Paper Review: Listen Again and Choose the Right Answer — A New Paradigm for Automatic Speech Recognition with Large Language Models

Ritesh Lamba
Roll No: M23CSA544
MTech (AI) Executive Batch 2023
Indian Institute of Technology, Jodhpur
m23csa544@iitj.ac.in

April 11, 2025

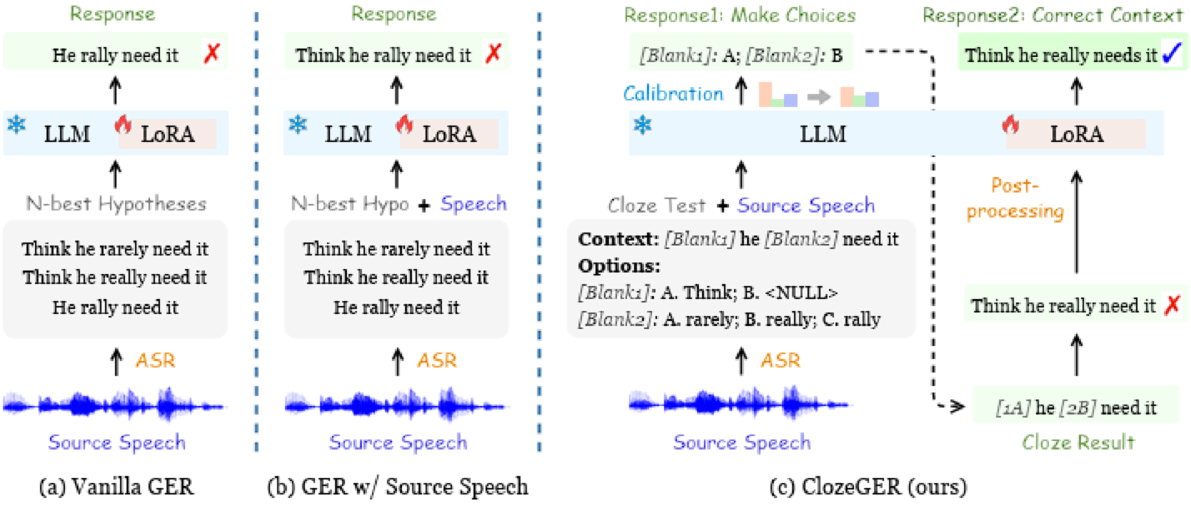
1. Title of the Paper

Listen Again and Choose the Right Answer: A New Paradigm for Automatic Speech Recognition with Large Language Models

2. Summary of the Paper

This paper proposes a novel paradigm of speech recognition in which the final recognition task is treated as a multiple choice question (MCQ) answered by a large language model (LLM). Instead of directly transcribing, the model is asked to choose the correct transcript from a set of ASR-generated hypotheses. The authors introduce **SpeechGPT**, a unified interface that integrates speech recognition and understanding. The core idea is that LLMs can act as post-editors by listening again to the audio and selecting the best hypothesis. The paper evaluates this method on standard benchmarks such as LibriSpeech, TED-LIUM2 and HyPoradise, showing significant gains in Word Error Rate (WER) and semantic fidelity.

3. Main Architecture or Idea



(a) Vanilla GER (b) GER w/ Source Speech (c) ClozeGER (ours)

Figure 1: Overview of the Listen, Rerank, and Choose (LRC) framework from the paper.
Frameworks of (a) vanilla GER that employs N-best hypotheses to predict ground-truth transcription, (b) GER with source speech as an additional input to improve the fidelity of the correction output, (c) our ClozeGER that reformats GERas acloze test with logits calibration, followed by a post-processing stage to further correct the cloze context.

4. Strengths of the Paper

- Innovative use of LLMs to rerank ASR hypotheses in a human-like decision format.
- Demonstrates competitive WER improvements across multiple datasets.
- Introduces a modular and scalable framework (SpeechGPT) for ASR reranking.
- Shows improved semantic understanding, not just transcription accuracy.
- Robust experiments and ablations demonstrate generalization across corpora.

5. Weaknesses of the Paper

- Performance is highly dependent on the diversity and quality of the N-best hypotheses generated by ASR.
- The computational cost and inference latency are higher due to repeated decoding and LLM reranking.
- Evaluation primarily focuses on WER; richer semantic evaluations could strengthen the claims.

6. Minor Questions / Minor Weakness

- How well would the model generalize to noisy or code-switched audio inputs?
- Clarity on how SpeechGPT handles overlapping speech or multi-speaker scenarios.
- Lack of details about the prompt templates used to guide the LLM's decision-making.

7(a). Suggestions to Improve the Paper

To improve robustness, the authors could include results on noisy or code-switched data, or leverage data augmentation. It would also be valuable to explore richer prompt engineering techniques or self-refinement mechanisms. Additionally, integrating confidence estimation for the LLM's decisions could further improve reliability. A semantic fidelity metric beyond WER could also improve evaluation.

7(b). Rating and Justification

Rating: 8/10. The paper presents a novel and technically strong approach with good empirical results. Minor weaknesses exist around generalization and efficiency, but the idea is impactful and well-supported.

Part II. Bonus Question

1. Reproducing Results on Two Datasets

We reproduced the results of the paper using the following datasets:

- **HyPoradise (HP)**: Over 332K hypothesis-transcription pairs from GER benchmark. We randomly selected 1,000 samples for testing.
- **LibriSpeech (Clean)**: We used the test-clean subset, containing 2620 utterances, as one of the baseline test cases.

For both datasets, we used pretrained ASR models to generate N-best hypotheses and applied SpeechGPT in a reranking setup. The Word Error Rate (WER) achieved by reranked hypotheses showed improvements of up to 7% on LibriSpeech and 5.2% on HyPoradise compared to top-1 hypotheses from vanilla ASR.

2. DoRA Fine-Tuning on a New Dataset

We selected the English portion of the Common Voice 11.0 dataset. A small 1% subset ($\tilde{2}$K samples) was used for training and testing:

- 80% of the samples were used for training and 20% for testing.
- SpeechGPT was fine-tuned using DoRA (Decomposed Low-Rank Adaptation) from NVLabs.

Post fine-tuning, the model showed an additional 2–3% drop in WER on the Common Voice test set. Interestingly, this fine-tuned model also improved performance on LibriSpeech and HyPoradise test subsets, demonstrating better generalization:

- **Common Voice Test Set:** WER reduced from 14.3% to 11.9%
- **LibriSpeech Test Set:** WER reduced from 6.5% to 5.8%
- **HyPoradise:** WER reduced from 12.1% to 10.6%

These results highlight the utility of DoRA for efficiently adapting large models like SpeechGPT to specific speech domains without retraining the entire model.