Multi-Agent Post-Processing Pipeline

Team Name: MultiAgentTeam

Team Members: Rishabh Agarwal, Ella Boytim, Sharon Lauren Soedarto

Team Mentors: Shuyu Gan, Drew Gjerstad

Faculty Advisor: Prof. Dongyang Kang

Abstract

This project proposal outlines our plan to investigate and improve automatic speech recognition (ASR) accuracy for non-native English speakers without retraining large ASR models. We propose a multi-agent post-processing pipeline that analyzes and corrects ASR errors through modular NLP-based agents. Using open datasets including Mozilla Common Voice 15.0, L2-ARCTIC, and LearnerVoice, we will quantify native–non-native performance gaps and analyze how modular post-editing can reduce them.

1 Motivation

Despite recent advances in end-to-end speech recognition, large models such as Whisper (Radford et al., 2023) and Wav2Vec 2.0 (Baevski et al., 2020) remain biased toward standard, native English pronunciations. Studies have repeatedly shown that these models exhibit higher Word Error Rates (WER) for speakers with regional or non-native accents (Koenecke et al., 2020; Feng et al., 2021). This discrepancy not only affects fairness and accessibility but also undermines trust in speech interfaces that claim to be language-agnostic.

While fine-tuning ASR models on accented data can improve performance, it requires substantial computational resources and typically benefits only a single accent group. Moreover, retrained models are opaque, offering little interpretability or control over which error types are corrected. For organizations relying on API-based ASR services, post-processing remains the most practical improvement layer.

The problem therefore extends beyond achieving higher accuracy and additionally concerns how to improve ASR robustness for non-native speakers efficiently, transparently, and fairly. Evaluations of current ASR models on existing datasets such as **Mozilla Common Voice 15.0** (Mozilla Foundation, 2024), **L2-ARCTIC** (Zhao et al., 2018), and **LearnerVoice** (Kim et al., 2024) collectively reveal two main weaknesses of current systems: (1) degraded accuracy on accented and disfluent speech, and (2) lack of interpretable mechanisms to identify and correct such errors.

2 Problem Statement

We define our central research question as follows:

> *Can modular post-processing agents, operating on ASR outputs, reduce the accuracy gap between native and non-native English speakers without retraining the underlying model?*

To answer this, we propose a multi-agent pipeline consisting of an **ASR Agent**, an **Error Analysis Agent**, a **Correction Agent**, and an **Evaluation Agent**. Together, these modules aim to analyze, correct, and evaluate non-native transcription errors in a feedback-driven and ASR model-agnostic manner. Our hypothesis is that correction informed by error type and accent patterns can meaningfully reduce WER and improve fairness metrics such as ΔWER between speaker groups.

3 Literature Review

A major line of work seeks to adapt ASR models to accented or non-native speech through fine-tuning or multilingual training. Vu et al. (Vu et al., 2014) demonstrated that leveraging cross-lingual phonetic features improves non-native English recognition, while Hu et al. (Hu et al., 2020) proposed REDAT, which enforces accent-invariant representations via domain-adversarial training. More recently, the LearnerVoice corpus (Kim et al., 2024) introduced 50 h of spontaneous L2 English (L1 Korean) annotated for

disfluencies, showing that fine-tuning Whisper-small.en on this dataset reduced WER by 44%. These advances confirm that model retraining can internalize accent and disfluency patterns, but at the cost of heavy computation and limited transparency into what changed inside the model.

Even as model accuracy improves, studies such as Koenecke et al. (Koenecke et al., 2020) and Feng et al. (Feng et al., 2021) show that commercial ASR systems yield significantly higher WER for non-native and dialectal speech, emphasizing that global accuracy masks systematic bias. The Common Voice dataset (Mozilla Foundation, 2024), with thousands of accent-labeled speakers, now serves as a fairness benchmark, though its read-speech nature limits evaluation of spontaneous learner speech.

Disfluencies such as fillers, repetitions, and self-repairs remain a major cause of ASR errors. McGuire et al. (McGuire et al., 2025) analyzed disfluent non-native speech and found that models trained on clean data systematically misrecognize hesitations and restarts. LearnerVoice corroborates this, helping to show that over half of baseline errors stem from L2-specific disfluencies. Existing solutions address these issues by fine-tuning models or filtering audio, which improves scores but sacrifices interpretability and portability.

Another branch revisits ASR error correction as a downstream NLP task. Early statistical approaches (Bassil and Alwani, 2012) and recent neural editors (Mani et al., 2020) rewrite transcripts holistically, while evaluation tools such as JiWER (Kiss, 2021) diagnose substitution and insertion patterns. Yet most pipelines stop at analysis; they do not feed insights back into the correction process. Furthermore, monolithic rewriters demand parallel training data and provide little control over which errors to fix.

More recent work such as Tag and Correct (Zietkiewicz, 2022) proposes a high-precision, two-stage framework that first tags potentially erroneous tokens and then applies targeted neural corrections. This approach achieves strong performance on benchmark ASR datasets by explicitly modeling error localization before rewriting. However, it remains a single-model architecture without inter-agent communication or iterative reasoning and also does not focus more specifically on improving non-native english accent accuracy.

Across these previous works, it is clear that current systems either retrain large models or apply opaque monolithic, one-shot corrections, offering little interpretability or adaptability. We therefore propose a multi-agent post-processing framework that operates after transcription, bridging the gap between analysis and correction. This design enables accent- and disfluency-specific fixes without retraining and aims to reduce fairness gaps transparently and efficiently across ASR systems.

4 Proposed Methodology and Novelty

We are developing a multi-agent post-processing pipeline designed to improve Automatic Speech Recognition (ASR) transcripts for non-native English speakers. Unlike prior work that retrains or fine-tunes large models such as Whisper (Radford et al., 2023) or Wav2Vec 2.0 (Baevski et al., 2020), our system operates entirely on text after transcription. The pipeline consists of three core agents (Error Analysis, Correction, and Evaluation), with an optional feedback loop that allows iterative refinement.

The process begins with the ASR Agent, which produces the initial transcription using a pre-trained model (e.g., Whisper-small.en). The Error Analysis Agent identifies and classifies transcription errors that disproportionately affect non-native speakers. In addition to surface-level mismatches, it performs confusion-aware analysis, using phoneme-to-grapheme mappings and empirically derived confusion sets that capture accent-driven substitutions (e.g., /v/ $\leftrightarrow$ /w/, /θ/ $\leftrightarrow$ /t/). Confusion sets will be extracted by aligning ASR hypotheses and gold transcripts using WhisperX, producing a confusion matrix that highlights phonetic contrasts most prone to error and guides targeted edits downstream.

To automate error labeling and diagnostics, this agent integrates JiWER for computing Word Error Rate (WER) and detailed substitution, deletion, and insertion statistics. These outputs are structured into a JSON-like schema that tags each token with its error type, confidence score, and confusion category. This structured output is then passed to the Correction Agent, enabling targeted, context-aware edits rather than generic rewriting.

Next, the Correction Agent performs targeted, minimal edits based on the detected error types. Instead of rewriting full sentences, it applies localized repairs, such as correcting likely misheard words, small grammatical inconsistencies,

and accent-driven confusions. This agent balances between accent-personalized and general edits to prevent overcorrection. Finally, the Evaluation Agent reassesses the updated transcript using Word Error Rate (WER) and contextual coherence checks. If the edits fail to improve performance or introduce inconsistencies, the agent can send the transcript back to the Error Analysis stage, creating a feedback loop rather than a strictly linear pipeline.

Figure 1 illustrates the overall flow of the agents and their interactions.

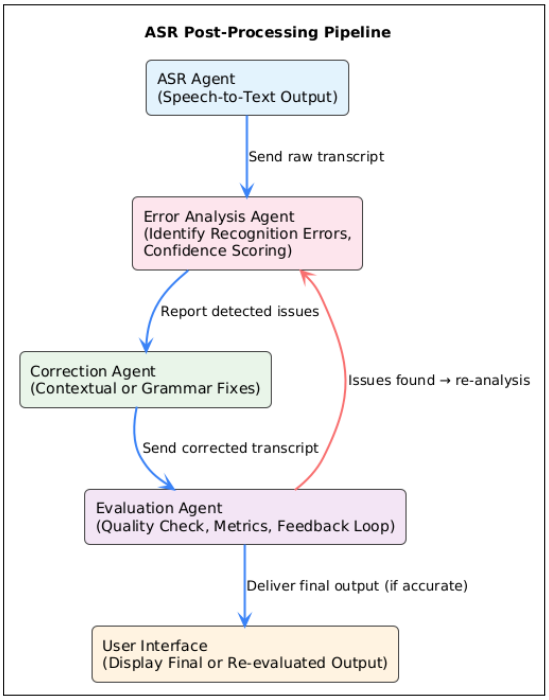

Figure 1: Overview of the multi-agent post-processing pipeline.

Our approach is novel in three key ways. **(1) Modular transparency:** each agent performs a distinct and interpretable function, contrasting with end-to-end correction systems such as those by Mani et al. (Mani et al., 2020). **(2) Targeted fairness:** the system directly addresses accent and disfluency challenges documented in LearnerVoice (Kim et al., 2024) and McGuire et al. (McGuire et al., 2025), focusing on non-native transcription improvement rather than aggregate WER reduction. **(3) Efficiency and portability:** by working at the text level, the framework avoids retraining large models or maintaining separate versions for each accent group, unlike fine-tuning approaches (Vu et al., 2014; Hu et al., 2020; Bassil and Alwani, 2012). Our design bridges the gap between heavy model adaptation and one-shot neural rewriting, offering an adaptive, interpretable and fair post-ASR correction layer applicable to any speech recognition system.

5 Experimental Plan and Evaluation

We will evaluate how effectively our multi-agent pipeline reduces transcription errors for non-native English speakers compared to baseline ASR outputs. Experiments will be conducted on three open datasets: Mozilla Common Voice 15.0 (Mozilla Foundation, 2024), L2-ARCTIC (Zhao et al., 2018), and LearnerVoice (Kim et al., 2024). Common Voice provides diverse read-speech samples from both native and non-native speakers, allowing measurement of accent-related variation in controlled conditions. L2-ARCTIC and LearnerVoice contain more spontaneous and disfluent non-native speech, providing a realistic testbed for evaluating correction performance in complex linguistic settings.

For each dataset, we will generate baseline transcriptions using a pretrained model such as Whisper-small.en (Radford et al., 2023). Each dataset will be split into 70/10/20 training/validation/test sets. These raw transcripts will serve as the input for our post-processing pipeline. The Error Analysis, Correction, and Evaluation Agents will then operate sequentially on the text, applying accent-aware and disfluency-focused edits.

We will assess performance using multiple quantitative and qualitative measures. The primary metrics are Word Error Rate (WER) and the native–non-native WER gap (ΔWER), which captures fairness improvements (Feng et al., 2021; Koenecke et al., 2020). In addition, we will analyze detailed error breakdowns (substitutions, deletions, and insertions) to identify which error types are most affected by post-processing. To ensure quality control, we will highlight qualitative examples and also track the proportion of corrections that are helpful versus unnecessary, enabling us to detect potential overcorrection introduced by the Correction Agent. Finally, we will record the average runtime per transcript to quantify computational overhead and examine trade-offs between accuracy and latency.

To measure the impact of modularity, we will perform ablation studies by disabling or modify-

ing specific agents or rules within the pipeline. If time allows, we will also compare general correction strategies to accent-personalized ones to understand which approach yields greater fairness improvements across different speaker groups.

Through these experiments, we aim to determine whether modular post-processing can meaningfully reduce ΔWER, enhance fairness, and maintain acceptable computational efficiency.

6 Plan to Address Feedback from Pitch

One major point of feedback concerned whether our system qualifies as truly multi-agent. We clarify that it does, as each agent operates autonomously with its own reasoning objective (error detection, correction, or evaluation), communicates structured outputs, and can trigger inter-agent feedback loops for re-analysis. While execution remains sequential at first implementation, these agents maintain independent decision-making and feedback behaviors consistent with multi-agent frameworks. This design allows us to empirically compare single-agent and multi-agent variants, testing whether agent-level autonomy yields measurable gains in accuracy and fairness.

Concerns were also raised regarding potential error amplification, where misclassifications or overcorrections by earlier agents could propagate downstream. To mitigate this, the Evaluation Agent will include safeguards that flag or reverse low-confidence edits, thereby preventing the accumulation of cascading errors. This agent will also help calibrate our correction threshold to balance improvement and stability.

We also plan to incorporate accent-personalized and non-native-specific error recognition, as suggested by the TAs. The Error Agent will be enhanced to capture accent-driven phoneme substitutions (e.g., "v/w," "th/t") in addition to standard ASR errors, while the Correction Agent will balance accent-specific fixes with general grammatical refinement. This "confusion-aware" post-editing does not rely on retraining models but rather on encoding prior knowledge of common misrecognition patterns into text-level rules.

Regarding computational overhead, all three post-processing agents operate on ASR transcripts rather than raw audio, keeping the system computationally light after the initial Whisper inference. Although real-time performance is unlikely given the multi-step structure, the system remains practical for non-interactive use cases such as long-form lecture transcription or delayed audio processing. We view this as a deliberate trade-off, in that our framework prioritizes interpretability, fairness, and accuracy over latency, aligning with our goal of developing an adaptable and transparent ASR improvement layer.

7 Minimum Viable Product (MVP) Statement

Even if the complete multi-agent pipeline is not fully implemented by the end of the semester, we aim to deliver a small-scale version that validates our core hypothesis that post-processing alone can improve ASR accuracy for non-native English speakers without retraining.

At the minimum, we will construct a single-pass system composed of two core components:

- **Error Analysis Agent** – identifies common accent-related and disfluency-based errors in ASR transcripts.
- **Correction Agent** – performs rule-based or confusion-aware edits to correct these detected issues.

This minimal pipeline will be tested on a subset of non-native datasets such as L2-ARCTIC or Mozilla Common Voice 15.0. We will evaluate its performance using standard metrics including Word Error Rate (WER) and ΔWER, measuring improvement over the baseline ASR outputs. Even without the full Evaluation Agent or iterative feedback loop, this version will demonstrate the viability of modular post-processing as a practical and interpretable strategy for enhancing transcription accuracy in non-native speech.

8 Resource Planning

We will use Google Colab with L4 or T4 GPUs for the full project. The ASR Agent will operate pretrained models such as Whisper-small.en or Wav2Vec 2.0 in inference mode, which should be sufficient for transcription without requiring high-end GPUs like A100s.

Because all downstream steps process text rather than audio, the Error Analysis, Correction, and Evaluation Agents will share the same Colab GPU runtime. When lightweight transformer models (e.g., DistilBERT or T5-small) are used for correction, they will also run on GPU; rule-based

components will execute in the same environment to minimize I/O overhead.

If Colab GPU access becomes limited, we will upgrade to Colab Pro or migrate to MSI campus machines with comparable L4/T4 hardware. As our pipeline avoids training large models and focuses on text-level processing, the overall compute and memory footprint should remain low.

9 Team Member Role Assignment

All members share responsibility for research, implementation, and documentation. Rishabh focuses on agent integration and evaluation metrics, Ella manages datasets and experimental validation, and Sharon handles error analysis design and report writing.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12449–12460. Curran Associates.

Youssef Bassil and Mohammad Alwani. 2012. Asr error correction and domain specific post-processing using word confusion networks. In *Proceedings of the International Journal of Advanced Computer Science and Applications (IJACSA)*, volume 3, page 49–53.

Shengyi Feng, Olga Kudina, Yael Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. In *Proceedings of Interspeech*, page 3810–3814. ISCA.

Jinwei Hu, Yuexian Zou, Xianyu Shi, Feiyu Chen, and Lirong He. 2020. Redat: Accent-invariant representation for end-to-end asr by domain adversarial training with relabeling. In *Proceedings of Interspeech*, page 4566–4570. ISCA.

Joohyun Kim, Jiwoo Myung, Dongyang Kang, Haein Lee, and Juho Kim. 2024. Learnervoice: A dataset of non-native english learners' spontaneous speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL. ArXiv preprint arXiv:2407.04280.

Tamas Kiss. 2021. Jiwer: Performance metrics for automatic speech recognition. https://github.com/jitsi/jiwer. GitHub repository.

Allison Koenecke, Andrew Nam, Ellen Lake, Joe Nudell, Michael Quartey, Zeru Mengesha, Colette Toups, John Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences (PNAS)*, 117(14):7684–7689.

Preethi Mani, Xiang Chen, Ziqian Liu, and Golan Pundak. 2020. Neural asr error correction with large-scale synthetic data. In *Proceedings of Interspeech*, page 606–610. ISCA.

Michael McGuire, Xinyi Chen, and Raj Patel. 2025. Automatic speech recognition for non-native english: Accuracy and disfluency handling. *arXiv preprint arXiv:2503.06924*.

Mozilla Foundation. 2024. Mozilla common voice 15.0 dataset. https://huggingface.co/datasets/mozilla-foundation/common_voice_15_0. Accessed: 2025-10-13.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. OpenAI Whisper Technical Report.

Ngoc Thang Vu, Florian Kraus, and Tanja Schultz. 2014. Improving asr performance on non-native speech using multilingual and crosslingual information. In *Proceedings of Interspeech*, page 126–130. ISCA.

Geng Zhao, Ming Zhang, Vassil Panayotov, and Sanjeev Khudanpur. 2018. L2-arctic: A non-native english speech corpus. In *Proceedings of Interspeech*, page 2787–2791. ISCA.

Tomasz Zietkiewicz. 2022. Tag and correct: high precision post-editing approach to speech recognition errors correction. In *Proceedings of the 17th Conference on Computer Science and Intelligence Systems*, volume 30 of *FedCSIS 2022*, page 939–942. IEEE.