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

Caleb Rascon[1]

1 Universidad Nacional Autónoma de México, Mexico

Potential competing interests: No potential competing interests to declare.

The authors provide a speech recognition system focused on dysarthric speech by fine-tuning several large language models, which also provides emotional context.

The manuscript has the following issues:

- The generalizability of the system is put into question. First off, each user/patient is expected to suffer from dysarthria in a different manner. Additionally, the training dataset is based on the Google Speech API, which provides a limited amount of voices from which to train, hinting at a high probability of overfitting to those specific voices. Thus, the authors need to acknowledge and establish the current level of generalizability of their system as part of the manuscript. This could be done by either testing with voices that were not in the training data or using other datasets that insert other types of speech effects from dysarthria.

- Furthermore, the authors mention in several parts their intent to use this system in multi-language scenarios. Unfortunately, emotional context is difficult to generalize between languages. The authors should test on the following datasets to validate results in multi-language scenarios:

Cao, Houwei, et al. "Crema-d: Crowd-sourced emotional multimodal actors dataset." IEEE Transactions on Affective Computing 5.4 (2014): 377-390.

Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS One 13.5 (2018): e0196391.

Adigwe, Adaeze, et al. "The emotional voices database: Towards controlling the emotion dimension in voice generation systems." arXiv preprint arXiv:1806.09514 (2018).

- Given that the main objective is to empower patients with dysarthria, it can be assumed that their system is to be used in real-time scenarios. Thus, the authors should provide the expected response times of their system, as well as the computational specifications (CPU number of cores and speed, GPU models, memory, etc.) required to achieve them.

- In Section 3.1, the authors should provide more information about the Torgo dataset that they used, such as the number of users, the number of dysarthric words, etc., as well as an actual reference for such a dataset.

- The authors state (page 7) that "Overall, GPT-4.o gave us good accuracy when compared to other Large Language Models when integrated into our workflow for predicting the correct sentence and recognizing the emotion behind the sentence, delivering a high inclusive communication solution for dysarthric users." They also state (page 8) that "LLaMA 3.1-70B's advances in efficiency, enhanced training data, and adaptability, combined with our fine-tuning process, allowed the model to outperform its predecessors in handling the complexities of dysarthric speech." The authors should provide these results as part of the manuscript.

- The authors state (page 12) that "For the emotion recognition task, we manually labeled the emotional context of each sentence in our dataset." How was the labeling process carried out? How many persons were tasked to label one sentence? If only one, was it verified by a psychological expert? If more than one, was there consensus between the different labelers?

- It seems that the emotional context was estimated from the text, not the audio signal. This reviewer believes there's much more information that can be used directly from the audio signal (prosody, frequency variation through time, rhythm, etc.) for emotion recognition. The authors should acknowledge if this is something that could be used in future versions of their system or justify its absence in their work.

-- Typos:

- abstract: "reconsturcts" -> "reconstructs"
- page 4: "the voice was could not be detected" -> "the voice could not be detected"
- page 5: "The great advantage of whisper is, it can convert speech to text in multiple languages, It is multilingual and it is open source." -> "The great advantage of whisper is it can convert speech to text in multiple languages and it is open source."
- page 7: "on hugging face, the reason we used" -> "on hugging face. The reason we used"
- page 9: "(Low-Rank Adaptation) [?]" reference missing