made per utterance; given the large sample size ($n = 736$) and deterministic or low-temperature decoding settings used by default, run-to-run variability was expected to be minimal. We note that this represents a single prompting strategy; alternative formulations (e.g., few-shot examples, domain-specific vocabulary hints, or multi-turn correction protocols) may yield different results and warrant future investigation.

3.4 Evaluation metrics

Lexical accuracy was assessed using WER and CER, computed as the sum of substitution, deletion, and insertion errors divided by reference length. *Semantic fidelity* was evaluated using SeMaScore [8] and BERTScore [28], embedding-based measures that compute similarity between hypothesis and reference in continuous semantic space. SeMaScore was computed using the all-MiniLM-L6-v2 sentence transformer model, whilst BERTScore used roberta-large embeddings; both configurations represent commonly used defaults for these metrics. Following the multi-dimensional evaluation frameworks discussed in Section 2.4, we include these semantic alternatives alongside WER and CER to capture meaning preservation even when surface forms differ. Operational efficiency was characterised by cost (US dollars per input minute) and latency (API round-trip time from request initiation to response receipt). We acknowledge that latency measurements may be influenced by network conditions; all API calls were made from the same network location during the evaluation window to minimise this variability.

3.5 Statistical analysis

All evaluations were conducted using TranscribeSight [29, 30], an open-source benchmarking toolkit implementing standardised scoring protocols. Metrics were computed per utterance; severity-stratified results are reported as mean plus or minus standard deviation. To assess prompting effects, we computed per-utterance differences (prompted minus non-prompted) for WER and BERTScore. Statistical analysis focused on the severe dysarthria subset for two reasons: (i) severe dysarthria represents the most challenging and underserved use case where intervention effects would be most practically significant, and (ii) the severe subset provided the largest sample size ($n = 279$ utterances) for robust statistical inference. Given non-normal

distributions characteristic of error metrics, paired Wilcoxon signed-rank tests (two-sided) with Bonferroni correction for multiple comparisons were applied. Effect sizes are reported as r (rank-biserial correlation), with $r < 0.3$ indicating small effects, 0.3 to 0.5 medium effects, and $r > 0.5$ large effects.

4 Results

4.1 Severity-dependent performance degradation

Table 2 presents lexical and semantic metrics across severity levels under default (non-prompted) configurations. For reference, published benchmarks for these systems on typical speech corpora (e.g., LibriSpeech [33], Common Voice [34]) report WER below 5% [18], establishing an implicit baseline against which dysarthric speech performance can be interpreted. As noted in Section 3.1, severity categories are dataset-relative labels provided by the TORGO corpus rather than clinically validated classifications; however, they represent consistent intelligibility tiers within the dataset and provide a meaningful basis for stratified accessibility evaluation. The data reveal a consistent pattern of severity-dependent degradation: recognition performance deteriorates progressively from mild through moderate to severe dysarthria across all systems and metrics. This graded degradation confirms the expected relationship between speech impairment severity and recognition difficulty, and, critically from a universal access perspective, demonstrates that individuals with greater communication needs face proportionally larger accessibility barriers.

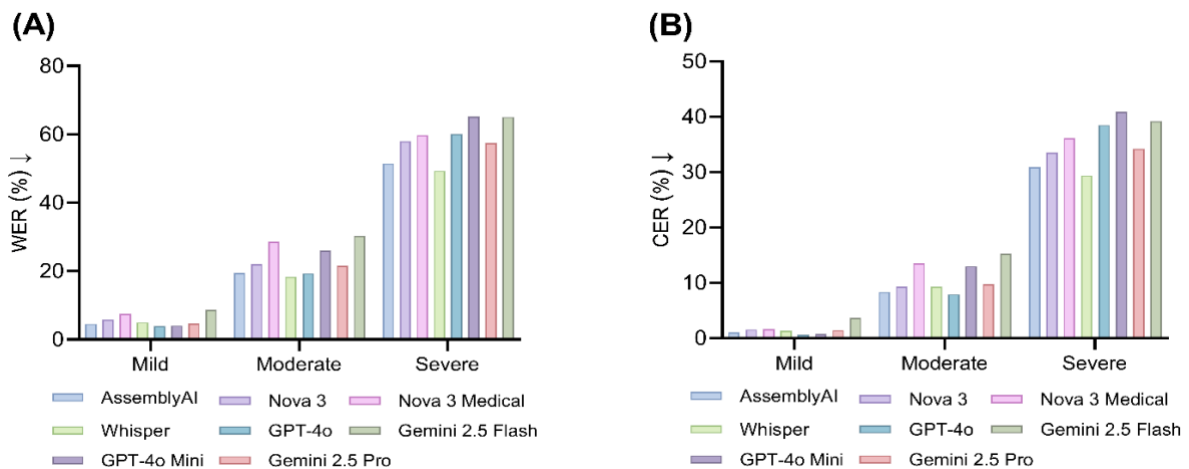


Fig. 2 Lexical accuracy across dysarthria severities under default configurations. Bars show mean WER (A) and CER (B) for conventional ASR (Deepgram Nova-3, Nova-3 Medical, AssemblyAI, Whisper) and MLLM systems (GPT-4o, GPT-4o Mini, Gemini 2.5 Pro, Gemini 2.5 Flash)

For *mild dysarthria*, system performance clustered within operationally acceptable ranges. GPT-4o achieved the lowest WER (3.87%), followed by GPT-4o Mini (4.12%) and AssemblyAI (4.58%). Among conventional ASR, Whisper demonstrated competitive performance (5.02% WER). Semantic preservation was high across systems (GPT-4o: 98.97% SeMaScore, 97.16% BERTScore), indicating that even when lexical errors occur, semantic intent is largely preserved. These results suggest that current commercial systems may provide adequate voice interface accessibility for users with mild dysarthria. It should be noted that these relatively low WER values reflect connected speech performance on utterances of three or more words; single-word command recognition, which was excluded from evaluation, may exhibit different accuracy characteristics.

For *moderate dysarthria*, performance declined substantially across all systems. Whisper achieved the lowest WER among conventional ASR (18.36%), whilst GPT-4o led among MLLMs (19.36%). Semantic metrics remained above 80% SeMaScore for leading systems, suggesting partial preservation of communicative function despite elevated lexical error rates. Notably, Deepgram Nova-3 Medical, ostensibly optimised for clinical contexts, underperformed the standard Nova-3 variant (28.57% vs. 22.12% WER). This unexpected finding may reflect the specific nature of medical vocabulary optimisation, which targets terminology rather than acoustic robustness to atypical speech; however, further investigation would be needed to draw definitive conclusions about domain-specific optimisation strategies.

For *severe dysarthria*, all systems exhibited substantial degradation, with WER exceeding 49% across the board. Whisper achieved the lowest WER (49.39%), followed by AssemblyAI (51.52%). MLLM-based systems, despite their contextual reasoning capabilities, did not outperform conventional ASR at this severity level under default configurations. Semantic preservation dropped substantially (SeMaScore range: 39 to 54%), indicating that approximately half of communicative intent is lost.

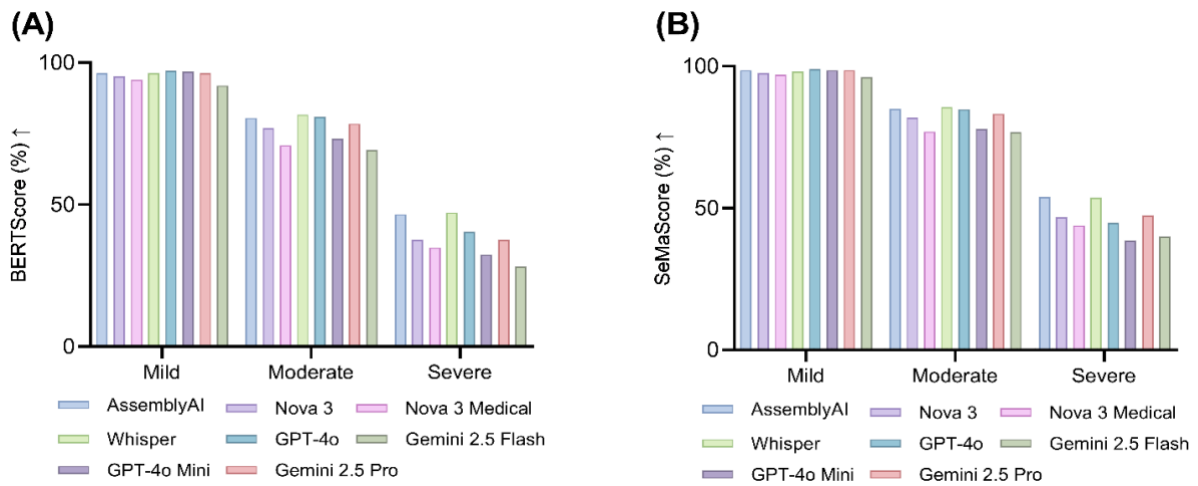


Fig. 3 Semantic preservation across dysarthria severities under default configurations. Bars show mean BERTScore (A) and SeMaScore (B) for all systems. Results degrade from mild to severe whilst preserving relative ranking between model families

Table 2 Performance of ASR and MLLM systems on TORGO under default (non-prompted) settings, reported as mean \pm SD

Severity	System	WER (%) ↓	CER (%) ↓	SeMaScore ↑	BERTScore ↑
Mild	AssemblyAI	4.58±9.42	1.03±3.04	98.64±3.79	96.43±8.30
	Whisper	5.02±10.44	1.33±3.64	98.24±5.67	96.37±8.51
	Deepgram Nova 3	5.76±10.80	1.58±3.86	97.52±9.40	95.25±11.41
	Nova 3 Medical	7.47±12.46	1.65±3.65	97.09±8.51	93.97±11.38
	GPT-4o	3.87±8.34	0.66±1.61	98.97±4.00	97.16±6.91
	GPT-4o Mini	4.12±8.45	0.76±1.85	98.64±5.48	96.94±7.01
	Gemini 2.5 Pro	4.64±9.49	1.44±3.85	98.66±3.84	96.33±8.36
	Gemini 2.5 Flash	8.73±14.63	3.66±9.66	96.15±10.63	91.99±14.65
Moderate	AssemblyAI	19.48±22.44	8.35±13.31	84.93±21.02	80.53±23.94
	Whisper	18.36±24.18	9.29±16.79	85.55±23.01	81.67±25.60
	Deepgram Nova 3	22.12±22.47	9.30±11.71	81.95±22.25	76.96±25.41
	Nova 3 Medical	28.57±26.06	13.54±20.71	76.97±27.50	70.84±29.92
	GPT-4o	19.36±28.20	7.93±13.90	84.91±25.98	81.08±29.31
	GPT-4o Mini	26.08±31.90	13.01±19.08	77.93±30.20	73.18±33.95
	Gemini 2.5 Pro	21.54±25.51	9.72±13.64	83.21±24.72	78.56±27.33
	Gemini 2.5 Flash	30.36±27.70	15.31±17.93	76.72±26.63	69.36±30.72
Severe	AssemblyAI	51.52±30.76	30.90±23.42	53.99±29.78	46.63±30.25
	Whisper	49.39±32.53	29.40±24.48	53.73±32.21	47.19±32.96

Severity	System	WER (%) ↓	CER (%) ↓	SeMaScore ↑	BERTScore ↑
	Deepgram Nova 3	58.03±26.96	33.57±20.77	46.82±28.53	37.57±28.43
	Nova 3 Medical	59.77±26.46	36.17±22.68	43.89±28.15	34.85±27.58
	GPT-4o	60.14±34.85	38.53±27.55	44.75±35.13	40.38±36.19
	GPT-4o Mini	65.26±35.04	40.91±26.38	38.66±34.37	32.38±36.17
	Gemini 2.5 Pro	57.42±30.23	34.18±21.75	47.41±32.17	37.63±33.18
	Gemini 2.5 Flash	65.09±27.57	39.21±22.30	39.95±28.91	28.19±29.27

4.2 Per-speaker analysis and macro-averaged results

To examine robustness of findings across individual speakers, Table 3 presents per-speaker WER for all systems under non-prompted conditions. This analysis reveals substantial within-severity variability, particularly for severe dysarthria. Among severe speakers, M02 achieved the lowest mean WER across systems (47.8%), whilst M04 exhibited the highest (70.1%), a difference of over 22 percentage points despite both being classified as severe. This variability likely reflects individual differences in dysarthria manifestation, speaking style, and the specific utterances recorded for each speaker.

Comparison of macro-averaged (per-speaker mean) and micro-averaged (per-utterance) results revealed generally consistent patterns, with differences typically below 2.5 percentage points. The largest divergence occurred for GPT-4o on severe dysarthria, where macro-averaging yielded 62.6% WER compared to 60.1% micro-averaged, reflecting the influence of speaker F01 who contributed only 20 utterances but exhibited high WER. These findings suggest that severity-level conclusions are robust to aggregation method, though practitioners should recognise that individual speaker performance may deviate substantially from severity-level averages.

Table 3 Per-speaker WER (%) ↓ under non-prompted conditions. Speakers grouped by severity; n indicates utterance count per speaker

Severity	Speaker	n	Asm	Whsp	DG	DGM	GPT	Mini	Pro	Flash
Mild	F04	100	4.1	4.9	6.6	8.3	3.9	3.9	4.9	8.4
	M03	95	5.1	5.2	4.9	6.6	3.9	4.4	4.4	9.1
Moderate	F03	138	13.4	14.4	14.3	17.9	14.1	19.3	16.0	24.8
	M05	124	26.7	23.1	31.5	40.5	25.7	34.2	28.2	37.1
Severe	F01	20	48.4	48.3	54.8	59.4	71.4	70.2	60.8	62.9
	M01	84	56.9	47.2	59.0	59.6	57.4	62.0	57.2	62.2
	M02	91	36.9	39.3	49.8	52.8	49.2	53.4	44.4	56.6
	M04	84	62.4	62.9	66.6	67.6	72.3	80.3	71.0	77.9

Note: Asm = AssemblyAI, Whsp = Whisper, DG = Deepgram Nova 3, DGM = Deepgram Nova 3 Medical, GPT = GPT-4o, Mini = GPT-4o Mini, Pro = Gemini 2.5 Pro, Flash = Gemini 2.5 Flash

4.3 Prompting effects on MLLM performance

The effect of the tested verbatim-transcription prompt on MLLM performance exhibited marked architecture dependence (Table 4, Fig. 4). OpenAI models demonstrated consistent improvement under prompting. GPT-4o achieved a WER reduction of 7.36 percentage points at the severe dysarthria level ($p < 0.001$, $r = 0.313$, medium effect), improving from 60.14% to 52.78%. BERTScore increased correspondingly by 6.75 percentage points ($p < 0.001$, $r = 0.319$, medium effect). GPT-4o Mini exhibited smaller but statistically significant improvements. Critically, per-speaker analysis revealed that GPT-4o improvements were consistent across all four severe dysarthria speakers: F01 improved by 16.7 percentage points, M01 by 10.3 points, M02 by 4.4 points, and M04 by 5.3 points. This 100% consistency across speakers strengthens confidence that the prompting benefit for GPT-4o is robust rather than driven by individual speaker characteristics.

In contrast, Google Gemini models exhibited adverse responses to the verbatim-transcription prompt. Gemini 2.5 Pro showed a non-significant WER increase of 1.49 percentage points ($p = 0.053$), whilst Gemini 2.5 Flash demonstrated a statistically significant WER increase of 1.92 percentage points ($p = 0.040$). Both Gemini variants showed significant semantic degradation under prompting. Per-speaker analysis revealed inconsistent effects: Gemini Pro improved for 2 of 4 severe speakers but degraded for the other 2, whilst Gemini Flash degraded for 3 of 4 speakers. We hypothesise that this divergent response may reflect architectural differences in how these systems balance acoustic evidence against linguistic inference; instructing models to avoid 'inferring or guessing' may conflict with core MLLM design principles that rely on contextual prediction. Future work should conduct prompt ablation studies (e.g., removing specific clauses such as 'rather than inferring or guessing') to isolate which instruction components trigger Gemini degradation. Notably, prompting effects also varied by severity level: for mild dysarthria, even GPT-4o showed WER increases under prompting (3.87% to 5.84%), suggesting that the verbatim-transcription instruction may be most beneficial for severe cases where over-correction by MLLMs is more problematic. This severity-dependent response warrants consideration when configuring systems for users with different impairment levels.

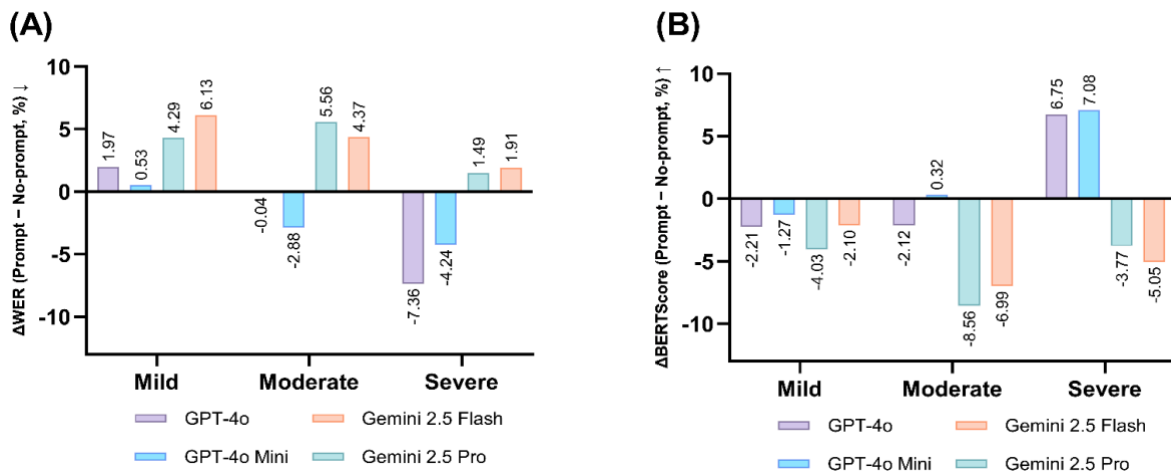


Fig. 4 Effect of prompting on MLLM performance. Differences (prompted minus non-prompted) in (A) WER (lower is better) and (B) BERTScore (higher is better) for GPT-4o, GPT-4o Mini, Gemini 2.5 Pro, and Gemini 2.5 Flash. Statistical significance testing (Table 5) was conducted on the severe dysarthria subset only

Table 4 Performance of MLLM systems on TORGO under prompted settings, reported as mean \pm SD. Prompting was not applicable to conventional ASR systems; therefore, only MLLM results are shown

Severity	System	WER (%) ↓	CER (%) ↓	SeMaScore ↑	BERTScore ↑
Mild	OpenAI GPT-4o	5.84 \pm 10.64	3.58 \pm 8.42	97.82 \pm 9.44	94.95 \pm 12.91
	OpenAI GPT-4o Mini	4.65 \pm 9.43	2.59 \pm 6.73	97.79 \pm 9.59	95.67 \pm 12.65
	Google Gemini 2.5 Pro	8.93 \pm 12.98	5.68 \pm 10.76	96.72 \pm 10.45	92.30 \pm 15.43
	Google Gemini 2.5 Flash	14.86 \pm 13.32	9.13 \pm 12.78	96.04 \pm 11.34	89.89 \pm 15.12
Moderate	OpenAI GPT-4o	19.32 \pm 24.02	10.72 \pm 13.88	83.20 \pm 26.66	78.96 \pm 29.26
	OpenAI GPT-4o Mini	23.20 \pm 28.10	7.61 \pm 12.30	78.36 \pm 30.10	73.50 \pm 32.81
	Google Gemini 2.5 Pro	27.10 \pm 25.65	13.48 \pm 14.55	76.94 \pm 28.75	70.00 \pm 32.12
	Google Gemini 2.5 Flash	34.73 \pm 25.88	11.37 \pm 13.77	72.45 \pm 28.59	62.37 \pm 33.22
Severe	OpenAI GPT-4o	52.78 \pm 32.33	30.99 \pm 22.59	53.22 \pm 32.37	47.13 \pm 33.42
	OpenAI GPT-4o Mini	61.02 \pm 32.32	37.12 \pm 25.50	45.35 \pm 33.02	39.46 \pm 33.87
	Google Gemini 2.5 Pro	58.91 \pm 27.32	32.15 \pm 20.06	46.93 \pm 30.36	33.86 \pm 31.33
	Google Gemini 2.5 Flash	67.00 \pm 24.10	38.36 \pm 20.13	38.94 \pm 26.43	23.14 \pm 27.32

Table 5 Paired Wilcoxon signed-rank test results for WER and BERTScore under prompted versus non-prompted conditions (severe dysarthria subset, $n = 279$). Effect sizes (r) and significance levels after Bonferroni correction are reported. Negative Δ WER and positive Δ BERTScore values indicate improvement

WER results

Model	Δ WER (%) ↓	Sig.	p-value	Effect (r)	Interpretation
GPT-4o	-7.36±25.24	***	<0.001	0.313	Medium
GPT-4o Mini	-4.24±21.54	**	0.0039	0.209	Small
Gemini 2.5 Pro	+1.49±17.93	†	0.0533	0.122	Small
Gemini 2.5 Flash	+1.92±16.63	*	0.0397	0.131	Small

BERTScore results

Model	Δ BERT (%) ↑	Sig.	p-value	Effect (r)	Interpretation
GPT-4o	+6.75±21.84	***	<0.001	0.319	Medium
GPT-4o Mini	+7.08±23.02	***	<0.001	0.339	Medium
Gemini 2.5 Pro	-3.77±18.65	***	<0.001	0.247	Small
Gemini 2.5 Flash	-5.05±16.29	***	<0.001	0.373	Medium

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ (not significant after Bonferroni correction). Effect size interpretation: Small ($r < 0.3$), Medium ($0.3 \leq r < 0.5$), Large ($r \geq 0.5$)

4.4 Cost and latency trade-offs

Operational characteristics varied substantially across systems (Fig. 5, Table 6). For context, human conversational studies suggest typical response latencies around 300-500 ms; delays beyond approximately 500 ms feel noticeably unnatural. In designing real-time AAC and voice interfaces, this range is often used as a target for natural interaction, though acceptable thresholds vary by application context and individual user expectations [35], though acceptable thresholds vary by application context and user expectations; asynchronous applications such as message composition can tolerate considerably longer processing times. Conventional ASR services occupied the efficient region of the cost-latency trade-off space, combining low cost with predictable processing times. Deepgram Nova-3 achieved US\$0.124 per minute at 688 seconds total corpus processing time, whilst Whisper demonstrated similar efficiency (US\$0.173 per minute, 518 seconds). AssemblyAI, despite strong accuracy performance, incurred higher cost and latency (US\$0.226 per minute, 1,741 seconds).

MLLM-based systems exhibited more variable efficiency profiles. Non-prompted Gemini variants offered the lowest absolute costs (Flash: US\$0.055 per minute, Pro: US\$0.069 per minute) but with substantially elevated latency (943 and 2,765 seconds, respectively). Prompting increased Gemini latencies further, with Pro rising to 5,470 seconds, without corresponding accuracy benefits. OpenAI models demonstrated more favourable prompting efficiency: GPT-4o latency actually decreased from 780 to 647 seconds under prompting, whilst cost increased moderately from US\$0.260 to US\$0.303 per minute. Given that typical AAC utterances are 2 to 5 seconds in duration, per-utterance costs would be approximately 3 to 8% of the per-minute figures reported, placing even premium services within practical budgets for individual use. We note that the latency estimates in Table 6 are proxy values derived from aggregate processing times; real-world latency will vary with client location, network conditions, and API load. The relative rank ordering of systems is likely more stable than the absolute millisecond values reported.

Table 6 Speaker-level estimated per-utterance latency (milliseconds). Median, interquartile range (IQR), and range computed across speaker-level estimates (n = 8 speakers)

System	Cond.	Median	IQR (Q25-Q75)	Min-Max
GPT-4o Mini	NP	1,567	1,462-1,733	1,431-2,138
Whisper	NP	2,150	2,071-2,428	1,705-2,527
Gemini 2.5 Flash	NP	2,452	1,630-3,315	1,528-15,949
Deepgram Nova 3 Medical	NP	2,480	2,210-2,780	1,980-3,520
Deepgram Nova 3	NP	2,691	2,384-3,009	2,181-3,814
GPT-4o	NP	3,197	2,913-3,383	2,816-3,990
AssemblyAI	NP	6,978	6,775-7,368	6,279-7,820
Gemini 2.5 Pro	NP	13,103	5,483-18,317	3,391-36,148
GPT-4o Mini	P	1,210	1,067-1,429	1,004-1,886
GPT-4o	P	2,319	1,962-2,993	1,797-3,402
Gemini 2.5 Flash	P	6,204	2,621-10,081	2,138-30,352
Gemini 2.5 Pro	P	17,148	7,839-26,544	4,913-39,027

Note: NP = Non-prompted, P = Prompted. Latency statistics derived from per-speaker aggregate processing times divided by utterance counts; Median and IQR reflect the distribution of these speaker-level estimates (n = 8 speakers), and Min-Max indicates the range across speaker-level means rather than individual utterance extremes. The 300-500 ms conversational latency range (see Section 4.4 for discussion) is not achieved by any evaluated system; all are better suited to asynchronous use cases where processing delays are tolerable.

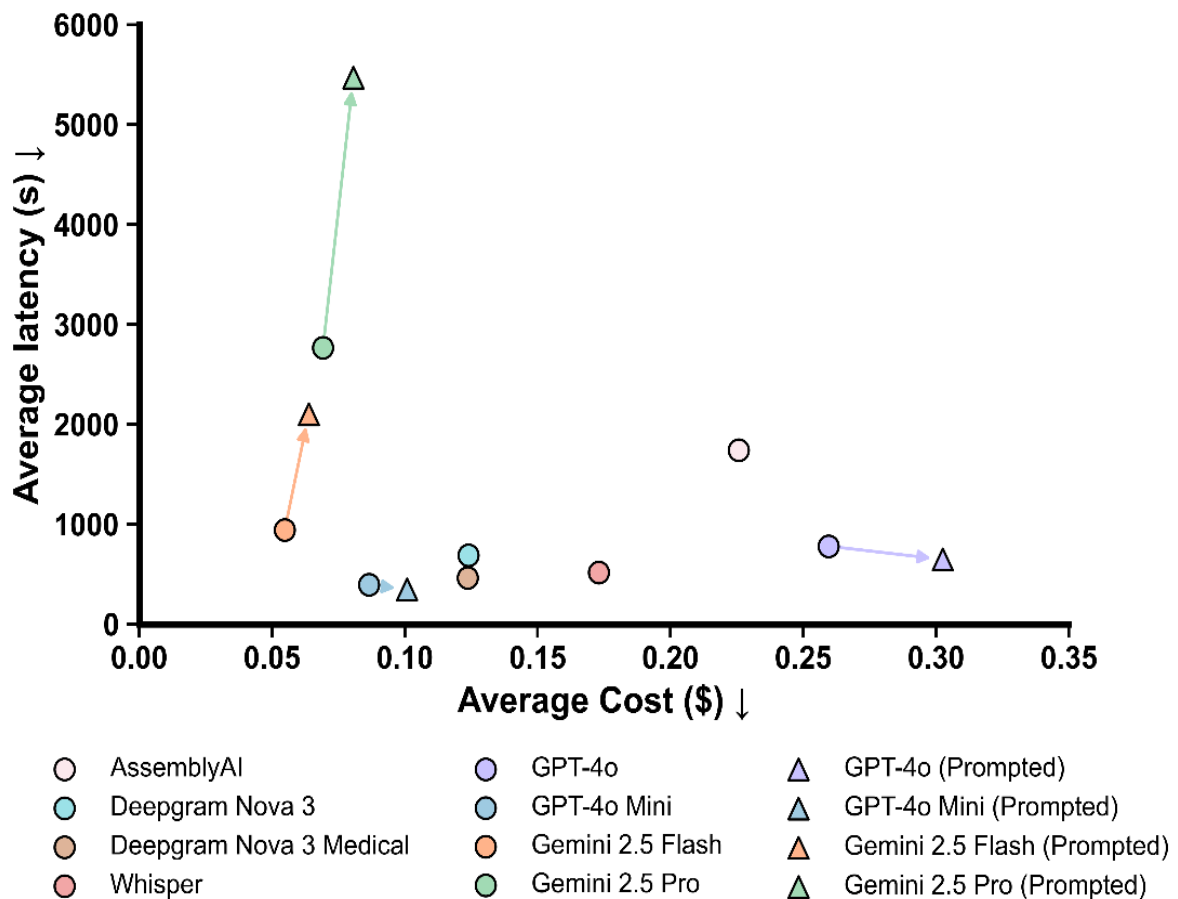


Fig. 5 Cost and latency characteristics of evaluated systems. Points show average cost (US dollars per minute) versus total corpus processing time (seconds) for non-prompted and prompted conditions

5 Discussion

5.1 Implications for universal access and inclusive design

The findings illuminate both the current state and future directions for inclusive voice interface design. The severity-dependent performance gradient confirms that commercial speech recognition systems, designed primarily for typical speech, systematically exclude users whose speech characteristics fall outside assumed norms. This exclusion is not merely a technical limitation but a manifestation of design choices that privilege acoustic homogeneity over diversity. From the perspective of design for all [6, 15], current voice technologies demonstrate limited success in proactively accommodating the full range of human speech variation.

However, the results also identify pathways toward more inclusive design. For users with mild dysarthria, multiple systems achieved recognition accuracy approaching typical-speech benchmarks, suggesting that voice interfaces can already serve this population with minimal accommodation. For moderate dysarthria, semantic preservation metrics indicate that user intent may be recoverable even when lexical accuracy degrades, supporting interface designs that incorporate confirmation mechanisms, semantic disambiguation, or intelligent error correction. For severe dysarthria, the substantial performance gap (exceeding 49% WER across all systems) underscores the continuing need for alternative and multimodal interaction strategies, with voice as one input among several rather than a sole modality.

The architecture-specific prompting effects observed in this study have direct implications for assistive technology practice. For practitioners deploying OpenAI GPT-4o in accessible voice interfaces, the verbatim-transcription prompt offers a low-cost, easily implementable intervention that yields statistically significant improvements without requiring technical expertise or model fine-tuning. per-speaker analysis in Section 4.3 shows GPT-4o improvements were consistent across all severe speakers, strengthening its practical relevance. For Gemini-based systems, default configurations outperform prompted variants, and practitioners should avoid applying prompting strategies that degrade rather than enhance accessibility. This divergent response may reflect architectural differences in how these systems balance acoustic evidence against linguistic inference; instructing models to avoid 'inferring or guessing' may conflict with core MLLM design principles that rely on contextual prediction, potentially explaining the degradation observed in Gemini models. This finding exemplifies a broader principle: accessibility interventions cannot be assumed to generalise across technological platforms and must be validated empirically.

5.2 Practical and design implications for assistive deployment

Drawing on our empirical findings, we offer consolidated guidance for practitioners selecting and configuring speech recognition for assistive applications. This section integrates deployment considerations (application type, cost, latency) with severity-stratified design recommendations aligned with design for all principles.

For users with mild dysarthria, current commercial systems can support voice-primary interfaces with minimal accommodation. Recognition accuracy (3-5% WER) approaches that achieved on typical speech. Recommended configuration: GPT-4o or Whisper in non-prompted mode. Interface design can follow typical voice assistant patterns with standard confirmation dialogs. Semantic similarity thresholds for intent confirmation can be set relatively high (>90% BERTScore) given strong preservation of meaning.

For users with moderate dysarthria, voice interfaces should incorporate explicit confirmation mechanisms driven by semantic similarity. Recommended configuration: GPT-4o with verbatim-transcription prompt for accuracy-critical applications, or Whisper for latency-critical scenarios. Interface design should present transcription results with confidence indicators and offer one-tap correction options. When semantic similarity exceeds 80% (BERTScore), systems can proceed with moderate confidence; below this threshold, explicit user confirmation should be required. These thresholds represent initial design heuristics derived from the present results and should be validated with end-users. Consider offering alternative input modalities (on-screen keyboard, predictive text) as readily accessible fallbacks.

For users with severe dysarthria, voice should function as one component within a multimodal interaction strategy rather than a standalone modality. Recommended configuration: prompted GPT-4o for best accuracy, accepting latency trade-offs. Interface design should implement: (i) multimodal input combining voice with touch, switch scanning, or eye gaze; (ii) partner-assisted confirmation workflows where a communication partner can verify intent; (iii) context-aware intent inference using conversation history and situational cues to disambiguate uncertain transcriptions; and (iv) progressive disclosure of correction options, from simple accept/reject to detailed word-level editing. Semantic similarity thresholds should be lowered (>60% BERTScore) to avoid excessive false rejections, accepting that some errors will require manual correction. As with moderate dysarthria, these threshold values are empirical heuristics that warrant validation in deployment contexts.

Application-type considerations: For real-time, interactive applications (AAC conversational interfaces, voice-controlled environmental systems, real-time captioning), none of the evaluated systems meet the conversational latency targets discussed in Section 4.4 (300-500 ms). Among available options, GPT-4o Mini offers the best latency profile (median 1,567 ms non-prompted, 1,210 ms prompted), followed by Whisper (median 2,150 ms). For asynchronous, intent-critical applications (message composition, email preparation, document dictation), prompted GPT-4o offers superior semantic preservation when processing time permits. For resource-constrained deployments, Gemini Flash offers the lowest cost, though accuracy limitations restrict suitability to non-critical applications with mild to moderate dysarthria users.

Engine selection decision flow: For real-time applications requiring latency under 2 seconds, use Whisper or prompted GPT-4o Mini. For asynchronous applications prioritising accuracy, use prompted GPT-4o. For cost-constrained deployments serving mild-to-moderate users, Gemini Flash offers acceptable performance at lowest cost. Avoid prompted Gemini variants regardless of use case. When user severity is unknown at design time, implement adaptive interfaces that adjust confirmation thresholds and fallback options based on observed recognition confidence over time. These findings are most directly applicable to open-vocabulary dictation and phrase-level AAC scenarios; generalisability to short-command smart-home control or constrained-vocabulary applications is limited given the exclusion of single-word utterances from evaluation.

5.3 Limitations and future directions

Several limitations constrain the generalisability of these findings and highlight directions for future research. *Corpus limitations*: The TORGO corpus, whilst established as a dysarthric speech benchmark, comprises a relatively small participant sample ($n = 8$ dysarthric speakers) speaking a single language variety (North American English). The severity classifications are

corpus-provided rather than derived from standardised clinical assessments, limiting comparability with clinical severity scales used in practice. The mixed aetiology sample (cerebral palsy and ALS) may obscure aetiology-specific effects, as these conditions produce qualitatively different speech impairments. Per-speaker analysis revealed substantial within-severity variability (over 22 percentage points WER difference between M02 and M04, both classified as severe), underscoring the importance of individual assessment over severity-level generalisations. Future research should extend evaluation to larger, more diverse corpora spanning multiple languages, aetiology types, and clinically validated severity assessment frameworks.

Absence of user-centred validation: This study employed quantitative performance metrics without direct input from dysarthric users. Whilst WER and semantic similarity provide useful technical characterisation, they do not capture the full complexity of functional communication success, user satisfaction, perceived autonomy, or trust in assistive technology. A system that achieves 30% WER might be perceived as highly useful by one user for whom it enables previously impossible communication, yet frustrating for another user with higher accuracy expectations. The practical significance of our findings therefore remains to be validated through user-centred evaluation. We strongly recommend that practitioners consider user feedback alongside technical metrics when selecting systems for deployment.

Semantic metric validity: The embedding-based semantic measures employed (SeMaScore, BERTScore) were developed and validated on typical speech transcription tasks. Their behaviour on dysarthric speech, where both reference utterances and recognition hypotheses may contain atypical lexical patterns or disfluencies, has not been extensively validated. High semantic similarity scores may not correspond to functional communicative success in all AAC contexts, and low scores may underestimate utility for users whose communication partners can interpret context. These metrics should be treated as reasonable proxies pending dedicated validation studies.

Methodological constraints: The zero-shot evaluation design, whilst reflecting realistic deployment constraints, does not explore the potential of speaker-adaptive or fine-tuned approaches that may offer superior performance for specific users. Commercial API configurations were fixed at provider defaults, limiting exploration of parameter spaces that might improve accessibility. The single prompting strategy tested may not represent optimal prompt design; alternative formulations warrant systematic investigation, though the per-speaker consistency of GPT-4o improvements suggests the tested prompt captures a meaningful intervention. Latency measurements were derived from per-speaker aggregate timing data; whilst this provides reliable central tendency estimates, individual utterance variability may be underestimated. Additionally, this study did not explicitly measure hallucination rates, wherein MLLMs generate plausible but incorrect words; this phenomenon may be particularly relevant for severe dysarthria where acoustic evidence is weakest and models may rely more heavily on linguistic inference.

Recording conditions: All evaluations used TORGO wav_headMic recordings captured with a head-mounted electret microphone in controlled laboratory conditions. Recognition performance may differ with array microphone recordings, ambient noise, or consumer-grade recording equipment typical of real-world AAC deployment. Future evaluations should assess robustness across recording conditions.

Ethical and privacy considerations: The use of commercial cloud-based APIs for assistive speech recognition raises important privacy and data governance considerations. Audio data

transmitted to third-party servers may be subject to retention, logging, or use for model improvement, with policies varying across providers and potentially changing over time. For clinical or healthcare-adjacent applications, practitioners must verify that selected services comply with applicable regulations (e.g., HIPAA in the United States, GDPR in the European Union) and obtain appropriate user consent for data transmission. Where privacy requirements are stringent, on-device alternatives such as locally deployed Whisper models may be preferable despite potential accuracy trade-offs. The TORGO corpus used in this study is publicly available with appropriate ethical approvals; no new human subjects data were collected.

Future research should address several directions identified by this study. First, systematic investigation of prompting strategies, including few-shot examples, domain-specific instructions, and multi-turn designs, may reveal additional interventions accessible to practitioners. Second, user-centred evaluations linking quantitative performance to perceived usability, trust, and communication success would complement technical assessment and validate the practical significance of observed differences. Third, hybrid approaches combining ASR with downstream semantic processing may bridge the gap between lexical accuracy and communicative adequacy. Fourth, evaluation on more recent and diverse dysarthric speech corpora would strengthen generalisability of findings. Fifth, characterisation of hallucination patterns in MLLM transcription of severely degraded speech would inform risk assessment for safety-critical applications. Sixth, evaluation of short-command recognition using appropriately designed protocols would address the gap created by excluding single-word utterances from the present analysis.

6 Conclusion

This study presents a severity-stratified, zero-shot evaluation of commercial speech recognition services for dysarthric speech, addressing an important gap in the universal access evidence base. Eight systems, comprising four conventional ASR and four MLLM-based services, were assessed across lexical accuracy, semantic fidelity, and operational efficiency dimensions using the TORGO dysarthric speech corpus. The findings establish empirical baselines that quantify the current accessibility gap: whilst users with mild dysarthria may achieve near-typical voice interface functionality, those with severe dysarthria face recognition failure rates exceeding 49% WER across all evaluated systems.

The study makes three primary contributions to the universal access literature. First, it provides actionable, severity-stratified performance data enabling evidence-based technology selection for assistive applications, including per-speaker analysis demonstrating substantial within-severity variability. Second, it characterises architecture-specific prompting effects, revealing that the tested verbatim-transcription prompt significantly improves OpenAI GPT-4o performance with 100% consistency across severe dysarthria speakers, whilst degrading Gemini systems, a finding with direct implications for deployment practice. Third, it situates speech recognition accessibility within the broader discourse on inclusive design, demonstrating how commercial voice technologies systematically privilege typical speech characteristics whilst excluding users with motor speech disorders.

Accessible voice interaction is not a peripheral concern but a matter of digital inclusion in an information society increasingly mediated by speech. As voice interfaces proliferate across domestic, occupational, and public domains, the performance gap documented here translates into systematic exclusion of individuals with dysarthria from the benefits of voice-mediated technology. Closing this gap requires continued research, commercial attention to diverse

speech populations, and design practices that treat human speech diversity not as exceptional but as fundamental.

References

- Stephanidis C. The universal access handbook. CRC Press; 2009.
- Jaddoh A, Loizides F, Lee J, Rana O. An interaction framework for designing systems for virtual home assistants and people with dysarthria. *Universal Access in the Information Society*. 2025;24(1):267–79.
- Duffy JR. Motor speech disorders: Substrates, differential diagnosis, and management. Elsevier Health Sciences; 2012.
- Park C, Chen M, Hain T. Automatic Speech Recognition System-Independent Word Error Rate Estimation. *ArXiv*. 2024;abs/2404.16743. doi: 10.48550/arxiv.2404.16743.
- Hui M, Zhang J, Mohan A. Enhancing AAC Software for Dysarthric Speakers in e-Health Settings: An Evaluation Using TORGO. *arXiv preprint arXiv:241100980*. 2024.
- Persson H, Åhman H, Yngling AA, Gulliksen J. Universal design, inclusive design, accessible design, design for all: different concepts—one goal? On the concept of accessibility—historical, methodological and philosophical aspects. *Universal access in the information society*. 2015;14(4):505–26.
- Singh S, Wang Q, Zhong Z, Mendes C, Hasegawa-Johnson M, Abdulla W, et al. Robust Cross-Etiology and Speaker-Independent Dysarthric Speech Recognition. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: IEEE; 2025. p. 1–5.
- Sasindran Z, Yelchuri H, Prabhakar T. SeMaScore: a new evaluation metric for automatic speech recognition tasks. *arXiv preprint arXiv:240107506*. 2024.
- Mair JL, Hashim J, Thai L, Tai ES, Ryan JC, Kowatsch T, et al. Understanding and overcoming barriers to digital health adoption: a patient and public involvement study. *Translational Behavioral Medicine*. 2025;15(1):ibaf010.
- Attaluri K, CHVS A, Chittepu S. Empowering Dysarthric Speech: Leveraging Advanced LLMs for Accurate Speech Correction and Multimodal Emotion Analysis. *arXiv preprint arXiv:241012867*. 2024.
- Cappellazzo U, Kim M, Chen H, Ma P, Petridis S, Falavigna D, et al. Large language models are strong audio-visual speech recognition learners. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: IEEE; 2025. p. 1–5.
- Hu S, Zhou L, Liu S, Chen S, Meng L, Hao H, et al. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:240400656*. 2024.
- Rudziej F, Namasivayam AK, Wolff T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*. 2012;46(4):523–41.
- Stephanidis C, Savidis A. Universal access in the information society: methods, tools, and interaction technologies. *Universal access in the information society*. 2001;1(1):40–55.
- Stephanidis C, Antona M. Universal access in the information Society (2001–2021): Knowledge, experience, challenges and new perspectives. *Universal Access in the Information Society*. 2022;21(2):329–31.
- Zheng W-Z, Han J-Y, Cheng H-L, Chu W-C, Chen K-C, Lai Y-H. Comparing the performance of classic voice-driven assistive systems for dysarthric speech. *Biomedical Signal Processing and Control*. 2023;81:104447.
- Qian Z, Xiao K. A survey of automatic speech recognition for dysarthric speech. *Electronics*. 2023;12(20):4278.
- Yuksel KA, Ferreira T, Gunduz A, Al-Badrashiny M, Javadi G. A reference-less quality metric for automatic speech recognition via contrastive-learning of a multi-language model with self-supervision. *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*: IEEE; 2023. p. 1–5.
- Zheng X, Phukon B, Hasegawa-Johnson M. Fine-Tuning Automatic Speech Recognition for People with Parkinson's: An Effective Strategy for Enhancing Speech Technology Accessibility. *arXiv preprint arXiv:240919818*. 2024.
- Hernandez A, Pérez-Toro PA, Nöth E, Orozco-Arroyave JR, Maier A, Yang SH. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. *arXiv preprint arXiv:220401670*. 2022.
- Al-Qatab BA, Mustafa MB, Salim SS. Maximum Likelihood Linear Regression (MLLR) for ASR Severity Based Adaptation to Help Dysarthric Speakers. *International Journal of Simulation—Systems, Science & Technology*. 2014;15(6).
- Geng M, Xie X, Deng J, Jin Z, Li G, Wang T, et al. Homogeneous Speaker Features for on-the-Fly Dysarthric and Elderly Speaker Adaptation and Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing*. 2025.
- Sapkota P, Sinha A, Kathania HK, Kadiri SR. Enhancing Traditional Kaldi Dysarthric Speech Recognition Using SSL-Features. *2025 National Conference on Communications (NCC)*: IEEE; 2025. p. 1–6.
- Yue Z, Xiong F, Christensen H, Barker J. Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: IEEE; 2020. p. 6094–8.
- Alsayegh A, Masood T. Leveraging generative AI for knowledge-driven information retrieval in the energy sector. *MATEC Web of Conferences: EDP Sciences*; 2024. p. 10008.
- Morris AC, Maier V, Green PD. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. *Interspeech2004*. p. 2765–8.
- Zhang Y, Ai J. Semantic-Weighted Word Error Rate Based on BERT for Evaluating Automatic Speech Recognition Models. *2024 11th International Conference on Dependable Systems and Their Applications (DSA)2024*. p. 189–98.
- Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. *ArXiv*. 2019;abs/1904.09675.

29. Alsayegh A, Ain NU, Abel A, Jia L, Masood T. TranscribeSight: redefining ASR evaluation for Industry 4.0 knowledge capture. *Procedia Computer Science*. 2025.
30. TranscribeSight: A comprehensive speech-to-text comparison platform. <https://github.com/toochii2/TranscribeSight> (2025). Accessed May 19 2025.
31. Kim H, Hasegawa-Johnson M, Perlman A, Gunderson JR, Huang TS, Watkin KL, et al. Dysarthric speech database for universal access research. *Interspeech2008*. p. 1741–4.
32. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. *International conference on machine learning: PMLR*; 2023. p. 28492–518.
33. Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an asr corpus based on public domain audio books. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP): IEEE; 2015. p. 5206–10.
34. Ardila R, Branson M, Davis K, Kohler M, Meyer J, Henretty M, et al. Common voice: A massively-multilingual speech corpus. *Proceedings of the twelfth language resources and evaluation conference2020*. p. 4218–22.
35. Bögels S, Kendrick KH, Levinson SC. Conversational expectations get revised as response latencies unfold. *Language, Cognition and Neuroscience*. 2020;35(6):766–79.