Qeios

Peer Review

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

Sebastian Roldan-Vasco[1]

1. Instituto Tecnológico Metropolitano, Colombia

General

The dataset was not described. Demographic information was not provided. This information is mandatory, even if you have used a public database. Otherwise, it is not possible to figure out the generalization capability of your results.

It's not clear what the real contribution of this work is, unless you compare the baseline performance of other models with dysarthric patients.

Additionally, you are evaluating intelligibility, but you use a very general performance measure (accuracy). In this kind of work, one expects to find problem-specific measures like Word Error Rate, Phoneme Error Rate, Short-Time Objective Intelligibility, Mean Opinion Score, etc. The only use of accuracy lacks rigor.

All the results are reported in individuals with an unknown level of dysarthria impairment, which hinders how generalizable (and clinically worthy) this work is. This kind of total approach (one single class with dysarthric individuals) ignores several clinical characteristics that should be analyzed in a paper that tries to contribute to the field of dysarthria assessment.

1. Introduction

There are no references.

2. Literature survey

Incorrect way to cite: The paper [1]..., This article [2]..., In this work [6] ...

The literature is insufficient. It's important to mention works that have addressed the automatic detection of dysarthria, not only the use of LLM for ASR.

There is a repeated paragraph.

The affirmations in this section are too vague.

Is the paper [6] yours? Why did you use "we"? Additionally, I don't get the point when summarizing the future work of other authors' papers. This information is irrelevant to your work.

doi.org/10.32388/N8S17E

The last paragraph is too general; focus on your dysarthria-related problem.

3. Methodology

Fig 1. The architecture was not described in detail.

You mentioned that an advantage of Whisper is the multilingual capability. Did you use audio from non-English speakers?

It seems like you used two ways to obtain text: Google Speech and Whisper. Why? It is confusing.

3.4 "Performance in our case": the reported accuracy should be moved to the results section. The same applies to the Mixtral architecture. Furthermore, why did you not report the performance in the GPT-4 model? You only mentioned that GPT-4 "gave us good accuracy." It's confusing. Additionally, it's not clear what the meaning of such accuracies is: are they related to emotion or dysarthria detection?

Is the goal of the work to show differences between LLM for dysarthric patients?

Describe how matrices A and B are computed.

The subsection 3.6.5 is irrelevant.

In section 3.6.4, you mentioned the performance for emotion detection, but you explain this task in the subsequent section 3.7. Maintain the order and coherence of the writing: explain the methodology prior to the achieved results.

It was not clear in the paper how automatic emotion recognition can be useful for dysarthria evaluation. It seems like an unrelated experiment with the same database.

5. Tables

Table 5.1: it's not clear if the shown sentence per model is such that achieved the highest accuracy. What do the bold words mean?

Declarations

Potential competing interests: No potential competing interests to declare.