Toward Equitable Machine Translation for Atypical Speech: An LLM Post-Correction Approach

Grace Pasion gpasion0@byu.edu
Ammon Shurtz acshurtz@byu.edu
Stephen D. Richardson srichardson@byu.edu
Department of Computer Science, Brigham Young University

Abstract

Individuals with speech disabilities rely on everyday technologies powered by Automatic Speech Recognition (ASR) systems, yet these systems consistently fail them–producing significantly higher error rates that undermine the usefulness of voice assistants, hands-free devices, and machine translation pipelines. We conduct a multi-stage evaluation of speech impairment effects in cascaded speech-to-text translation, examining two distinct conditions: real dysarthric speech and simulated rhotacism. For dysarthria, we quantify ASR error rates. For rhotacism, we quantify error rates from a minimal-pair text substitution. We then analyze how impairment-induced errors propagate through downstream machine translation across three language directions (English to Spanish, Ukrainian, Khmer), and propose a training-free LLM-based post-correction methodology as an accessible intervention. We find that larger LLMs (70B parameters or more) consistently improve downstream translation quality even when surface-level corrections made by those LLMs remain modest, while smaller models lack the capacity to do so reliably. These results reveal a promising but scale-dependent path toward more equitable speech technology for users with atypical speech.

1 Introduction

Voice-user interfaces (VUIs) have become deeply embedded in everyday life, powering applications ranging from virtual assistants to accessibility tools. These technologies rely on sophisticated Automatic Speech Recognition (ASR) systems to interpret user intent. However, for a significant portion of the global population, these systems fail to perform reliably due to speech disorders and disabilities. According to a 2012 study, approximately 10% of U.S. adults have a speech, language, and/or voice disability (Morris et al., 2016), and more than 80 million people worldwide stutter.[1] Despite this substantial user base, those with speech disorders continue to encounter significantly higher error rates when using ASR-interfaced systems (Jefferson, 2019; Tobin et al., 2024), highlighting a critical gap between the populations these technologies serve and those they are designed to accommodate.

ASR is an essential component for a range of downstream NLP tasks that rely on spoken input. One prominent example is Speech-to-Speech Machine Translation (S2ST), in which a cascaded pipeline integrates ASR, Machine Translation (MT), and Text-to-Speech (TTS) components to produce a fully end-to-end spoken translation system. In such pipelines, the performance of each component is inherently dependent on the quality of its predecessor, meaning that ASR errors propagate downstream and compound throughout the system (Matusov et al., 2005; Min et al., 2025). This cascade effect makes atypical speech particularly problematic: recognition errors introduced at the ASR stage can fundamentally distort the meaning of subsequent translations, disproportionately affecting speakers with speech disorders.

[1] https://www.stutteringhelp.org/faq

While cascaded S2ST systems have seen considerable advancement in recent years, the robustness of such pipelines to atypical speech remains poorly understood. Despite substantial progress in ASR robustness, speech impediment modeling, and LLM-based error correction, existing work largely studies these components in isolation. Additionally, prior research on speech-impaired ASR output primarily focuses on evaluating recognition accuracy without examining how impairment-induced errors propagate to downstream tasks, like S2ST. Similarly, work on MT robustness to ASR noise typically assumes generic or synthetic noise patterns rather than systematic articulatory errors caused by speech impairments.

Large Language Models (LLMs) have recently emerged as a promising post-processing solution for ASR error correction (Pu et al., 2023; Tang et al., 2025; Iudin et al.), offering a potential remedy that does not require retraining or adapting the underlying recognition model. These studies rarely consider speech impairments explicitly, compare correction at different stages of the pipeline, or evaluate downstream translation quality. Whether input corrections effectively mitigate the compounding effects of speech-impairment noise in a cascaded translation pipeline remains an open question.

Thus, in this work, we aim to address these gaps through a systematic, multi-stage evaluation of speech-impairment effects in cascaded speech-to-text translation. We analyze both real dysarthric speech and a simulated rhotacism condition, comparing three open-source LLMs of varying sizes in their ability to correct impairment-induced ASR errors. A simulated rhotacism dataset is employed due to the lack of high-quality real datasets for this specific speech impairment. We then quantify how these corrections affect downstream translation quality from English into three target languages representing high, medium, and low-resource settings.

Overall, all three LLMs struggle to substantially correct dysarthria-induced ASR errors, achieving only up to a 1.4% relative improvement in Word Error Rate (WER) (Marzal and Vidal, 2002) compared to the original ASR output. In contrast, simulated rhotacism errors yield a modestly higher average improvement of up to 2.6% WER. Despite these limited gains at the recognition level, we observe a more nuanced pattern downstream. For both dysarthric speech and simulated rhotacism, the two largest open-source LLMs evaluated — LLAMA3 (70B) and Qwen3 (235B) — consistently improve translation performance across all three language directions. In contrast, the smallest model (OLMo-3 7B) does not demonstrate sufficient capacity to produce consistent downstream gains. These findings suggest that modest surface-level ASR improvements can nonetheless yield measurable translation benefits, but that model scale plays a critical role in effective correction.

We release the code and datasets used here: `https://github.com/byu-matrix-lab/project-atypical-speech-mt`.

2 Related Works

2.1 Robustness of ASR

Automatic speech recognition (ASR) systems are sensitive to intrinsic speaker variations, such as pitch and speaking rate (Meyer et al., 2011), and to extrinsic acoustic noise, such as the background noise generated by public crowds (Moser et al., 2025).

Early work on ASR robustness focused on auditory processing-based features (Maganti and Matassoni, 2014), spectrum enhancement with sparse coding (He et al., 2015), and LSTM-based speech enhancement with feature fusion (Weninger et al., 2015). More current approaches leverage unsupervised self-training for noise and accent robustness (Hu et al., 2024), context-aware noise representation learning (Lee et al., 2024), and two-stage deep neural network spectral mapping (Wang et al., 2020). Other approaches focus on error correction via machine translation (Mani et al., 2020), error corrections via transformer-based models (Lin et al., 2025), and training-free intelligibility-guided fusion (Li et al., 2026) to make ASR more robust.

While these techniques are generally effective, they remain less effective when systematic errors arise due to speech impediments, necessitating further downstream correction and robustness strategies.

2.2 Robustness of MT to ASR Errors

ASR errors propagate directly to MT systems in traditional cascaded speech-to-text translation. Consequently, these errors often degrade translation quality.

Early work explored adapting an NMT model on ASR transcripts to improve model robustness towards these errors (Di Gangi et al., 2019). More

recent research proposed a unified framework, Whisper-UT, that improves robustness in multimodal MT by simulating ASR errors (Xiao et al., 2025). Other work has focused on mitigating ASR errors in speech-to-text translation through cascaded systems that incorporate multiple ASR candidates and self-supervised speech features (Min et al., 2025). Additionally, some approaches for end-to-end models use regularized cross-attention to transfer alignment knowledge from ASR and MT tasks (Zhao et al., 2024). Finally, low-resource approaches exist that leverage synthetic data and model regularization to improve both cascaded and end-to-end ST performance (Li et al., 2025).

2.3 ASR Errors Induced by Speech Impediments

ASR performance degrades substantially when applied to speech affected by disorders such as dysarthria, apraxia, or articulation impairments. For example, commercial ASR platforms are limited in their ability to decipher dysarthric speech, for which a 80%-90% error rate has been reported (De Russis and Corno, 2019). Other research emphasizes the gap between ASR errors from users with speech impediments and those without one (Jefferson, 2019; Tobin et al., 2024). Clearly, there is work to be done to improve ASR models for users with disabilities.

2.4 Speech Impediment Correctors

Several methods have been developed to improve ASR performance for individuals with speech impairments. Techniques include EEG-assisted deep learning (Krishna et al., 2021), neural network–based speech assistance (Mounir et al., 2017), cross-lingual self-supervised representations (Hernandez et al., 2022), and two-stage data augmentation (Bhat and Strik, 2025). Further, other advances leverage transformer architectures (Shahamiri et al., 2023), empirical mode decomposition with CNNs (Sidi Yakoub et al., 2020), acoustic and lexical model adaptation (Mengistu and Rudzicz, 2011), hybrid feature extraction with evolutionary optimization (Rughani and Shivakrishna, 2015), and speaker-dependent parameter tuning (Marini et al., 2021) to improve recognition accuracy and robustness. Finally, methods such as fine-tuning the Wav2Vec 2.0 transformer model (Sapkota et al., 2026) and Mel spectrogram–based transfer learning for dysarthric ASR (Dhanalakshmi et al., 2025) further improve accuracy.

2.5 Large Language Models as ASR Correctors

Large Language Models (LLMs) have recently been applied as post-processing tools to improve ASR output. For general ASR errors, studies have explored multi-stage reasoning approaches (Pu et al., 2023), full-text error correction (Tang et al., 2025), post-correction fine-tuning (Iudin et al.), task-activating prompting (Yang et al., 2023), and zero-shot correction (Ma et al., 2025).

In addition to general ASR errors, LLMs have been used to correct errors in speech-impairment contexts, including dysarthric and stuttering speech. For example, La Quatra et al. (2025); Aboeitta et al. (2025); Bhat and Strik (2025); Lou et al. (2025) used LLM-based post-correction to improve speech-impaired ASR, while Hui et al. (2025) applied LLM-augmented methods in AAC software for dysarthric speakers. LLMs have also been applied to related tasks such as stuttering detection (Huang et al., 2025). Others have used either user-adapted, fine-tuned, or instruction-tuned LLMs to correct disordered speech (Hsu and Chang, 2025; Hernandez et al., 2025b,a). All in all, through previous research, LLMs have been shown to be robust at speech corrections for both ASR errors and ASR speech impediment errors.

3 Datasets

We employ a combination of clinical, simulated, and multilingual datasets to study speech impairments and induced rhotic errors.

3.1 Dysarthria

For our experiments, we use the Torgo dataset which contains audio recordings from individuals with speech impairments due to cerebral palsy or amyotrophic lateral sclerosis (Rudzicz et al., 2012). The dataset includes those with dysarthria and those without a speech impairment. The audio from those in the control group was filtered out. Additionally, the dataset was obtained by using a head mounted and directional microphone from the same session. We chose to use the head mounted data to better approximate real world conditions. Moreover, because Torgo consists of individual words and phrases, we extracted only the complete sentences for analysis.

Sentence transcription was performed using a large Whisper model, which is a state-of-the-art open-source ASR system widely used in research (Radford et al., 2023). After aligning the ASR output with the

original prompts, we were left with 691 lines of text.

3.2 Rhotacism

Current rhotacism datasets have their limitations, as many do not include the original prompts or full sentences. Consequently, we instead chose to simulate rhotic errors via text-based minimal-pair substitution. While designed to model potential ASR performance, this synthetic data does not represent genuine ASR transcripts.

3.2.1 Language Pairs

For our simulation, we use parallel sentences pulled from the Tatoeba[1] corpus as described by Tiedemann (2020). Tatoeba was selected because its large parallel corpora allowed us to filter for sentences containing sufficient rhotic phonemes while preserving aligned reference translations. In contrast, smaller benchmarks would have provided fewer usable examples after rhotacism-specific filtering. Additionally, Tatoeba is widely used in machine translation research, making it a suitable choice for our study. We pulled a random subset from three different datasets which include the English-to-Spanish, English-to-Ukrainian, and English-to-Khmer language pairs. Spanish, Ukrainian, and Khmer were chosen to compare the possible differences between a high-resource, a medium-resource, and a low-resource language.

To ensure the datasets were suitable for modeling induced rhotic errors, we filtered the data to retain only lines containing five "r"s in the English–Spanish and English–Ukrainian files, and two "r"s per line for English–Khmer. The resulting datasets included 1000 lines of English-to-Spanish, 999 lines of English-to-Ukrainian, and 386 lines of English-to-Khmer lines.

3.2.2 Simulation Techniques

Rhotacism is defined as a speech impediment that affects the production of the phoneme /r/, frequently leading to substitutions that resemble /l/ or /w/ (Flipsen Jr, 2021). Thus, in order to simulate rhotacism we utilize minimal pairs which are words that differ by one phoneme. For example, right (R AY1 T) and white (W AY1 T) are considered minimal pairs. To craft a minimal pair dictionary, we leverage CMUDict which is an open-source pronunciation dictionary (Carnegie Mellon University Speech Group, 2014).

To generate the simulated-rhotacism dataset, we first processed the original English sentences and identified minimal pairs within each sentence. Whenever a matching word was found, it was replaced with its corresponding minimal pair. To replicate real-world noise, there was a 1% probability that a word would remain unchanged. This simulation resulted in three aligned files of the original English-Target language pair and the aligned English rhotacism simulated file.

4 Methodology

We now define the technical architecture, algorithms, and data flow used to evaluate how speech-impairment-induced ASR errors affect downstream machine translation, and whether input-level correction mitigates these effects.

4.1 System Architecture Overview

Our system follows a cascaded speech-to-text translation pipeline consisting of four primary components:

1. Speech Recognition or Text Simulation
2. Optional Input Correction
3. Machine Translation
4. Evaluation

For dysarthria, the pipeline begins with impaired speech audio, while for rhotacism, it begins with clean English text that is algorithmically modified to simulate articulatory errors. In both cases, the output is English text that contains systematic recognition or articulation errors. Additionally, this text may optionally be corrected before being passed into a machine translation model. The translated outputs are then evaluated using automatic metrics.

4.2 Input-Level LLM Correction

To mitigate impairment-induced errors prior to translation, we apply input-level correction using instruction-tuned large language models. Each ASR output or simulated sentence is processed independently using a fixed prompt that instructs the model to correct errors caused by the specific speech impairment. The prompt used is included in Appendix A.

We evaluate three open-source LLMs:

- LLaMA-3.3-70B-Instruct
- Qwen3-235B-A22B

- OLMo-3-7B-Instruct

These models were selected to represent current state-of-the-art open-source systems across different parameter scales and training strategies (Grattafiori et al., 2024; Yang et al., 2025; Olmo et al., 2025).

LLMs are known to produce variable outputs or be sensitive to prompt formatting even under constrained prompting, with minor changes in prompt wording or format leading to substantially different responses across multiple models (Chatterjee et al., 2024). Thus, it is necessary that we apply deterministic post-processing rules. If the output has tags or formatting difficulty such as *"The corrected sentence is: [THE SENTENCE],"* it was cleaned to not include the extra format. Additionally, if the LLM output has no meaningful sentence to extract or there was a failure for any other reason such as model refusal for inappropriate sentences, the original sentence is returned. Finally, if the outputted sentence is a legible, extractable sentence it is simply returned. These rules applied to all correction files.

As a result from the input correction process, our erroneous English sentences were transformed to corrected English sentences.

4.3 Machine Translation

To evaluate downstream effects on translation, all English source variants are translated using the same MT system: facebook/nllb-200-distilled-600M (Costa-Jussà et al., 2022). Decoding uses the Hugging Face default generation settings (greedy decoding; no beam size or length penalty was explicitly specified), and these settings are held constant across all conditions. This model was selected due to its strong multilingual performance, coverage of low-resource languages, and open-source availability.

We translate all source conditions into Spanish, Ukrainian, and Khmer, representing high, medium, and low resource target languages. Translation is performed using identical decoding settings across all conditions to ensure comparability.

5 Experiments

5.1 Experimental Setup

Baselines and Comparisons For our experiments, we compare the following conditions:

- Clean Prompt Translation (Upper Bound): The original English Prompt translated directly
- Raw Impaired Translation (Baseline): The uncorrected ASR output or simulated rhotacism text translated
- Input-Corrected Translation: The LLM-corrected English input translated

This design isolates the effect of speech impairment and evaluates whether correction prior to MT recovers translation quality.

Data Splits and Processing All datasets are evaluated in a sentence-level, one-pass setting without train/dev/test splits, as no model parameters are trained. All processing is deterministic given fixed prompts and decoding parameters.

5.2 Evaluation Methodology

5.2.1 Speech Impediment Effect

To measure the severity of impairment-induced text corruption, we compare raw impaired text against the original English prompts using Word Error Rate (WER) which is a normalized edit distance computed from insertions, deletions, and substitutions between two sequences (Marzal and Vidal, 2002). WER directly measures recognition errors common in impaired speech. This metric allows us to quantify impairment severity and assess the effectiveness of input-level correction independently of translation.

5.2.2 Machine Translation Effect

For rhotacism, we compute BLEU scores using the sacrebleu library (Post, 2018). System outputs are compared against aligned reference translations within the same language. For our evaluation, Khmer uses the *flores200 tokenizer* (Costa-Jussà et al., 2022) and all the other languages uses the default *13a tokenizer* which is the sacreBLEU implementation of the mteval-13a tokenization script. All other settings were left to the defaults of sacreBLEU. Scores are only computed for datasets with available reference translations. Because the TORGO dataset only has the aligned text prompt and the corresponding audio file, BLEU cannot be calculated for the dysarthria dataset.

For both rhotacism and dysarthria, we use MetricX-24-Hybrid-XXL[2] and XCOMET-XL[3] (Juraska et al., 2024; Guerreiro et al., 2024). These

[2] https://huggingface.co/google/metricx-24-hybrid-xxl-v2p6
[3] https://huggingface.co/Unbabel/XCOMET-XL

models estimate translation quality using the source and hypothesis alone and have been shown to correlate strongly with human judgments. MetricX[4] emphasizes semantic adequacy, while XCOMET captures overall translation quality using cross-lingual contextual representations.

Together, these metrics provide a comprehensive evaluation of how speech impairments propagate through MT and how effectively input correction mitigates their impact.

6 Results & Discussion

Using the dysarthria and simulated rhotacism data described in Section 3, we first evaluate the effect of speech impairments and the ability of the three open-source LLMs to correct the errors. After this, we evaluate the downstream Machine Translation performance comparing the uncorrected to the corrected sentences.

All Word Error Rates (WER) are presented as percentages.

6.1 ASR Results

In Table 1, we show the WER for the dysarthric ASR output and the simulated rhotacism condition when compared to the correct reference sentences. The dysarthria transcript has a very large WER, above 100 indicating extremely poor ASR performance. Meanwhile the three rhotacism datasets achieve a small WER percentage appropriate for the proportion of words with /r/ sounds. These values act as baselines for our LLM correction experiments.

Dataset	WER ↓
Dysarthria	113.07
Rhotacism - Khmer	6.80
Rhotacism - Ukranian	4.08
Rhotacism - Spanish	4.27

Table 1: Word Error Rates (WER) calculated against the correct transcriptions, showing authentic ASR performance for dysarthria alongside results from the three simulated rhotacism datasets.

6.2 LLM Correction Results

Table 2 presents the WER of each LLM-corrected output relative to the original ASR output, where a negative Δ indicates an improvement over the uncorrected baseline. For both dysarthria and rhotacism, model scale is a strong predictor of correction quality. Qwen3 (235B) achieves the best overall performance, attaining the lowest post-correction WER in nearly every setting and producing the largest reductions for both dysarthria ($\Delta = -1.40$) and the rhotacism conditions (Δ ranging from -2.47 to -2.73). LLaMA (70B) follows closely behind, also improving over the uncorrected baseline across all rhotacism datasets. OLMo (7B), by contrast, consistently increases WER relative to the uncorrected input, suggesting the model lacks sufficient capacity to reliably distinguish and repair impairment-induced errors.

Model	Dataset	WER ↓	Δ ↓
OLMo	Dysarthria	114.04	0.97
OLMo	Rhotacism - Khmer	9.64	2.84
OLMo	Rhotacism - Ukranian	6.55	2.47
OLMo	Rhotacism - Spanish	7.74	3.47
LLaMA	Dysarthria	113.29	0.22
LLaMA	Rhotacism - Khmer	5.22	-1.58
LLaMA	Rhotacism - Ukranian	3.05	-1.03
LLaMA	Rhotacism - Spanish	3.08	-1.19
Qwen3	Dysarthria	111.67	-1.40
Qwen3	Rhotacism - Khmer	4.07	-2.73
Qwen3	Rhotacism - Ukranian	1.53	-2.55
Qwen3	Rhotacism - Spanish	1.80	-2.47

Table 2: WER of each LLM-corrected output compared to the correct transcriptions. Δ denotes the change relative to the uncorrected ASR baseline from Table 1, where negative values indicate improvement.

For dysarthria specifically, all three models struggle, with corrections yielding only marginal changes — even the best result from Qwen3 represents a modest 1.4% relative improvement. This is unsurprising given the severity of the dysarthria WER baseline (113.07), where the ASR output is heavily degraded and the correction signal is correspondingly weak. The rhotacism conditions, while

[4]We use MetricX-24 rather than MetricX-25 because, although MetricX-25 has been described in the literature, the model was not publicly available at the time of our experiment Juraska et al. (2025).

Experiment	Target Language	XCOMET ↑	MetricX ↓
Uncorrected	Khmer	0.6466	10.4437
OLMo Corrected	Khmer	0.6438	10.2847
LLaMA Corrected	Khmer	**0.6518**	**10.1358**
Qwen3 Corrected	Khmer	0.6512	10.1901
Clean	Khmer	0.8014	6.2781
Uncorrected	Ukrainian	0.7321	9.4699
OLMo Corrected	Ukrainian	0.7347	9.3444
LLaMA Corrected	Ukrainian	**0.7414**	**9.1619**
Qwen3 Corrected	Ukrainian	0.7388	9.2161
Clean	Ukrainian	0.9212	4.0137
Uncorrected	Spanish	0.7677	8.9878
OLMo Corrected	Spanish	0.7645	8.8321
LLaMA Corrected	Spanish	0.7716	**8.3588**
Qwen3 Corrected	Spanish	**0.7739**	8.4062
Clean	Spanish	0.9388	2.3965

Table 3: Downstream MT results for Dysarthria across three target languages. *Clean* denotes translation from ground-truth transcriptions; *Uncorrected* denotes translation from unmodified ASR output.

lower in overall error rate, show a clearer correction benefit, particularly for Qwen3 and LLaMA, whose post-correction WERs approach near-clean levels in several cases.

6.3 Downstream MT Results

Tables 3 and 4 present the downstream MT results for dysarthria and rhotacism respectively. In each table, the *clean* row represents the theoretical performance ceiling (translation from ground-truth transcriptions), while the *uncorrected* row represents the floor, reflecting translation directly from unmodified ASR output. The LLM-corrected rows fall between these two bounds and are the focus of this analysis.

Dysarthria. Despite the modest WER improvements observed in Table 2, LLM-based correction still yields consistent downstream translation gains for the two largest models. LLaMA achieves the best XCOMET and MetricX scores in two of the three language directions (Khmer and Ukrainian), while Qwen3 has the best XCOMET scores for Spanish. The two models are largely comparable, with differences falling within a narrow range. OLMo, consistent with its correction results, fails to improve over the uncorrected baseline XCOMET scores for Khmer and Spanish, though it does show minor MetricX gains across all three directions. Overall, improvements remain modest relative to the gap between the uncorrected and clean conditions, which is expected given the limited correction gains at the ASR stage.

Rhotacism. The rhotacism results are more clear. Qwen3 generally yields the highest translation scores across conditions, though its advantage over LLaMA is typically modest and not consistent across all settings. Notably, Qwen3-corrected output approaches near-clean translation performance in several cases — for Ukrainian and Spanish in particular, the gap between Qwen3-corrected and clean narrows substantially across both XCOMET and MetricX. This is consistent with the near-zero post-correction WERs observed for Qwen3 in Table 2, suggesting that when surface-level errors are systematic and correctable, the effect of strong LLM correction can propagate meaningfully through the downstream pipeline. OLMo again lags behind, showing only marginal gains over the uncorrected baseline and failing to match LLaMA or Qwen3 in any direction.

6.4 Summary

Taken together, these results highlight two consistent themes. First, model scale matters: Qwen3 and LLaMA outperform OLMo at both the correction and translation stages across nearly every condition, while OLMo frequently fails to improve over — or

Experiment	Target Language	BLEU ↑	XCOMET ↑	MetricX ↓
Uncorrected	Khmer	8.76	0.8663	4.9233
OLMo Corrected	Khmer	8.74	0.8748	4.6139
LLaMA Corrected	Khmer	9.65	0.9046	3.8716
Qwen3 Corrected	Khmer	**9.54**	**0.9122**	**3.6724**
Clean	Khmer	9.71	0.9240	3.2711
Uncorrected	Ukrainian	34.00	0.9231	3.7415
OLMo Corrected	Ukrainian	34.42	0.9362	3.2167
LLaMA Corrected	Ukrainian	35.69	0.9544	2.6289
Qwen3 Corrected	Ukrainian	**36.37**	**0.9584**	**2.4417**
Clean	Ukrainian	37.55	0.9660	2.1251
Uncorrected	Spanish	47.43	0.9275	3.3913
OLMo Corrected	Spanish	48.07	0.9420	2.5573
LLaMA Corrected	Spanish	50.78	0.9565	1.9746
Qwen3 Corrected	Spanish	**51.22**	**0.9632**	**1.4978**
Clean	Spanish	52.73	0.9690	1.1615

Table 4: Downstream MT results for simulated rhotacism across three target languages. *Clean* denotes translation from ground-truth transcriptions; *Uncorrected* denotes translation from unmodified simulated ASR output.

actively degrades — the uncorrected baseline. Second, even modest surface-level corrections can yield measurable downstream benefits.

For dysarthria, where WER reductions are small, LLM correction still produces consistent translation gains for the two larger models. For rhotacism, where the errors are more systematic and more correctable, this effect is amplified — with Qwen3-corrected output approaching clean translation quality in the three language directions examined. These findings underscore both the promise and the current limitations of LLM-based ASR post-correction as a drop-in component for cascaded speech translation pipelines serving speakers with speech impairments.

Representative transcript examples illustrating both ASR errors and LLM corrections are provided in Appendix A.

7 Conclusion

In this work, we presented a systematic evaluation of speech impairment effects in cascaded speech-to-text translation, examining both real dysarthric speech and a text-based simulated rhotacism condition. We compared three open-source LLMs of varying scale as ASR post-correction tools and measured the downstream impact on translation quality across three language directions.

Overall, our results demonstrate that while LLM-based correction yields only modest WER improvements–particularly for dysarthria–these surface-level gains can nonetheless produce consistent downstream translation benefits when model scale is sufficient. Qwen3 and LLaMA both improve over the uncorrected baseline across nearly all conditions, with Qwen3-corrected rhotacism output approaching clean translation quality in the three language directions examined. OLMo, at 7B parameters, lacks the capacity to reliably correct either condition. Taken together, these findings highlight both the potential and the current limitations of LLM post-correction as an accessible, training-free intervention for improving cascaded pipeline robustness for atypical speech.

Several directions remain open for future work. Most directly, our rhotacism experiments rely on simulated ASR errors rather than real recordings; collecting and evaluating on authentic rhotacism audio would provide a stronger empirical foundation. More broadly, extending this methodology to other speech disorders–such as stuttering or apraxia–would help characterize how impairment type and severity interact with downstream pipeline performance.

Thus, future work may benefit from speech-disability resources such as the Speech Accessibility Project , which contains substantially larger collections of authentic atypical speech Hasegawa-Johnson et al. (2024). Finally, it remains an open question whether other NLP technologies that consume spoken input, such as spoken dialogue systems or speech-based information retrieval, are similarly affected by impairment-induced ASR errors, and whether LLM post-correction offers comparable benefits in those settings.

Limitations

While our results offer meaningful insights into the effects of speech impairments on cascaded translation pipelines, several limitations should be noted.

Our rhotacism experiments rely on simulated ASR errors via minimal pair substitution rather than real speaker recordings, and future work should validate these findings using authentic rhotacic speech. The Torgo dataset also lacks reference translations, limiting MT evaluation to reference-free neural metrics (XCOMET and MetricX) and preventing direct comparison with rhotacism results. Because Tatoeba is publicly available, we cannot completely rule out the possibility of training-data overlap. We therefore interpret absolute translation scores cautiously and focus primarily on relative differences between conditions evaluated under identical settings. Beyond data constraints, our study examines only two speech impairments, and findings should not be assumed to generalize to conditions such as stuttering or apraxia of speech. The LLM correction experiments are similarly constrained by a single fixed prompt per condition; given known prompt sensitivity, our results likely represent a lower bound on what post-correction can achieve.

References

Ahmed Aboeitta, Ahmed Sharshar, Youssef Nafea, and Shady Shehata. Bridging asr and llms for dysarthric speech recognition: Benchmarking self-supervised and generative approaches. *arXiv preprint arXiv:2508.08027*, 2025.

Chitralekha Bhat and Helmer Strik. Two-stage data augmentation for improved asr performance for dysarthric speech. *Computers in Biology and Medicine*, 189: 109954, 2025.

Carnegie Mellon University Speech Group. *The CMU Pronouncing Dictionary*. Carnegie Mellon University, 2014. Version 0.7b, available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. Posix: A prompt sensitivity index for large language models, 2024. URL https://arxiv.org/abs/2410.02185.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Luigi De Russis and Fulvio Corno. On the impact of dysarthric speech on contemporary asr cloud platforms. *Journal of Reliable Intelligent Environments*, 5(3):163–172, 2019.

M Dhanalakshmi, S Bharathipriya, E Vasantha Surya, and R Krishna Raj. Automatic speech recognition system for dysarthria leveraging mel spectrograms and transfer learning. In *2025 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT)*, pages 1–6. IEEE, 2025.

Matti Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. Robust neural machine translation for clean and noisy speech transcripts. In Jan Niehues, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico, editors, *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, November 2-3 2019. Association for Computational Linguistics. URL https://aclanthology.org/2019.iwslt-1.32/.

Peter Flipsen Jr. *Remediation of/r/for speech-language pathologists*, volume 1. Plural Publishing, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet:

Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwilling, Marie Moore Channell, Laura Mattie, Heather Hodges, et al. Community-supported shared infrastructure in support of speech accessibility. *Journal of Speech, Language, and Hearing Research*, 67(11):4162–4175, 2024.

Yongjun He, Guanglu Sun, and Jiqing Han. Spectrum enhancement with sparse coding for robust speech recognition. *Digital Signal Processing*, 43:59–70, 2015.

Abner Hernandez, Paula Andrea Pérez-Toro, Elmar Nöth, Juan Rafael Orozco-Arroyave, Andreas Maier, and Seung Hee Yang. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition, 2022. URL https://arxiv.org/abs/2204.01670.

Abner Hernandez, Tomás Arias-Vergara, Andreas Maier, and Paula Andrea Pérez-Toro. Enhancing asr accuracy for speakers with parkinson's disease using instruction-tuned llms. In *International Conference on Text, Speech, and Dialogue*, pages 153–164. Springer, 2025a.

Abner Hernandez, Tomás Arias Vergara, Andreas Maier, and Paula Andrea Pérez-Toro. Confidence-guided error correction for disordered speech recognition. *arXiv preprint arXiv:2509.25048*, 2025b.

Yu-Chen Hsu and Yue-Shan Chang. A dysarthric speech recognition system with personal style embedding. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3770–3775. IEEE, 2025.

Yuchen Hu, Chen Chen, Chao-Han Yang, Chengwei Qin, Pin-Yu Chen, Eng-Siong Chng, and Chao Zhang. Self-taught recognizer: Toward unsupervised adaptation for speech foundation models. *Advances in Neural Information Processing Systems*, 37:29566–29594, 2024.

Shangkun Huang, Jing Deng, Jintao Kang, and Rong Zheng. Leveraging llm for stuttering speech: A unified architecture bridging recognition and event detection. *arXiv preprint arXiv:2505.22005*, 2025.

Macarious Hui, Jinda Zhang, and Aanchan Mohan. Enhancing aac software for dysarthric speakers in e-health settings: an evaluation using torgo. In *ICC 2025-IEEE International Conference on Communications*, pages 3673–3679. IEEE, 2025.

Artyom Iudin, Dmitrii Korzh, Matvey Skripkin, Dmitrii Masnyi, and Oleg Rogov. Enhancing speech recognition with llms in post-correction settings.

Madeline Jefferson. Usability of automatic speech recognition systems for individuals with speech disorders: Past, present, future, and a proposed model. 2019.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.35. URL https://aclanthology.org/2024.wmt-1.35/.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. Metricx-25 and gemspaneval: Google translate submissions to the wmt25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, 2025.

Gautam Krishna, Mason Carnahan, Shilpa Shamapant, Yashitha Surendranath, Saumya Jain, Arundhati Ghosh, Co Tran, Jose Del R Millan, and Ahmed H Tewfik. Brain signals to rescue aphasia, apraxia and dysarthria speech recognition. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 6008–6014. IEEE, 2021.

Moreno La Quatra, Alkis Koudounas, Valerio Mario Salerno, and Sabato Marco Siniscalchi. Exploring generative error correction for dysarthric speech recognition. *arXiv preprint arXiv:2505.20163*, 2025.

Wonjun Lee, San Kim, and Gary Geunbae Lee. Enhancing dialogue speech recognition with robust contextual awareness via noise representation learning. In Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 333–343, Kyoto, Japan, September 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.30. URL https://aclanthology.org/2024.sigdial-1.30/.

Haoyang Li, Changsong Liu, Wei Rao, Hao Shi, Sakriani Sakti, and Eng Siong Chng. Training-free intelligibility-guided observation addition for noisy asr. *arXiv preprint arXiv:2602.20967*, 2026.

Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. KIT's low-resource speech translation systems for IWSLT2025: System enhancement with synthetic data and model regularization. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 212–221, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.20. URL https://aclanthology.org/2025.iwslt-1.20/.

Ye Bhone Lin, Thura Aung, Ye Kyaw Thu, and Thazin Myint Oo. Asr error correction in low-resource burmese with alignment-enhanced transformers using phonetic features. In *2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6. IEEE, 2025.

Haowei Lou, Chengkai Huang, Hye-young Paik, Yongquan Hu, Aaron Quigley, Wen Hu, and Lina Yao. Speechagent: An end-to-end mobile infrastructure for speech impairment assistance. *arXiv preprint arXiv:2510.20113*, 2025.

Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill. Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Hari Krishna Maganti and Marco Matassoni. Auditory processing-based features for improving speech recognition in adverse acoustic conditions. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):21, 2014.

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348. IEEE, 2020.

Marco Marini, Nicola Vanello, and Luca Fanucci. Optimising speaker-dependent feature extraction parameters to improve automatic speech recognition performance for people with dysarthria. *Sensors*, 21(19):6460, 2021.

Andres Marzal and Enrique Vidal. Computation of normalized edit distance and applications. *IEEE transactions on pattern analysis and machine intelligence*, 15(9): 926–932, 2002.

E Matusov, S Kanthak, and Hermann Ney. On the integration of speech recognition and statistical machine translation. In *Proc. Interspeech 2005*, pages 3177–3180, 2005.

Kinfe Tadesse Mengistu and Frank Rudzicz. Adapting acoustic and lexical models to dysarthric speech. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4924–4927. IEEE, 2011.

Bernd T Meyer, Thomas Brand, and Birger Kollmeier. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. *The Journal of the Acoustical Society of America*, 129(1):388–403, 2011.

Anna Min, Chenxu Hu, Yi Ren, and Hang Zhao. When end-to-end is overkill: Rethinking cascaded speech-to-text translation. *arXiv preprint arXiv:2502.00377*, 2025.

Megan A Morris, Sarah K Meier, Joan M Griffin, Megan E Branda, and Sean M Phelan. Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey. *Disability and health journal*, 9(1):140–144, 2016.

Denis Moser, Nikola Stanic, and Murat Sariyar. Benchmarking speech-to-text robustness in noisy emergency medical dialogues: an evaluation of models under realistic acoustic conditions. *JAMIA open*, 8(6):ooaf147, 2025.

Ramy Mounir, Redwan Alqasemi, and Rajiv Dubey. Speech assistance for persons with speech impediments using artificial neural networks. In *ASME International Mechanical Engineering Congress and Exposition*, volume 58363, page V003T04A056. American Society of Mechanical Engineers, 2017.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch,

Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Matt Post. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191, 2018.

Jie Pu, Thai-Son Nguyen, and Sebastian Stüker. Multi-stage large language model correction for speech recognition. *arXiv preprint arXiv:2310.11532*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46(4):523–541, 2012.

Megha Rughani and D. Shivakrishna. Hybridized feature extraction and acoustic modelling approach for dysarthric speech recognition, 2015. URL https://arxiv.org/abs/1506.02170.

Paban Sapkota, Hemant Kumar Kathania, Shrikanth Narayanan, and Sudarsana Reddy Kadiri. Addressing dysarthric speech variability with severity-aware fine-tuning of transformer-based wav2vec2 asr models: P. sapkota, hk kathania, s. narayanan, sr kadiri. *Circuits, Systems, and Signal Processing*, pages 1–41, 2026.

Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3407–3416, 2023.

Mohammed Sidi Yakoub, Sid-ahmed Selouani, Brahim-Fares Zaidi, and Asma Bouchair. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1):1, 2020.

Zhiyuan Tang, Dong Wang, Shen Huang, and Shidong Shang. Full-text error correction for chinese speech recognition with large language model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Jörg Tiedemann. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*, 2020.

Jimmy Tobin, Phillip Nelson, Bob MacDonald, Rus Heywood, Richard Cave, Katie Seaver, Antoine Desjardins, Pan-Pan Jiang, and Jordan R Green. Automatic speech recognition of conversational speech in individuals with disordered speech. *Journal of Speech, Language, and Hearing Research*, 67(11):4176–4185, 2024.

Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Complex spectral mapping for single-and multi-channel speech enhancement and robust asr. *IEEE/ACM transactions on audio, speech, and language processing*, 28: 1778–1787, 2020.

Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International conference on latent variable analysis and signal separation*, pages 91–99. Springer, 2015.

Cihan Xiao, Matthew Wiesner, Debashish Chakraborty, Reno Kriz, Keith Cunningham, Kenton Murray, Kevin Duh, Luis Tavarez-Arce, Paul McNamee, and Sanjeev Khudanpur. Whisper-ut: A unified translation framework for speech and text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3000–3016, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chen-

gen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

Xiaohu Zhao, Haoran Sun, Yikun Lei, and Deyi Xiong. Regularizing cross-attention learning for end-to-end speech translation with asr and mt attention matrices. *Expert Systems with Applications*, 247:123241, 2024.

A Appendix

A.1 Prompt Template

Prompt for Corrections

This is a sentence in English that was transcribed by an ASR system. The person who spoke it has a speech impairment (**[INSERT IMPAIRMENT TYPE]**). Fix any errors caused by the speech impairment and ASR. If the sentence is not in English, correct it to proper English based on how it sounds. If the sentence appears to be accurate, return it as is with nothing extra. Return only the corrected sentence. The sentence: **[INSERT SENTENCE]**

A.2 Examples from Transcripts

Rhotacism Example 1 (Line 43)

English
Original: Fresh fruit is good for you.
Uncorrected: flesh flute is good fall you.
OLMo Corrected: flesh flute is good for you.
LLaMA Corrected: Fresh fruit is good for you.
Qwen3 Corrected: Fresh fruit is good for you.

Spanish Translation
Original: Las frutas frescas son buenas para ti.
Uncorrected: La flauta de carne es buena para ti.
OLMo Corrected: La flauta de carne es buena para ti.
LLaMA Corrected: Las frutas frescas son buenas para ti.
Qwen3 Corrected: Las frutas frescas son buenas para ti.

Rhotacism Example 2 (Line 137)

English
Original: I am writing to express my dissatisfaction.
Uncorrected: I am lighting to express my dissatisfaction.
OLMo Corrected: I am lighting out to express my dissatisfaction.
LLaMA Corrected: I am writing to express my dissatisfaction.
Qwen3 Corrected: I am writing to express my dissatisfaction.

Spanish Translation
Original: Estoy escribiendo para expresar mi insatisfacción.
Uncorrected: Estoy encendida para expresar mi insatisfacción.
OLMo Corrected: Estoy encendida para expresar mi insatisfacción.
LLaMA Corrected: Estoy escribiendo para expresar mi insatisfacción.
Qwen3 Corrected: Estoy escribiendo para expresar mi insatisfacción.