StreamCorrect

Self-Supervised Error Correction for Real-Time ASR with Offline Recognizers

Tung D. Vu[1,3], Zeying Xie[2], Weiwei Zhao[2], Victor Junqiu Wei[1], Di Jiang[2], Yi R. Fung[1]

[1] The Hong Kong University of Science and Technology [2] WeBank [3] VinUniversity

1 Introduction

Offline recognizers achieve high accuracy with full audio context, whereas streaming deployment restricts access to future context, causing early recognition errors to **propagate** through the confirmed transcript and affect subsequent chunks.

We introduce **StreamCorrect**, a SpeechLM-based error corrector trained on **self-generated ASR errors,** which mitigates error propagation in streaming recognition without modifying or retraining the underlying ASR backbone**.**

3 Approach

3.1 Problem Definition

For each chunk, we predict the text to append using the **confirmed prefix**, **top-k hypotheses** and available audio. We append the corrected segment to the prefix, improving context for subsequent decoding.

3.2 Overview of StreamCorrect

We first synthesize streaming errors with an offline-based streaming ASR and recover chunk-level targets using **Prefix-to-Reference Alignment (PRA).** At inference, the trained corrector refines each segment before committing it to the confirmed prefix.

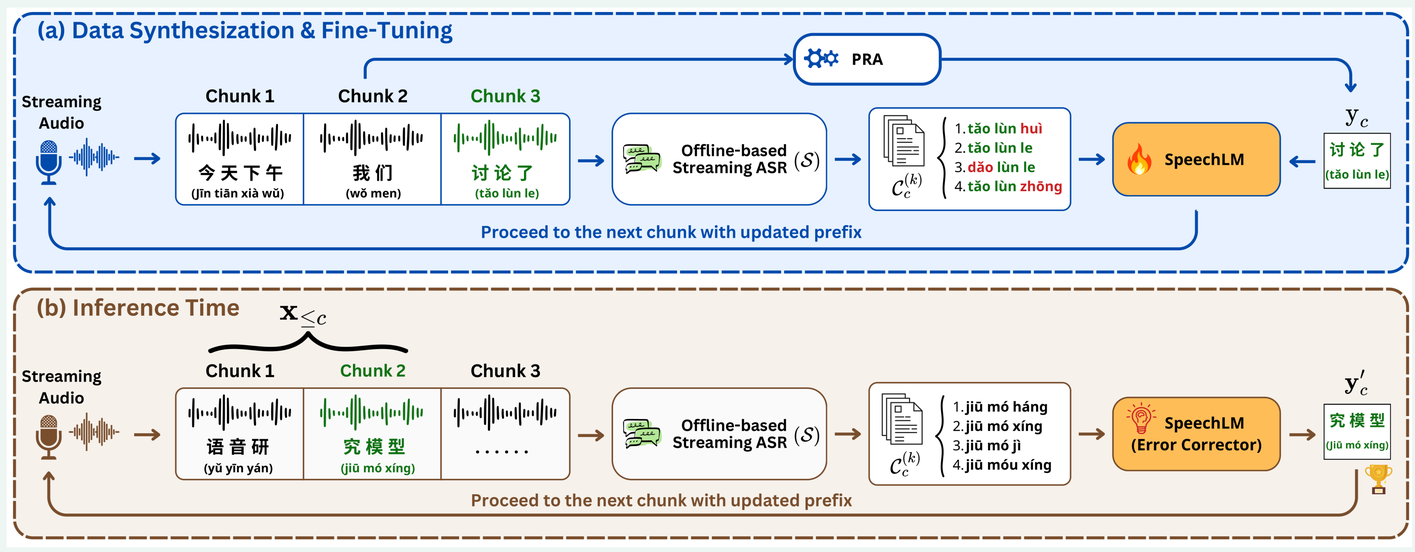
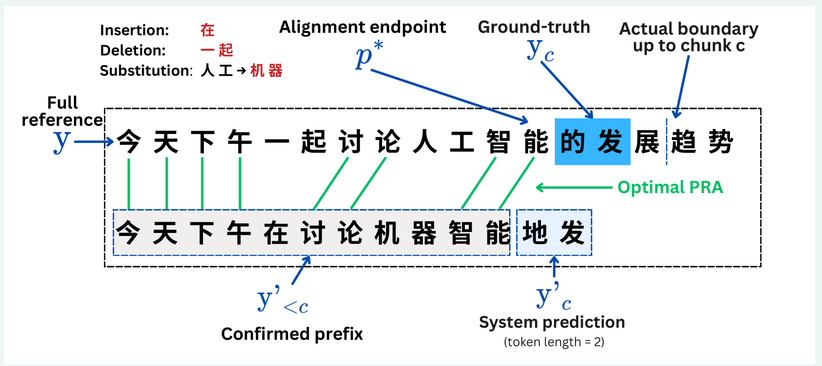


Figure 1. Overview of StreamCorrect. **(a)** Data synthesis & fine-tuning: streaming errors are generated by an offline-based ASR, and PRA recovers timestamp-free chunk-level supervision. **(b)** Inference: the fine-tuned SpeechLM refines top-k hypotheses using available context

Prefix-to-Reference Alignment, a timestamp-free boundary recovery approach



PRA aligns the confirmed prefix with the reference via edit distance and selects the minimum-cost endpoint, breaking ties with the earliest endpoint to preserve recoverable content.

PRA extracts the next reference tokens using the emitted segment length. We use utterance transcripts, with **no extra chunk timestamps**.

4 Experiments

4.1 Benchmark and Experimental Setup

We evaluate on AIShell-1 (Mandarin), a 30-min banking-domain dataset, and WSYue (9.5 h, Cantonese), using Belle Whisper and Whisper Medium Yue as ASR backbones. We LoRA-tune Ultravox-v0.5-Llama-3.2-1B and compute CER and first/last-token latency (FTL/LTL) metrics.

4.2 Statistics of Self-Generated Training Data

We generate **~1M samples** using 100–1500 ms chunks and beams of 1, 2, 4 and 8. **Substitutions dominate**, with ~22% of substitutions due to homophone confusion.

Table 1. AIShell-1 training-set CER (%) by chunk size and error type.

Chunk (ms)	Substitution	Insertion	Deletion	CER (%)
100	13.1	6.8	1.7	**21.6**
500	8.4	6.6	3.2	**18.2**
1000	7.0	4.5	3.2	**14.7**
1500	6.2	3.4	3.2	**12.8**

4.3 Results and Analysis

StreamCorrect reduces CER across all datasets and chunk sizes (Table 2).

Table 2. CER (%) by chunk size: SimulStreaming / StreamCorrect (beam 4).

Dataset	100 ms	500 ms	1000 ms
AIShell-1	27.3 / 11.7	16.3 / 9.8	13.9 / 7.9
Industrial	38.7 / 26.8	27.8 / 20.7	24.1 / 18.5
WSYue	21.2 / 13.9	15.6 / 10.2	14.1 / 8.2

At 500 ms on AIShell-1, StreamCorrect reduces CER from **16.3% to 9.8%**, with 61 ms additional FTL and 82 ms additional LTL.

4.4 Error Propagation Effect

Early chunks benefit little, and chunk 2 can degrade through over-correction. **Later chunks gain more** as the corrected prefix limits accumulated errors.

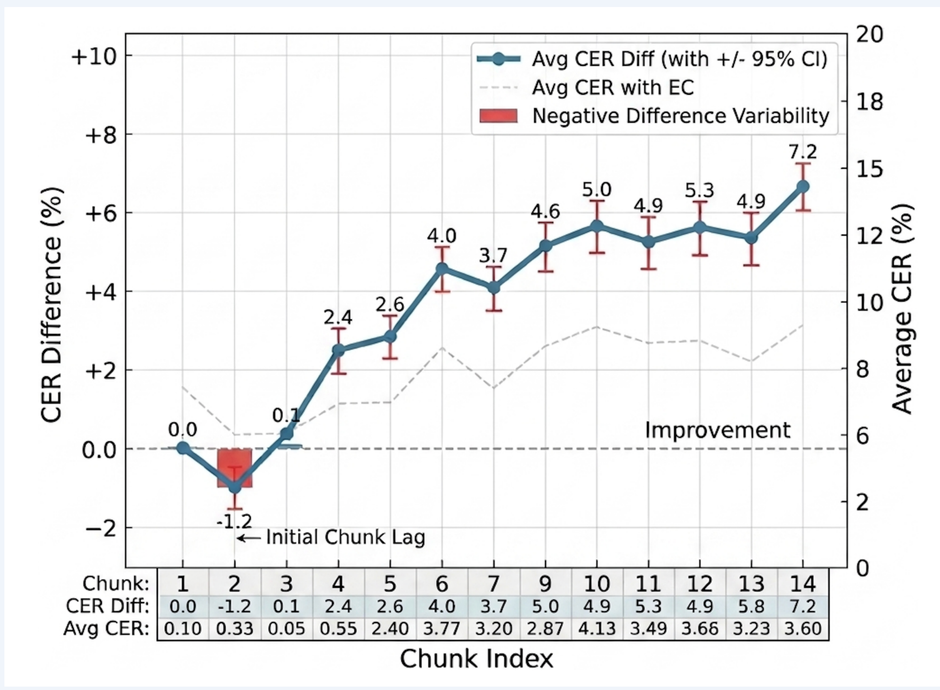


Figure 4. AIShell-1 chunk-wise performance, excluding utterances shorter than 4 s. Positive CER difference indicates improvement. We observe gains up to 7.2 percentage points in later chunks.

4.5 Industrial Integration

In the banking pipeline, the ASR remains unchanged**,** with the corrector using top-4 hypotheses. Correction is capped at **8 new tokens per 500 ms chunk** to bound computation. Each GPU worker hosts a complete ASR-corrector replica, with **~6.9 GB of FP16/BF16 model weights per replica**, excluding runtime states.

5 Conclusion

StreamCorrect enhances offline-based streaming ASR with a **separately trainable error corrector**. **PRA** enables chunk-level supervision, while **corrected prefixes mitigate error propagation** over time. Its modular design supports independent development and deployment of the recognizer and corrector.

References

Limitations

Evaluation is limited to Mandarin and Cantonese, with broader linguistic and domain generalization left for future work.

Ethical Considerations

Deployment in sensitive or high-stakes domains requires privacy-aware data governance, domain-specific validation, and appropriate human oversight.

Code: github.com/vu-duy-tung/StreamCorrect **Contact:** dtvu@connect.ust.hk / yrfung@ust.hk

Acknowledgment: Web24EG01 and HKUST Research Travel Grant.