Review of: "Empowering Dysarthric Speech: Leveraging Advanced LLMs for Accurate Speech Correction and Multimodal Emotion Analysis"

Elakkiya R[1]

[1] Birla Institute of Technology and Science Pilani, Pilāni, India

Potential competing interests: No potential competing interests to declare.

This manuscript describes a framework using LLMs to improve the interaction of dysarthric patients. Using OpenAI's Whisper model, the authors transcribe dysarthric speech and then perform sentence prediction using fine-tuned LLaMA 3.1 and Mistral 8x7B models with emotion recognition.

Strengths:

1. The authors investigate the application of state-of-the-art LLMs to perform both speaker-dependent correction and emotion analysis of dysarthric speech, contributing to a major sense of empowerment and liberty of communication for those with the condition.
2. The work combines various techniques such as speech-to-text conversion, sentence prediction, and emotion recognition to enhance the communication experience for dysarthric users.
3. The authors took the TORGO dataset, filtered the voice samples with Google Speech data, and manually labelled emotional contexts, leading to a custom dataset for training and testing. The manuscript discloses accuracy scores for sentence prediction and emotion recognition but does not provide a comprehensive statistical evaluation.

Improvements:

1. The custom dataset contains a mixture of the TORGO dataset and Google Speech data but lacks diversity in accents, speech patterns, and most importantly, languages. This restriction may impact the models' ability to be generalized to a wider population of users of dysarthric speech.
2. The paper reports accuracy without providing detailed statistical data like confidence intervals, precision, recall, or F1-scores. The absence of other metrics to assess multiple aspects of performance dilutes the strength of the performance evaluation.
3. By not providing comparisons of the proposed method with existing state-of-the-art methods for dysarthric speech correction or for emotion recognition, the manuscript is less informative. This lack of comparative context limits our ability to evaluate the true uniqueness and benefits of the authors' proposed framework.
4. Though it is novel to leverage fine-tuned LLMs such as LLaMA and Mistral, the problems surrounding the deployment compute of it are not solved. The description of whole environments challenges the practical feasibility of the framework, especially for low-cost assistive devices.

5. There is no error analysis in the manuscript to see what kind of errors the system generates in correcting the speech or recognizing the emotion. Knowledge of such limitations is essential for enhancing the system's effectiveness and dependability.
6. Though the framework combines speech correction and emotion recognition, the relationship between the two and how they influence each other is not statistically investigated.
7. The differences are in the data on which the models are evaluated, on datasets curated but not tested in real-world scenarios, e.g., real-time deployment on assistive devices. This gap calls into question the framework's practical utility.