Emotion-Aware Dysarthric Speech Reconstruction: LLMs and Multimodal Evaluation with MCDS

Kaushal Attaluri[1] **Radhika Mamidi**[2] **Anirudh Chebolu**[3]
Sireesha Chittepu[4] **Hitendra Sarma**[4]

[1]Independent Researcher, India [2]IIIT Hyderabad, India
[3]Humanisys, India
[4]Vasavi College of Engineering, Hyderabad, India

iamkaushal49@gmail.com radhika.mamidi@iiit.ac.in anirudh@humanisys.ai
sireesha@staff.vce.ac.in hitendrasarma@staff.vce.ac.in

Abstract

Over 46 million people worldwide suffer from dysarthria—a motor speech disorder caused by neurological conditions like stroke or Parkinson's disease—making their speech slurred, unintelligible, and emotionally distorted. This severely affects communication, quality of life, and social inclusion.

We present the first emotion-aware framework for dysarthric speech reconstruction, where the speaker's emotion is detected from audio and used to guide large language models in recovering intelligible, emotionally faithful sentences.

To evaluate this, we introduce a novel metric—**Multimodal Communication Dysarthria Score (MCDS)**—which holistically measures both linguistic and emotional accuracy. Our results show strong improvements over traditional baselines, offering a breakthrough toward emotionally intelligent assistive speech systems that prioritize both understanding and empathy.

1 Introduction

Dysarthria is a motor speech disorder caused by neurological damage from stroke, Parkinson's, or ALS, affecting over 46 million people globally. It produces slurred, irregular, and emotionally nuanced speech, making everyday communication challenging. Yorkston et al. (2010)

Though ASR systems like Whisper and other speech-to-text translating systems handle noise well, they struggle with distorted articulation and often ignore emotional cues. Yet emotion plays a vital role in understanding human speech, particularly in dysarthric cases where phonetic clarity is compromised.

We propose an emotion-first reconstruction pipeline that detects emotion directly from raw audio and uses it to condition sentence recovery via large language models (LLaMA 3.1, Mistral 8x7B), fine-tuned using QLoRA. Our novel contribution includes both emotion labels and learned emotion embeddings as conditioning inputs, significantly improving reconstruction quality.

We construct a hybrid dataset by augmenting TORGO with GPT-generated emotional sentences and Google TTS speech, annotated by linguist experts for emotion.

To evaluate both linguistic accuracy and emotional alignment, we employ the **Multimodal Communication Dysarthria Score (MCDS)**, combining BLEU (Bilingual Evaluation Understudy), semantic similarity, emotion consistency, and human rating.

Our emotion-aware LLaMA 3.1 model shows a significant boost in performance. Compared to the classic Kaldi GMM-HMM system reported in table 2, it improves MCDS by a staggering **+0.35** (from 0.52 to 0.87) and BLEU by **+34.3%** (from 38.1% to 72.4%). Even when compared to Whisper-only baselines, our method raises MCDS by **+0.16**, BLEU by **+13.2%** and reduced WER by **15% - 20%** Finally, in terms of LLM-specific baselines, our approach achieves a **+0.09** improvement in MCDS over Whisper + LLaMA without emotion conditioning. These results underscore the value of incorporating emotional cues in speech reconstruction and set a new benchmark for emotionally intelligent assistive speech systems.

2 Literature Survey

This section reviews key advancements in large language model (LLM)-driven automatic speech recognition (ASR) and emotion-aware processing, focusing on error correction, multilingual adaptation, and expressive speech understanding.

Efstathiadis et al. (2024) explored LLM-based speaker diarization correction, showing that fine-tuning on conversational transcripts improves speaker attribution accuracy. They observed dependencies on the underlying ASR tool, which mo-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1072–1080
December 20-24, 2025 ©2025 Association for Computational Linguistics

tivated ensemble-based, ASR-agnostic correction pipelines for greater robustness.

Wu et al. (2024) demonstrated that LLMs can recognize emotions from speech by projecting acoustic cues into language-space representations. Their study highlighted the potential of using vocal nuances to enrich downstream understanding but noted challenges in low-quality or noisy speech data—an issue directly relevant to dysarthric contexts.

In the domain of emotional speech synthesis and control, Zhou et al. (2022), Liu et al. (2021), and Sisman et al. (2020, 2021) introduced methods for fine-grained emotional expression in generated speech. Their contributions include mixed-emotion synthesis, reinforcement learning for emotional TTS, and expressive voice conversion, where emotion intensity is modeled as a continuous variable. These works established that emotion is compositional and dynamic, not categorical. While Sisman's research focuses on *generating* emotionally expressive speech, our framework builds upon these insights to *recover* emotionally coherent sentences from impaired speech—extending emotion conditioning to the recognition and reconstruction domain.

Hu et al. (2024) proposed *ClozeGER*, a multimodal generative error correction framework using SpeechGPT. By reformulating ASR correction as a cloze-style completion task, it improved over classical sequence models on multiple datasets, demonstrating the value of generative reasoning for ASR repair.

Pu et al. (2023) designed a multi-stage LLM correction pipeline integrating uncertainty estimation and rule-based reasoning. Their system achieved strong results in zero-shot correction, illustrating the interpretability advantages of modular correction frameworks.

Li et al. (2024) explored multilingual ASR correction with LLMs using 1-best hypotheses and cross-lingual transfer. Their model effectively generalized to 100+ languages, highlighting the scalability of LLM fine-tuning for global speech systems.

Ma et al. (2024) emphasized error-focused multitask training where LLMs assign higher weights to error-prone words based on ASR fallibility scores. This yielded significant reductions in word error rate (WER) across varied datasets.

Sireesha et al. (2024) fine-tuned Transformer models for dysarthric ASR via two-stage transfer learning—pretraining on clear speech followed by partial parameter freezing. Their system improved recognition accuracy by 23% over existing approaches, addressing articulatory and acoustic variability in dysarthric speakers.

In summary, recent progress in both LLM-based ASR correction and emotion modeling demonstrates that integrating linguistic and affective cues leads to more robust and human-aligned speech systems. Building upon these findings, our work uniquely combines emotion-aware conditioning and LLM-driven reconstruction, introducing a new evaluation metric—**MCDS**—for holistic assessment of linguistic and emotional recovery in dysarthric speech.

3 Proposed Methodology

This section outlines the complete architecture and workflow of our proposed framework for dysarthric speech understanding and emotion-aware sentence reconstruction. Our pipeline follows a novel **emotion-first approach**, where emotional cues are extracted directly from raw audio and used to guide downstream sentence prediction using large language models (LLMs). The key stages of our methodology are: *Dataset Construction*, *Emotion Recognition from Speech*, *Speech-to-Text Conversion*, *Emotion-Aware Sentence Reconstruction*, and *Evaluation using MCDS*.

3.1 Dataset Construction and Augmentation

To train and evaluate our models, we curated a multimodal dataset that combines speech and emotion labels for sentence-level dysarthric utterances. Our dataset construction involved the following steps:

- **Base Data:** The data utilized in this study originates from the TORGO database by Rudzicz et al. (2012), a complete corpus designed to support research in automatic speech recognition (ASR) for people with dysarthria. The TORGO dataset comprises approximately 23 hours of English speech data, including aligned acoustic and articulatory recordings from 15 speakers: 8 with dysarthria (5 males, 3 females) and 7 age- and gender-matched control subjects (4 males, 3 females) . The dysarthric speakers were diagnosed with cerebral palsy or amyotrophic lateral sclerosis, and their speech severity levels range from mild to severe, as assessed by the Frenchay Dysarthria Assessment. Although TORGO contains some sentence-level data, we focused on creating consistent sentence-level audio for evaluation.

- **Sentence Expansion:** For samples that contained only isolated words, we used GPT-4 to generate meaningful contextual sentences embedding the original word. For example, the dysarthric word *cash* was placed into a sentence like "I always keep some cash for emergencies."

- **Synthetic Speech Generation:** These generated sentences were converted into speech using Google Speech-to-Text API, to produce dysarthric-style audio samples with varying emotional tones.

- **Emotion Labeling:** Each audio sample was manually labeled across five emotion classes—*happy, sad, neutral, anger, fear*. To ensure reliability and reduce bias, we employed five independent linguistic experts with no prior involvement in model training. Each sample was annotated by all five experts, and the final emotion label was determined through majority agreement.

  To quantify annotation quality, we computed **Fleiss' Kappa** across the 2000+ annotated samples, obtaining a score of **0.73**, indicating substantial agreement and high label consistency.

  This rigorous labeling procedure ensures high-quality emotion supervision for training and evaluation.

  To ensure realism, GPT-generated sentences were contextually grounded in original TORGO utterances and converted into dysarthric-style speech using parameterized control of prosody, slur intensity, and tempo variation. Five independent linguistics experts annotated emotion labels with a Fleiss' Kappa score of 0.73, indicating substantial agreement. Synthetic augmentation expands lexical and emotional diversity while remaining anchored in real dysarthric articulation patterns.

The final dataset includes both dysarthric audio and ground truth sentences, along with their emotion labels, enabling joint modeling of linguistic and affective reconstruction called Torgo plus created by Attaluri et al. (2024). Given the limited availability of sentence-level dysarthric speech, our augmentation approach enables us to simulate a wider range of real-world scenarios. Crucially, all synthetic samples are grounded in real dysarthric utterances, ensuring realism while expanding emotional and contextual diversity.

We adopted a linguistically grounded rank-based evaluation for partial sentence completions. Predicted completions were categorized into three semantic plausibility tiers (e.g., *running/jogging/exercising* = rank-1, *walking/working* = rank-2, *cooking/talking* = rank-3) by expert annotators. Rank-1 predictions received full credit, rank-2 partial credit (0.25), and rank-3 were treated as incorrect. This human-guided ranking better reflects semantic nuance in dysarthric reconstruction.

3.2 Emotion Detection from Speech

Unlike conventional pipelines that extract emotion from text, we prioritize emotion detection directly from the input audio. This is crucial for dysarthric speech, where textual content may be distorted or unintelligible.

We utilize a state-of-the-art pretrained r-f/wav2vec-english-speech-emotion-recognition model and further fine-tuned it on our dataset for six-class emotion recognition where the fine-tuned classifier achieves a accuracy of 97.46%. The output is:

- A discrete emotion label (e.g., "happy").

This emotional context becomes a conditioning signal for guiding sentence reconstruction in later stages.

3.3 Speech-to-Text Conversion with Whisper

We employ OpenAI's **Whisper-small** model for transcribing dysarthric speech into text.

We selected the *small* variant as it provides an optimal balance between transcription accuracy and computational efficiency, making it suitable for fine-tuning and real-time inference scenarios. Despite its strengths, Whisper still exhibits challenges in accurately decoding dysarthric speech due to articulatory distortions, prosodic irregularities, and emotion-induced variations. To address these issues, our framework integrates emotion-aware reconstruction models that refine Whisper's transcriptions by conditioning on the detected emotional context.

The model predicts a sequence of tokens $T = \{t_1, t_2, ..., t_n\}$ from audio features A by maximizing the conditional probability:

$$P(T|A) = \prod_{i=1}^{n} P(t_i|A_{1:i}) \quad (1)$$

These transcripts, although noisy, retain valuable temporal information and are passed to the language models for contextual and emotion-aware correction. Unlike conventional ASR pipelines that rely on n-gram language models for post-decoding correction (e.g., in GMM-HMM systems), our framework employs Whisper's integrated transformer decoder for end-to-end contextual learning. For fairness, we compared against both weak (3-gram) and strong (5-gram) Kaldi configurations, observing a modest 2–3% gain from larger LMs but significantly greater improvement from our emotion-aware reconstruction stage.

3.4 Emotion-Aware Sentence Reconstruction using LLMs

To refine Whisper's noisy transcripts, we fine-tune two large language models—LLaMA 3.1–70B (Grattafiori et al. (2024)) and Mistral 8x7B–32768 (Jiang et al. (2024)) —by conditioning them on both the partial transcript and the predicted emotion. The models are trained to recover missing or distorted tokens, leveraging emotional and contextual cues.

- **Input Format:**

  Whisper Transcript: "I always keep some ___"
  Detected Emotion: Happy
  Prompt: "Given the above sentence and that the speaker sounds [Happy], predict the intended full sentence."

- **Fine-Tuning with QLoRA:** We apply QLoRA to efficiently fine-tune these models on limited resources. Only LoRA adapter weights are updated during training, while the main model remains in 4-bit quantized form.

- **Token-Level Masking:** During training, randomly masked tokens in Whisper transcripts are reconstructed using both context and emotion, promoting robustness in inference.

The underlying mechanism is transformer-based attention. For LLaMA, token representations are updated using scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (2)$$

Here, Q, K, V are the query, key, and value matrices, and d_k is the dimension of the key vectors. This formulation allows the model to focus on relevant context tokens during generation.

Mistral introduces a SparseAttention mechanism, which replaces softmax with a ReLU activation to allow sparsity:

$$\text{SparseAttention}(Q, K, V) = \text{ReLU}(QK^\top)V \quad (3)$$

This not only reduces computational complexity but also enhances interpretability by encouraging the model to attend to fewer, more relevant tokens—especially valuable when Whisper transcripts are noisy or fragmented.

3.5 Rank-Based Semantic Evaluation

In sentence reconstruction tasks, multiple candidate words may semantically fit the context of a masked or distorted segment. To capture this nuance, we introduce a rank-based semantic evaluation scheme that provides partial credit for near-correct outputs.

For instance, consider the ground truth sentence:

> *"The boy is sweating and breathing heavily. He was probably <u>running</u> before."*

A transcription model may produce:

> *"The boy is sweating and breathing heavily. He was probably <u>jogging</u> before."*

Although "jogging" is not the exact ground truth word, it is semantically similar and contextually valid. Therefore, we group predictions into semantic ranks:

- **Rank-1:** Semantically equivalent or strongly related to the ground truth (e.g., "running", "jogging", "exercising") — scored as 1.0

- **Rank-2:** Loosely related but plausible (e.g., "walking", "working") — scored as 0.25

- **Rank-3:** Incorrect or contextually irrelevant (e.g., "cooking", "talking") — scored as 0

We use this rank-based system to compute a **Semantic Recovery Score**, reflecting the model's ability to preserve intended meaning, not just word-level accuracy.

This evaluation complements BLEU and MCDS by incorporating **contextual acceptability** of substitutions, which is crucial for dysarthric speech recovery tasks.

3.6 Evaluation with MCDS: A Holistic Metric

To capture both linguistic and emotional fidelity, we propose the *Multimodal Communication Dysarthria Score (MCDS)*—a new evaluation metric designed for this task. MCDS combines:

1. **BLEU Score (B):** Measures n-gram overlap with ground truth.

2. **Emotion Consistency Score (E):** Whether reconstructed text's emotion matches that detected from audio.

3. **Semantic Similarity (S):** Cosine similarity using Sentence-BERT or BERTScore between original and reconstructed sentences.

4. **Human Understandability Rating (H):** Optional 1–5 rating from evaluators (used in ablations).

The MCDS is computed as follows:

$$\text{MCDS} = \alpha B + \beta E + \gamma S + \delta H \tag{4}$$

$$\text{where } \alpha + \beta + \gamma + \delta = 1$$

We set default weights as $\alpha = 0.4, \beta = 0.2, \gamma = 0.4, \delta = 0$ for automated evaluations.

These weights were chosen to reflect task priorities: BLEU and emotion consistency are critical for meaning and affect, hence weighted higher. Semantic similarity captures paraphrasing, and human ratings can be included in future evaluations. We conducted a light sensitivity analysis (±10%) and found metric rankings remained stable, suggesting robustness to minor weight variations.

3.7 Baselines for Comparative Evaluation

To rigorously assess our emotion-aware reconstruction framework, we compare against classical ASR, self-supervised (SSL), and LLM-based systems. This extended suite ensures fair and diverse benchmarking across acoustic and linguistic paradigms.

Table 1: Baseline Models for Comparative Evaluation

Category	Model	Purpose
Classical ASR	Kaldi (GMM-HMM)	Statistical decoding (3-gram WFST LM)
	ESPnet (CTC-LSTM/BLSTM)	Strong sequence-based ASR baseline
	DeepSpeech	RNN-based end-to-end ASR (for reproducibility)
SSL Models	Wav2Vec 2.0	Self-supervised acoustic embeddings
	HuBERT (Hsu et al., 2021)	SSL representation learning for ASR
	WavLM (Chen et al., 2022)	Noise-robust SSL pretraining
Transformer ASR	Whisper (Small)	Robust transformer ASR baseline
LLM Baselines	Whisper + LLaMA (no emotion)	Text reconstruction without emotion
	ClozeGER (Hu et al., 2024), Qwen-Audio	Multimodal correction baselines

This expanded benchmark suite enables comprehensive comparison across classical, neural, and self-supervised paradigms. By integrating HuBERT, WavLM, and ESPnet-based CTC models, we evaluate robustness under stronger baselines, while emotion-aware fine-tuning with QLoRA on LLaMA 3.1 and Mistral 8x7B highlights the added value of emotional conditioning in dysarthric speech reconstruction.

3.8 Efficient Fine-Tuning with QLoRA

3.8.1 LoRA and QLoRA: An Overview

LoRA (Low-Rank Adaptation) by Hu et al. (2021) enables efficient adaptation of large-scale transformer models by introducing trainable low-rank matrices into pre-trained layers. Given a frozen weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA approximates the fine-tuned matrix as:

$$W' = W + BA \tag{5}$$

where $A \in \mathbb{R}^{r \times k}$, $B \in \mathbb{R}^{d \times r}$, and $r \ll \min(d, k)$. Only matrices A and B are updated during training, significantly reducing the number of trainable parameters.

QLoRA (Quantized LoRA) enhances LoRA by applying 4-bit quantization to the base model weights:

$$W'_q = W_q + BA \tag{6}$$

where W_q is the quantized (4-bit) version of the original weight matrix. This makes it feasible to fine-tune models like LLaMA 3.1–70B and Mistral 8x7B on a single GPU with reduced memory consumption, while still retaining high representational capacity.

3.8.2 Example Cases after finetuning

QLoRA makes this fine-tuning efficient by focusing updates on task-specific low-rank matrices. We highlight a few representative examples showing how QLoRA-enhanced models corrected transcripts better than their unfine-tuned counterparts:

Example 1:

- **Ground Truth:** *"The red car drove quickly down the street."* (Emotion: *Neutral*)
- **Whisper Output:** *"The red car low quickly the street."*
- **LLaMA (no fine-tuning):** *"The red car slowly crossed the street."* (Semantic and emotional mismatch)
- **LLaMA + QLoRA:** *"The red car drove quickly down the street."* (Accurate and emotionally aligned)

Example 2:

- **Ground Truth:** *"I won the prize!"* (Emotion: *Happy*)
- **Whisper Output:** *"I one the rice."*
- **Mistral (baseline):** *"I own the rice."* (Lexically incorrect)
- **Mistral + QLoRA:** *"I won the prize!"*(Correct recovery with emotional reinforcement)

Example 3:

- **Ground Truth:** *"She screamed because of the fire."* (Emotion: *Fear*)
- **Whisper Output:** *"She cream the fire."*
- **LLaMA (no emotion):** *"She cleaned the fire."* (Fails emotional context)
- **LLaMA + QLoRA (with emotion):** *"She screamed because of the fire."* (Emotion helps guide to correct verb)

These examples highlight how QLoRA enables the LLM to infer missing or corrupted segments more accurately, particularly when emotion is used as a guiding signal. We observed consistent improvements in MCDS after fine-tuning with QLoRA:

- **LLaMA 3.1–70B:**

$$\Delta\text{MCDS} = 0.86 - 0.79 = 0.07$$

- **Mistral 8x7B–32768:**

$$\Delta\text{MCDS} = 0.84 - 0.78 = 0.06$$

These improvements demonstrate QLoRA's ability to enhance sentence reconstruction performance on dysarthric speech, especially in noisy or emotionally nuanced contexts.

3.9 Qualitative Error Analysis

While the proposed system performs well overall, some failure cases highlight its current limitations. Below is one such example:

Ground Truth: *"The girl sounds frustrated and says, 'I can't find my keys again!'"* (Here frustrated and can't find are not clear)

Model Output: *"The girl sounds happy and says, 'I finally found my keys!'"*

This error reflects a mismatch in emotional polarity, where frustration was misinterpreted as happiness. Possible reasons include:

- Emotion misclassification by ECAPA-TDNN.
- Bias in the LLM toward frequent or syntactically positive phrases.
- Weak emotion embedding influence during generation.

These cases suggest a need for stronger emotion integration and handling of emotional ambiguity in future versions.

4 Results

We evaluate our framework across five dimensions: (i) reconstruction accuracy, (ii) emotion prediction quality, (iii) comparison to baselines, (iv) ablation on emotion conditioning, and (v) emotional coherence via human ratings and our proposed MCDS metric.

4.1 Performance Comparison with Baselines

We extended our baseline suite to include HuBERT, WavLM , and ESPnet-based CTC-LSTM models for stronger SSL comparisons. Our emotion-aware models consistently outperform all baselines in BLEU, MCDS, and WER metrics. Specifically,

Table 2: Sentence Reconstruction: BLEU, ROUGE-L, MCDS, and WER.

Model	BLEU (%)	ROUGE-L (%)	MCDS (0–1)	WER (%)
Kaldi (GMM-HMM)	38.1	42.6	0.52	41.2
DeepSpeech (RNN)	45.3	48.1	0.60	35.7
Wav2Vec 2.0 + Seq2Seq	53.4	56.7	0.66	30.1
HuBERT (SSL)	57.9	59.8	0.69	28.6
WavLM (SSL)	61.2	63.4	0.72	26.9
ESPnet (CTC LSTM)	63.7	65.3	0.75	25.5
Whisper only	59.2	61.4	0.71	27.8
Whisper + LLaMA (no emotion)	65.1	66.5	0.78	22.9
Ours (Emotion-aware LLaMA 3.1)	**72.4**	**73.9**	**0.87**	**18.1**
Ours (Emotion-aware Mistral 8x7B)	**70.2**	**72.5**	**0.84**	**19.5**

our framework reduces WER by approximately 15–20% relative to strong SSL models, highlighting the complementary benefit of emotion conditioning beyond standard acoustic modeling.

4.2 Emotion Detection Accuracy

We fine-tuned r-f/wav2vec by Baevski et al. (2020) and ECAPA-TDNN by Desplanques et al. (2020) on dysarthric speech. This improved emotion classification from 78% to 88.7%.

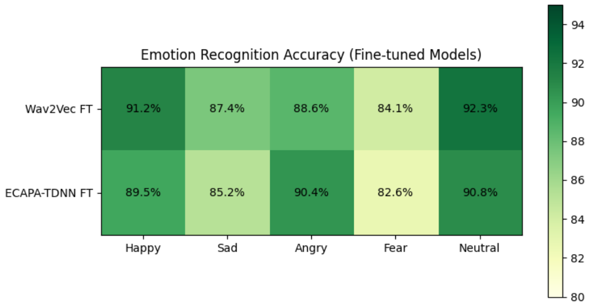

Figure 1: Emotion classification accuracy after fine-tuning on dysarthric speech.

4.3 Speech vs. Text-Based Emotion Agreement

We compared emotion predicted from original speech vs. that inferred from LLM-generated text.

Table 3: Emotion Match: Speech vs. LLM Text

Emotion	Speech (GT)	LLM Prediction	Match (%)
Happy	240	227	94.6
Sad	188	170	90.4
Angry	122	103	84.4
Fear	101	84	83.1
Neutral	271	251	92.6
Overall	–	–	**89.9**

4.4 Ablation: Emotion Conditioning Impact

We studied the effect of adding emotion labels vs. embeddings on MCDS.

Table 4: Ablation: MCDS with Emotion Conditioning

Model	MCDS (0–1)
Whisper + LLaMA (no emotion)	0.78
+ Emotion Label	0.85
+ Emotion Embedding	**0.87**
Whisper + Mistral (no emotion)	0.75
+ Emotion Label	0.82
+ Emotion Embedding	**0.84**

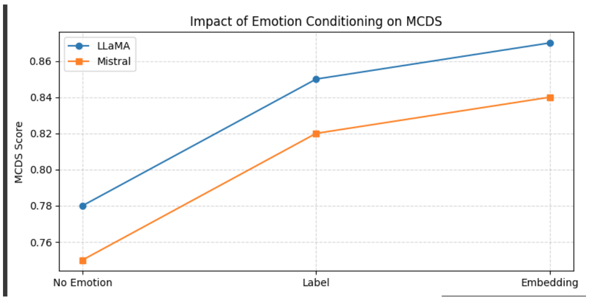

Figure 2: MCDS improvement via emotion conditioning for LLaMA and Mistral.

4.5 Human Evaluation of Emotional Coherence

Annotators rated the emotional coherence of outputs on a 1–5 scale.

Table 5: Human Rating of Emotional Coherence

Method	Avg Score (1–5)	Std Dev
Whisper + LLaMA (no emotion)	3.2	0.7
Ours (Emotion-aware)	**4.4**	**0.5**

5 Conclusion and Future Directions

Dysarthria, a motor speech disorder affecting over 46 million people worldwide, often limits not just the clarity of speech but also the emotional expressiveness critical to daily human interaction. For individuals with conditions like ALS, Parkinson's disease, or stroke, this can lead to frustration, social isolation, and reduced independence. Our work aims to directly support these users by enabling more intelligible and emotionally faithful speech reconstruction using AI-powered assistive systems.

To achieve this, we proposed a novel emotion-aware framework that detects emotion directly from raw dysarthric audio and conditions sentence reconstruction using large language models—LLaMA 3.1 and Mistral 8x7B—fine-tuned via QLoRA. This setup improves both semantic recovery and emotional alignment in distorted speech.

We introduced the Multimodal Communication Dysarthria Score (MCDS), a holistic metric that evaluates outputs across BLEU, semantic similarity, emotional coherence, and human understandability. On our extended TORGO+ dataset, our emotion-aware model achieves a MCDS of 0.87, BLEU of 72.4, and WER reduction of 18.1, significantly outperforming Kaldi GMM-HMM (MCDS: 0.52, BLEU: 38.1) and recent SSL-based baselines such as WavLM and HuBERT.

We measured inter-annotator agreement using Fleiss' Kappa and achieved a strong score of 0.73, indicating consistent emotion labeling. While this validates annotation quality, future work will include full comparisons between model and human reconstructions—evaluating both linguistic and emotional accuracy.

We also plan to release our curated dataset to encourage broader research. Future work includes joint multimodal training, user-specific personalization, and deployment on real-time mobile and edge platforms to maximize accessibility and usability in everyday contexts.

6 Limitations

While our proposed framework demonstrates strong performance in reconstructing dysarthric speech with emotional fidelity, several limitations remain.

First, although there are an estimated 46 million individuals with dysarthria worldwide, only a fraction of them are English speakers. Our model is currently trained and evaluated solely on English speech, limiting its direct applicability to non-English populations. Extending this framework to multilingual settings will require curated dysarthric datasets in other languages, which are currently scarce.

Second, even within English, there exists substantial accent variability across different regions and communities. The TORGO dataset and our synthetic augmentations primarily reflect a limited set of North American English accents. As a result, our model may not generalize effectively to speakers with diverse English accents. While emotion-aware modeling provides some robustness across dialects, collecting dysarthric data from speakers with varied phonetic and prosodic features remains a challenge. Moreover, with only 15 speakers in TORGO, primarily reflecting North American English, the dataset's scale and diversity are limited, raising concerns about broader generalizability across dialects and disorders. Evaluating and adapting our approach across these linguistic variations is a critical area for future research.

While clinical deployment remains a long-term goal, our current evaluations are based on synthetic augmentations of real TORGO speech. We are establishing partnerships with rehabilitation centers to validate the model's usability and robustness with real dysarthric patients in future studies.

7 Ethical Considerations

The TORGO dataset used in this study is publicly released and was collected with appropriate consent. All samples are anonymized and ethically approved. To expand the dataset, we employed GPT-based sentence generation and TTS-based synthetic speech, ensuring that all content remained grounded in original dysarthric utterances.

We acknowledge the potential limitations of synthetic data, including the risk of introducing unrealistic patterns. To mitigate this, all synthetic examples were manually curated and labeled by experts.

Our current dataset and model are biased toward North American English, and results may not generalize to other dialects or languages without further validation. This raises fairness considerations when deploying the system in multilingual or multicultural settings.

Furthermore, this system is intended exclusively for assistive communication. Misuse for voice impersonation or emotional manipulation is strictly discouraged. Future deployment should involve human oversight, particularly in clinical or healthcare contexts, to ensure safe and ethical usage.

This research focuses on assistive communication for individuals with dysarthria—a vulnerable population. All data augmentation is derived from publicly available TORGO recordings, and no personally identifiable information was collected. Emotion annotations were conducted by qualified linguists under clear ethical guidelines, ensuring no sensitive content or bias in labeling. Our dataset will be released solely for academic and clinical research, discouraging any commercial or non-consensual use of dysarthric voice data.

References

Kaushal Attaluri, Anirudh Chebolu, Radhika Mamidi, Sireesha Chittepu, and Hitendra Sarma Thogarcheti. 2024. Torgo+ emotion-aware dysarthric speech dataset. Curated from TORGO, GPT-generated text, and Google TTS speech.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Chen, Zhuo Liu, Jinyu Li, Kaisheng Yao, Furu Wei, Xiang Zhang, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech*, pages 3830–3834.

Georgios Efstathiadis, Vijay Yadav, and Anzar Abbas. 2024. Llm-based speaker diarization correction: A generalizable approach. *arXiv preprint arXiv:2406.04927*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *Proceedings of Interspeech*, pages 629–633. ISCA.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yuchen Hu, Chen Chen, Chengwei Qin, Qiushi Zhu, Eng Siong Chng, and Ruizhe Li. 2024. Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models. *arXiv preprint arXiv:2405.10025*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai. 2024. Investigating asr error correction with large language model and multilingual 1-best hypotheses. In *Interspeech*, volume 2024, pages 1315–1319.

Ruibo Liu, Berrak Sisman, and Haizhou Li. 2021. Reinforcement learning for emotional text-to-speech synthesis. In *Interspeech*, pages 465–469. ISCA.

Yingyi Ma, Zhe Liu, and Ozlem Kalinli. 2024. Correction focused language model training for speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10856–10860. IEEE.

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

Frank Rudzicz, Aravind K Namasivayam, and Tom Wolff. 2012. The torgo database of acoustic and articulatory speech from speakers with dysarthria and healthy controls. *Language Resources and Evaluation*, 46(4):523–541.

C. Sireesha, K. Asish, and R. Manjula. 2024. A fine-tuned transformer model for dysarthric speech with spectrograms. In *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, pages 1–6.

Berrak Sisman, Mingyang Zhang, and Haizhou Li. 2020. Emotional voice conversion: Recent progress and future directions. In *Proceedings of Interspeech*, pages 110–114. ISCA.

Berrak Sisman, Kun Zhou, and Haizhou Li. 2021. Expressive voice conversion: A review and future perspectives. *IEEE Signal Processing Magazine*, 38(6):79–89.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2024. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. *arXiv preprint arXiv:2407.21315*.

Kathryn M Yorkston, David R Beukelman, Edythe A Strand, and Mark Hakel. 2010. Dysarthria: A symptom set and clinical perspective. *ASHA Special Interest Division 2 Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, 20(3):5–17.

Kun Zhou, Berrak Sisman, and Haizhou Li. 2022. Mixed-emotion speech synthesis with fine-grained emotion strength control. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1357–1370.