**Reviewer #1**

**Questions**

* **2. Reviewer confidence By selecting a value, you confirm that you are competent to evaluate this paper. If you have doubts about your ability to assess it, do not complete this form but instead use "Contact Meta-Reviewer".**
  + 2: Confident
* **3. Originality score**
  + 1: Not novel
* **4. Comments on originality Please, comment on the originality of the work. Are the tasks or methods new or a novel combination of previously-known techniques? Is the work a description of a new publicly released dataset? Is related work adequately cited and the new contributions explained?**
  + The paper proposes a method using wav2vec 2.0 and phoneme embedding for Vietnamese Mispronunciation Detection, achieving sota performance in the VLSP Vietnamese Mispronunciation Detection 2023 challenge. This work demonstrates the best performance on a recent dataset in the field.   
      
    However, the paper does not introduce very innovative concepts; it combines well-performing components such as Wav2Vec 2.0 and optimizes it in the context of the VLSP challenge.
* **5. Technical correctness score**
  + 4: Technically solid
* **6. Comments on technical correctness Please evaluate the scientific and/or technical correctness of the work. If experiments are presented please consider if enough details are provided on the datasets, baseline and experimental design to allow the experiments to be reproduced or equivalent experiments run. Are measures of statistical significance presented? Are the author's responses to the checklist satisfactory?**
  + The paper is technical solid. It provides detailed explanations of datasets, metrics, experimental configurations, and extensive baselines, enabling reproducibility and comparison with equivalent experiments. The analysis and ablation studies are presented clearly, enhancing the understanding of the methodology and results.
* **7. Clarity of presentation score**
  + 3. Clear enough, could benefit from some revision
* **8. Comments on clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing, organization, and quality of figures, tables etc. The English does not need to be flawless, but the text should be understandable.**
  + The text is well written and figure is clear
* **10. Overall recommendation**
  + 2: Weak Reject: I am leaning to reject this paper
* **11. Summary Please, summarise in one or two sentences the main motivation for your overall recommendation.**
  + The paper is technical solid, but limited innovations

**Reviewer #2**

**Questions**

* **2. Reviewer confidence By selecting a value, you confirm that you are competent to evaluate this paper. If you have doubts about your ability to assess it, do not complete this form but instead use "Contact Meta-Reviewer".**
  + 3. Very Confident
* **3. Originality score**
  + 4: Very novel
* **4. Comments on originality Please, comment on the originality of the work. Are the tasks or methods new or a novel combination of previously-known techniques? Is the work a description of a new publicly released dataset? Is related work adequately cited and the new contributions explained?**
  + This is the first time I see the addition of the canonical phoneme sequence to an end2end system for MDD
* **5. Technical correctness score**
  + 4: Technically solid
* **6. Comments on technical correctness Please evaluate the scientific and/or technical correctness of the work. If experiments are presented please consider if enough details are provided on the datasets, baseline and experimental design to allow the experiments to be reproduced or equivalent experiments run. Are measures of statistical significance presented? Are the author's responses to the checklist satisfactory?**
  + no comments
* **7. Clarity of presentation score**
  + 3. Clear enough, could benefit from some revision
* **8. Comments on clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing, organization, and quality of figures, tables etc. The English does not need to be flawless, but the text should be understandable.**
  + There are many English typos in the paper. Althought I am tempted to say that this is nice to see in the era of GPTs, still I believe that it would be more readable if corrected.
* **9. Additional comments Please, provide any additional or final comments about the paper. You may leave this section empty if you have already provided all your comments in previous sections.**
  + In section 2 I see lots of decisions taken on architecture that are not motivated well. It would be great if you could add, wherever possible, some info on why each decision is taken.
* **10. Overall recommendation**
  + 4: Accept: I think this paper should be accepted
* **11. Summary Please, summarise in one or two sentences the main motivation for your overall recommendation.**
  + I enjoyed reading the paper, thank you.  
    I like the proposed algorithm. I find that the way the end2end system is constrained on the canonical phoneme sequence is very clever and brings the powerful end2end models closer to old style Viterbi decoding using a decoding graph for forced alignment.   
    As next step, please think how you can enter transcription variations instead of just 1 single canonical transcription.

**Reviewer #5**

**Questions**

* **2. Reviewer confidence By selecting a value, you confirm that you are competent to evaluate this paper. If you have doubts about your ability to assess it, do not complete this form but instead use "Contact Meta-Reviewer".**
  + 2: Confident
* **3. Originality score**
  + 2: Minor novelty
* **4. Comments on originality Please, comment on the originality of the work. Are the tasks or methods new or a novel combination of previously-known techniques? Is the work a description of a new publicly released dataset? Is related work adequately cited and the new contributions explained?**
  + The novelty of this paper is limited. But this paper gives very detailed results and is useful for other researchers.
* **5. Technical correctness score**
  + 3: Minor issues but credible results
* **6. Comments on technical correctness Please evaluate the scientific and/or technical correctness of the work. If experiments are presented please consider if enough details are provided on the datasets, baseline and experimental design to allow the experiments to be reproduced or equivalent experiments run. Are measures of statistical significance presented? Are the author's responses to the checklist satisfactory?**
  + This paper gives very clear description of their experimental setups. They obtained SOTA results on their task.   
    Q1: This paper improved the FFN module and experimental result shows that it is very helpful for this task, but this paper gives little explanation why it works. I also wonder if this improvement works well for other task.
* **7. Clarity of presentation score**
  + 4. Very well written
* **8. Comments on clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing, organization, and quality of figures, tables etc. The English does not need to be flawless, but the text should be understandable.**
  + This paper is well written.
* **10. Overall recommendation**
  + 3: Weak Accept: I am leaning to accept this paper
* **11. Summary Please, summarise in one or two sentences the main motivation for your overall recommendation.**
  + The novelty of this paper is limited. But the results are great and the authors give very detailed experimental verification.

**Reviewer #6**

**Questions**

* **2. Reviewer confidence By selecting a value, you confirm that you are competent to evaluate this paper. If you have doubts about your ability to assess it, do not complete this form but instead use "Contact Meta-Reviewer".**
  + 2: Confident
* **3. Originality score**
  + 2: Minor novelty
* **4. Comments on originality Please, comment on the originality of the work. Are the tasks or methods new or a novel combination of previously-known techniques? Is the work a description of a new publicly released dataset? Is related work adequately cited and the new contributions explained?**
  + The authors claim to have extended some wav2vec software   
    for pronunciation error detection.  
    For me it is hard to see what has been done because the authors  
    do not give an simple and clear information of what they measure.
* **5. Technical correctness score**
  + 2: Significant flaws that may call results into question
* **6. Comments on technical correctness Please evaluate the scientific and/or technical correctness of the work. If experiments are presented please consider if enough details are provided on the datasets, baseline and experimental design to allow the experiments to be reproduced or equivalent experiments run. Are measures of statistical significance presented? Are the author's responses to the checklist satisfactory?**
  + I just do not know, because I could not understand what the authors did. There is not a single statement about acoustic  
    or linguistic conditions (e.g. audio sampling rate, frame rate,  
    phoneme vocabulary). Are the authors talking about free recognition or does the system know the text that has been spoken?
* **7. Clarity of presentation score**
  + 1. Difficult to follow
* **8. Comments on clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing, organization, and quality of figures, tables etc. The English does not need to be flawless, but the text should be understandable.**
  + The description of what the authors did is masked by the jargon  
    on neural network software. There are no principal statements  
    about the task formulation. What are the training data? Test data/  
    What is the typical length of an utterance?  
    How are the pronunciation errors measured?
* **10. Overall recommendation**
  + 1: Reject: I think this paper should be rejected
* **11. Summary Please, summarise in one or two sentences the main motivation for your overall recommendation.**
  + I suppose that the speakers are asked to read a specific text,  
    and then based on their voice signal, the pronunciation errors  
    are measured. This interpretation is not clear at all from the text.  
    If this is the case, then this task is only of very much limited interest to the IS community. In addition, the presentation is so unclear that not much can be learned from the paper.

**Reviewer #7**

**Questions**

* **2. Reviewer confidence By selecting a value, you confirm that you are competent to evaluate this paper. If you have doubts about your ability to assess it, do not complete this form but instead use "Contact Meta-Reviewer".**
  + 2: Confident
* **3. Originality score**
  + 2: Minor novelty
* **4. Comments on originality Please, comment on the originality of the work. Are the tasks or methods new or a novel combination of previously-known techniques? Is the work a description of a new publicly released dataset? Is related work adequately cited and the new contributions explained?**
  + The paper proposes a minor change in wav2vec 2.0 by combining it with a linguistic encoder where both systems feed into an additional module consisting of a multi-head attention followed by a feedforward network. The authors claim that these changes are expected to provide a balanced trade-off between having the correct phoneme sequence output from the canonical phoneme and having the right pronunciation scores.  
    These changes are derived from well-known techniques but are applied in somewhat a novel way to solve the Mispronunciation Detection problem.
* **5. Technical correctness score**
  + 3: Minor issues but credible results
* **6. Comments on technical correctness Please evaluate the scientific and/or technical correctness of the work. If experiments are presented please consider if enough details are provided on the datasets, baseline and experimental design to allow the experiments to be reproduced or equivalent experiments run. Are measures of statistical significance presented? Are the author's responses to the checklist satisfactory?**
  + The papers focuses heavily on the engineering of the system while rarely providing good arguments or a clear motivation behind the proposed approach or the choices behind it. While the former is critical for reproducibility, understanding the thought process and the reasons behind it is rather more important. I think that the authors failed to a large extent in covering this aspect.
* **7. Clarity of presentation score**
  + 2. Difficult to read, requires major revision
* **8. Comments on clarity of presentation Please evaluate the clarity of the presentation of the work. Take into account the writing, organization, and quality of figures, tables etc. The English does not need to be flawless, but the text should be understandable.**
  + The clumsy use of English in many passages combined with the large number of typos and errors become very distracting and leave a lot to desire when it comes to the clarify of the paper and the ideas it presents. I would recommend the authors to do a major revision of this paper. e.g., "which mean we can use this to guiding for the phoneme recognition....", "From that view, APL [9], introduce phonetic encoder, complement with acoustic encoder, which using the phonetics embeddings from pre- trained native ASR models, to help the learning of the model on limited amounts of dataset..."
* **10. Overall recommendation**
  + 2: Weak Reject: I am leaning to reject this paper
* **11. Summary Please, summarise in one or two sentences the main motivation for your overall recommendation.**
  + The paper needs a major revision in my opinion and the authors need to step back and provide more insights about the motivation behind the choices of the architecture as well as some theoretical foundation justifying them.In particular, the introduction should also be rewritten with less focus on the details of previous architectures and engineering while putting more focus on the foundation and reasons behind the approach and how it relates to previous work on this level.

---

Based on the feedback from Reviewers #1, #2, #5, #6, and #7, we can group their comments into different criteria for evaluating your paper on Vietnamese Mispronunciation Detection using wav2vec 2.0 and phoneme embedding. Here are the advantages and areas for improvement based on their feedback:

Advantages:

1. \*\*Technical Solidity\*\*:

- Reviewers #2 and #4 noted that the paper is technically solid, providing detailed explanations of datasets, metrics, experimental setups, and baselines.

- Reviewer #5 appreciated the clear description of the experimental setups and the achievement of state-of-the-art results.

2. \*\*Detailed Experimental Verification\*\*:

- Reviewer #6 acknowledged the detailed experimental verification provided by the authors, indicating a strong foundation for the results presented in the paper.

- Reviewer #5 also highlighted the detailed results presented in the paper, which can be useful for other researchers.

Areas for Improvement:

1. \*\*Originality\*\*:

- Reviewers #1, #4, and #6 mentioned that the novelty of the paper is limited, suggesting a minor contribution to the existing literature.

- Reviewer #2 noted that while the addition of the canonical phoneme sequence to the end-to-end system is innovative, there is room for more originality in the work.

2. \*\*Clarity of Presentation\*\*:

- Reviewers #1, #3, and #6 expressed concerns about the clarity of presentation, with comments ranging from difficult to follow, requiring major revision, to clear enough but could benefit from some revision.

- Reviewer #7 specifically highlighted issues with English usage, typos, and errors that detract from the clarity of the paper.

3. \*\*Explanation of Methodology\*\*:

- Reviewers #6 and #7 pointed out a lack of clear information on what was measured in the experiments and a need for more detailed explanations of the tasks and methods employed in the study.

4. \*\*Motivation and Theoretical Foundation\*\*:

- Reviewer #7 emphasized the importance of providing insights into the motivation behind architectural choices and a theoretical foundation justifying them, suggesting a lack of depth in these aspects in the paper.

In your rebuttal, you can address these points by emphasizing the technical solidity of your work, highlighting the detailed experimental verification provided, and acknowledging the need for more originality, clarity in presentation, detailed methodology explanations, and stronger motivation and theoretical foundations. Consider providing additional explanations, clarifications, and justifications to address the reviewers' concerns and improve the overall quality and impact of your paper.