INTERSPEECH2025

Large Language Models based ASR Error Correction for Child Conversations

Anfeng Xu∗1, Tiantian Feng∗1, So Hyun Kim2, Somer Bishop3, Catherine Lord4, Shrikanth Narayanan1

Viterbi School of Engineering, University of Southern California, USA, School of Psychology, Korea University, South Korea, Weill Institute for Neurosciences, University of California, San Francisco, USA, Semel Institute of Neuroscience and Human Behavior, University of California, Los Angeles, USA

2025.10.02

HyorinJung

Index

- **Introduction**
- **Background**
- **Analyze & Delineate**
    - **Experiments**
        - Setting
        - Result
- **Limitation & Future Direction**
- **Conclusion**
- **Proposal**

Introduction

Speech Foundation Models(SFM)

- **End-to-End Supervised Models :** Learn acoustic and language features jointly from large labeled datasets (ex) Whisper, Parakeet)
- **Self-Supervised Learning (SSL) Models :** Learn representations from unlabeled audio(ex) Wav2vec 2.0, HuBERT, WavLM)

Introduction

Problem with Children’s Speech

- Error rates are **10–19× higher than adults**
- still **6× higher** even after adaptation

Why Challenging? -> Children’s speech differs from adults in :

- Acoustic-phonetic characteristics
- Vocabulary usage
- Prosodic(억양) features (intonation, rhythm)
- Conversational dynamics

Introduction

Large Language Models (LLMs) in ASR

- LLMs with audio or speech encoders enables improved ASR performance
- By leveraging language structure, context, and semantic relationships, LLMs can effectively correct ASR errors, considering both narrow linguistic patterns and broader semantic context
- selecting and refining N-best ASR hypotheses using LLMs significantly improves transcription quality

👉 Explore Methods to Correct Errors in Child Speech Recognition (ASR) Using Large Language Models (LLMs)

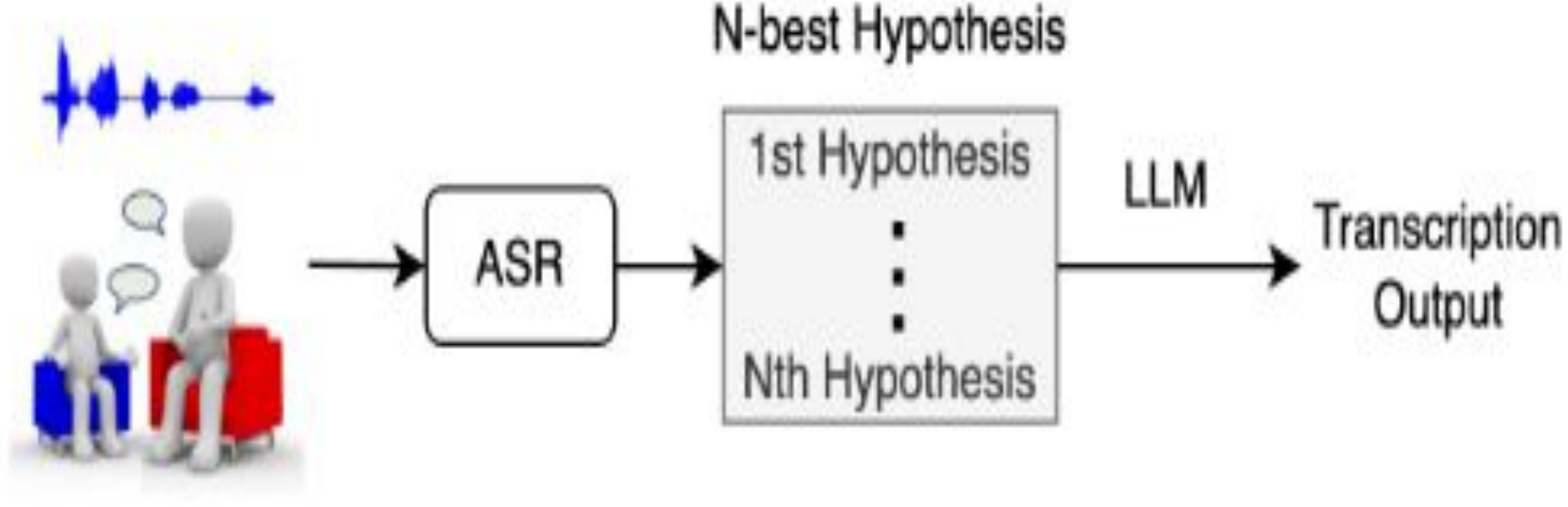


Figure 1: *Overall pipeline for ASR with LLM error correction.*

Background

Whisper

- Attention-based encoder–decoder ASR model
- Trained on 680k hours of multilingual speech (weak supervision)
- Robust across diverse recording conditions; strong ASR benchmark results
- Models used: WSP-S, WSP-L, WSP-L-T
- Zero-shot outputs from all, fine-tuned outputs from WSP-L-T
- Beam search decoding (beam size = 60)

Background

WavLM

- SSL model building on Wav2vec 2.0 & HuBERT
- Learns universal speech reps via masked prediction pre-training
- State-of-the-art on SUPERB benchmark
- Only fine-tuned outputs used for ASR experiments
- Model: WavLM-L, fine-tuned with CTC loss
- Beam search decoding (beam size = 10)

Background

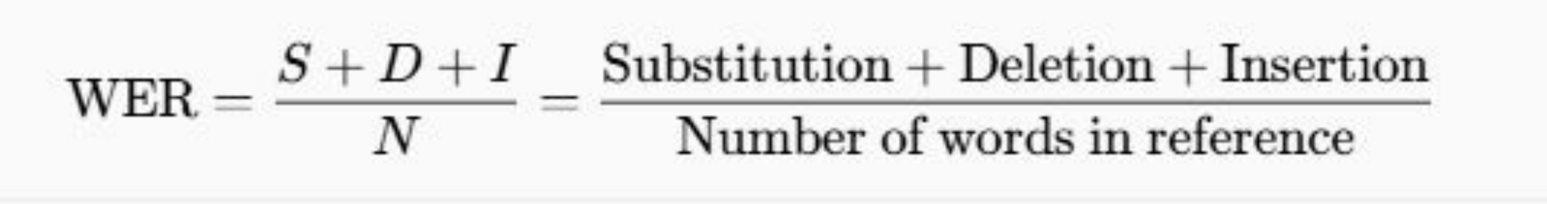
ASR Fine-tuning & LLM Error Correction

- Goal: Test LLM correction on **fine-tuned ASR outputs** (not only zero-shot)
  Generated fine-tuned outputs for train & test sets
- 2-fold cross-validation: validation outputs → training data for LLM correction
- Test set: fine-tuned models trained on full training set
- ASR models: WSP-L-T, WavLM-L

Background

WER(Word Error Rate)

- Preprocessing with Whisper Normalizer
    - Applied to both ground-truth transcripts and ASR outputs before WER calculation
    - Normalizes text for consistency
    - Ensures fair and accurate comparison
- WER is Standard metric for ASR performance



$$\mathrm{WER} = \frac{S + D + I}{N} = \frac{\text{Substitution} + \text{Deletion} + \text{Insertion}}{\text{Number of words in reference}}$$

Experiments : Setting

Datasets

- **MyST (Children’s Tutoring Corpus)**
  a. Grade 3–5 children (8–12 yrs) with virtual tutors
  b. 8 science topics (e.g., biology, physics)
  c. Used only for ASR error correction *without contex*
  d. Official train/test split
- **ADOS-Mod3 (Autism Diagnostic Sessions)**
  a. 352 sessions from 180 children (2–13 yrs)
  b. Data from two medical centers: Chicago (train), Michigan (test)
  c. Contains both child & adult speech
  d. Average: 25.9 child utterances / 30.0 adult utterances per session

Experiments : Setting

ASR Models

- **Whisper & WavLM**

Implementation Details

- Whisper (WSP-L-T)
  - *Fine-tuning setup:*
    - *2000 steps*
    - *Learning rate: 1e-6 (0.000001)*
- WavLM (WavLM-L)
  - Fine-tuning setup:
    - 30000 steps
    - Learning rate: 3e-4 (0.0003)
- Common Settings
  - Batch size: 32
  - Optimizer: Adam

Experiments : Setting

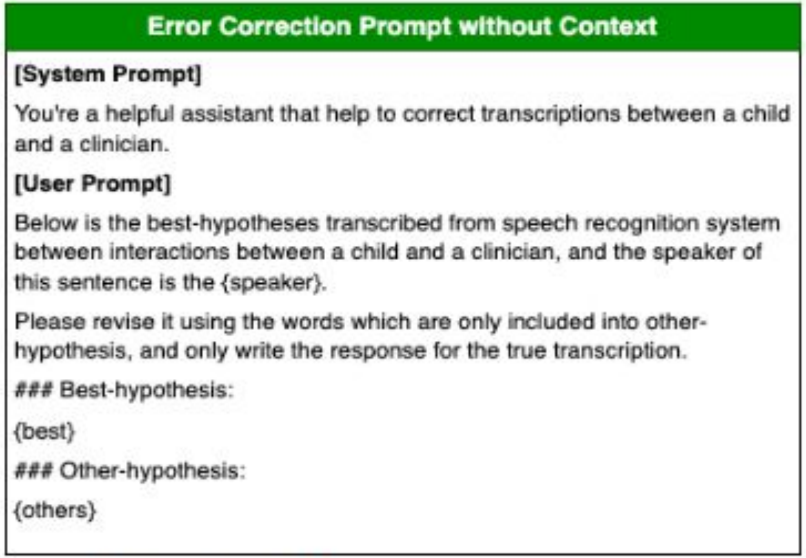
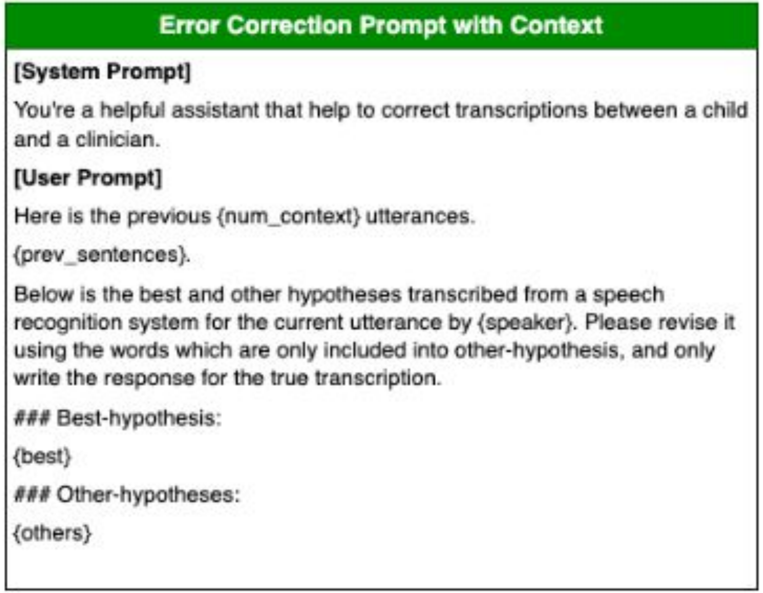
LLM Models

- **LLaMa 3.1-8B and LLaMa 3.2-1B**

Implementation Details

- Training epochs: 5 epochs (MyST), 10 epochs (ADOS)
- Learning rate: 5e−4
- Prompts:
    - *MyST*: “You are a helpful assistant that helps to correct child transcripts.”
    - *ADOS*: Task-specific system prompt(next slide prompt)
- Inference: temperature = 0.2 -> high accuracy
- Safeguard: Use best ASR hypothesis if LLM output exceeds it by >3 words

prompt

Context-Free ASR Error Correction	Context-Aware ASR Error Correction
 **Error Correction Prompt without Context** **[System Prompt]** You're a helpful assistant that help to correct transcriptions between a child and a clinician. **[User Prompt]** Below is the best-hypotheses transcribed from speech recognition system between interactions between a child and a clinician, and the speaker of this sentence is the {speaker}. Please revise it using the words which are only included into other-hypothesis, and only write the response for the true transcription. ### Best-hypothesis: {best} ### Other-hypothesis: {others} Figure 2: *LLM prompt without context.*	 **Error Correction Prompt with Context** **[System Prompt]** You're a helpful assistant that help to correct transcriptions between a child and a clinician. **[User Prompt]** Here is the previous {num_context} utterances. {prev_sentences}. Below is the best and other hypotheses transcribed from a speech recognition system for the current utterance by {speaker}. Please revise it using the words which are only included into other-hypothesis, and only write the response for the true transcription. ### Best-hypothesis: {best} ### Other-hypotheses: {others} Figure 3: *LLM prompt with context.*

Inference Temperature

- Controls randomness vs. consistency in text generation
- Applied in the softmax function

$$P(w_i|\text{context}) = \frac{\exp(\text{logit}(w_i)/T)}{\sum_j \exp(\text{logit}(w_j)/T)}$$

- Effects of temperature:
  - High T (>1): flatter distribution → more diverse, creative, but risk of irrelevant outputs
  - Low T (0<T<1): sharper distribution → more predictable, consistent, but possibly repetitive
  - T → 0: almost deterministic, always picks the highest-probability word

👉 **T = 0.2: accurate ASR error correction**

Experiments : Results

RQ1: Can LLMs Improve zero-shot Child ASR Results?

- Consistent WER reductions across **ALL** Whisper models when **using LLaMA 3.1-8B**

- Smaller model (LLaMA 3.2-1B) shows less improvement

👉 larger LLMs perform better in ASR error correction

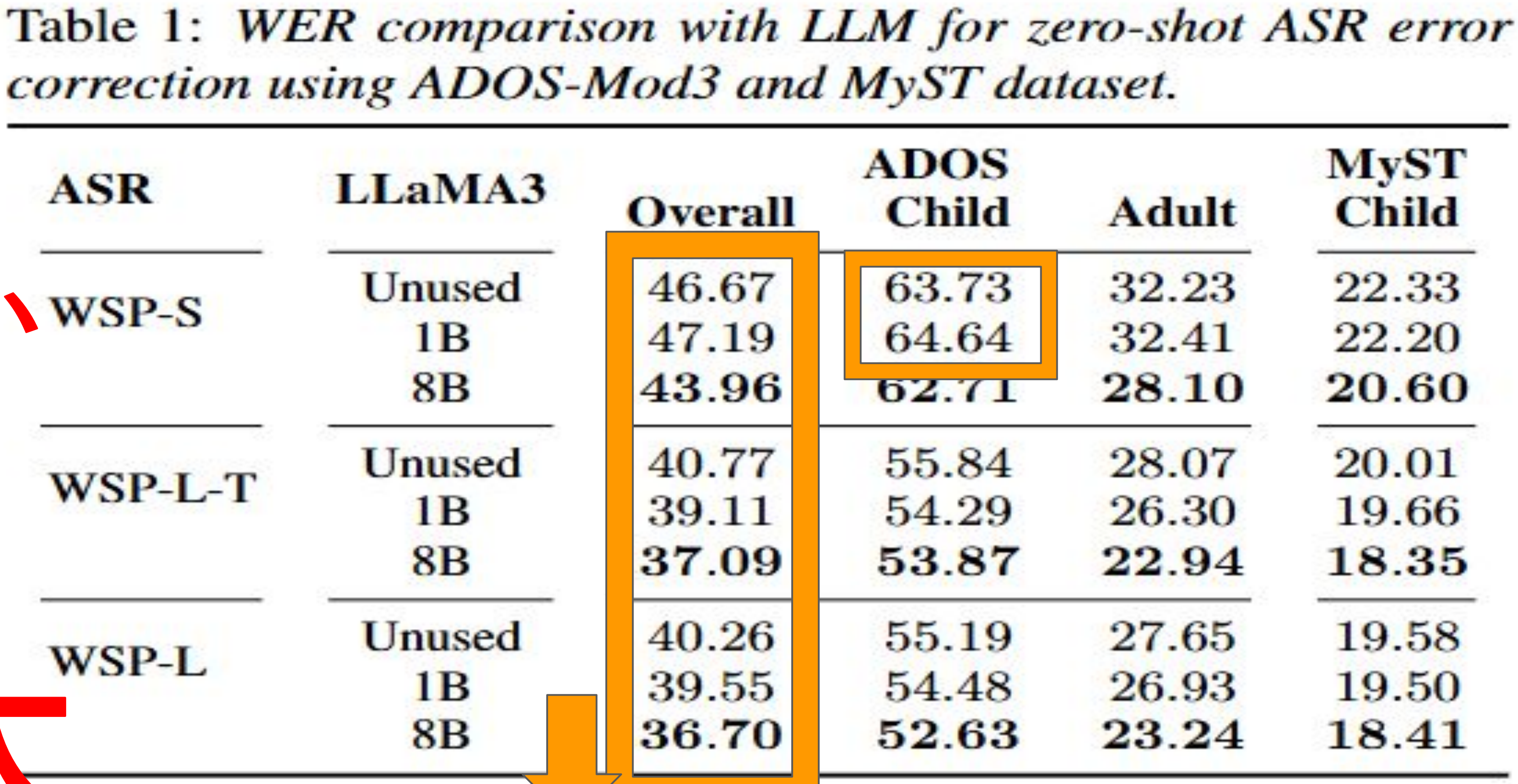


Table 1: *WER comparison with LLM for zero-shot ASR error correction using ADOS-Mod3 and MyST dataset.*

小

ASR	LLaMA3		ADOS		MyST
		Overall	Child	Adult	Child
WSP-S	Unused	46.67	63.73	32.23	22.33
	1B	47.19	64.64	32.41	22.20
	8B	**43.96**	**62.71**	**28.10**	**20.60**
WSP-L-T	Unused	40.77	55.84	28.07	20.01
	1B	39.11	54.29	26.30	19.66
	8B	**37.09**	**53.87**	**22.94**	**18.35**
WSP-L	Unused	40.26	55.19	27.65	19.58
	1B	39.55	54.48	26.93	19.50
	8B	**36.70**	**52.63**	**23.24**	**18.41**

大

Experiments : Results

RQ2: Can LLMs Improve Fine-tuned Child ASR Results?

- WERs for MyST are relatively high due to keeping all utterances without strict filtering(removing higher than WER 50%)
- WSP-L-T consistently outperforms WavLM-L across both datasets
- Whisper (autoregressive BPE tokens) → fewer spelling mistakes
- WavLM (CTC-based) → more spelling errors, especially with children's unclear pronunciation

👉Substantial improvements for WavLM outputs (spelling correction)

👉 Limited benefits for Whisper outputs (similar autoregressive decoding, no access to speech features)

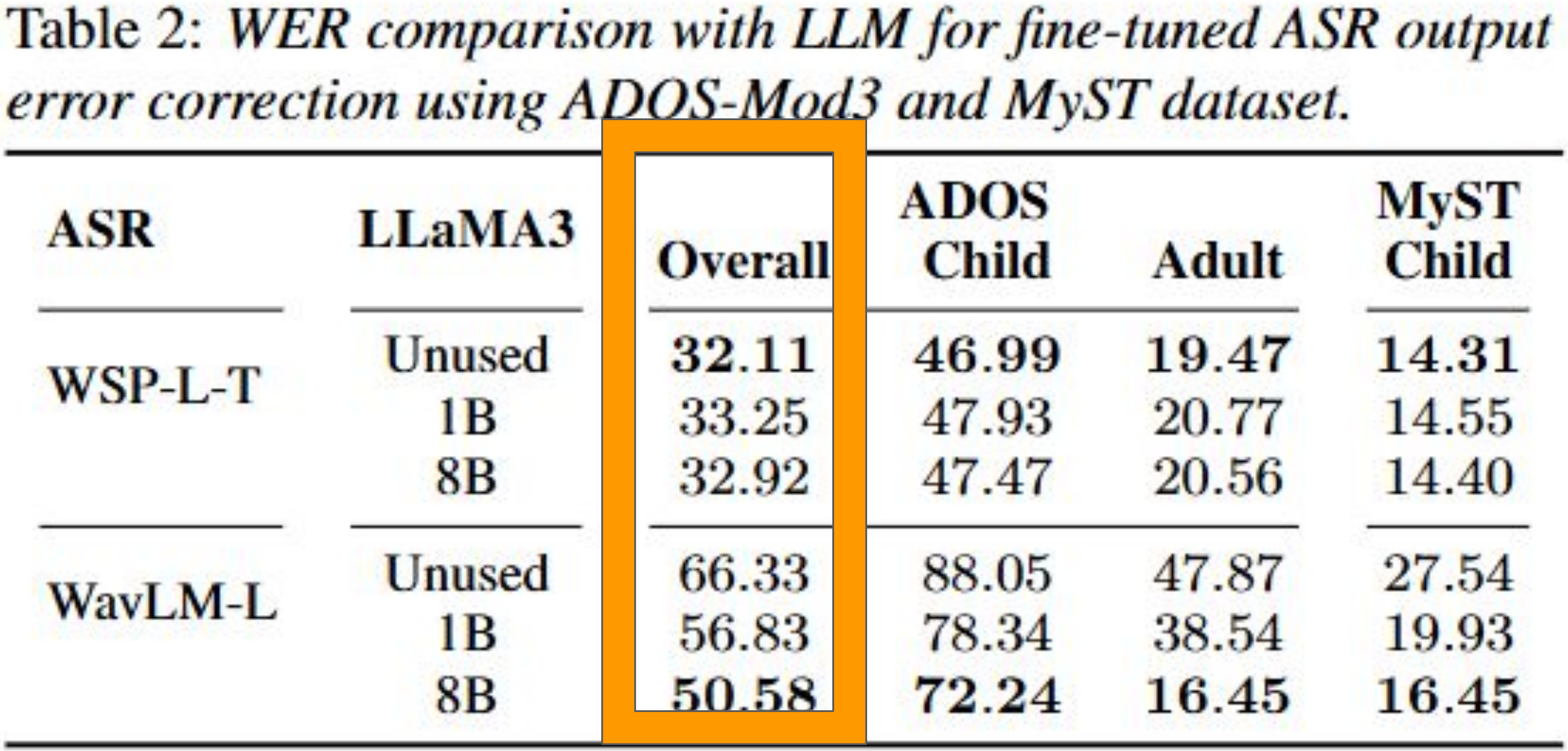


Table 2: *WER comparison with LLM for fine-tuned ASR output error correction using ADOS-Mod3 and MyST dataset.*

ASR	LLaMA3		ADOS		MyST
		Overall	Child	Adult	Child
WSP-L-T	Unused	**32.11**	**46.99**	**19.47**	**14.31**
	1B	33.25	47.93	20.77	14.55
	8B	32.92	47.47	20.56	14.40
WavLM-L	Unused	66.33	88.05	47.87	27.54
	1B	56.83	78.34	38.54	19.93
	8B	**50.58**	**72.24**	**16.45**	**16.45**

Experiments : Results

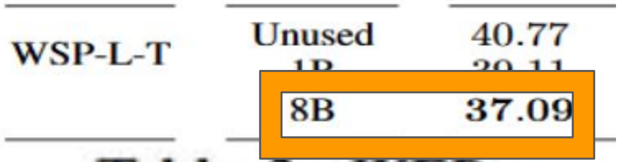
RQ3: Does Context Improve LLM Error Correction?

- Adding previous utterances (1 or 3) unexpectedly increased WERs
- Context of 3 utterances caused even higher error rates than 1 utterance
- **Main reason:** **Error propagation**

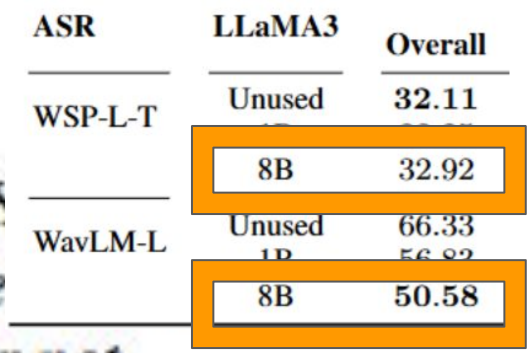
Experiments : Results

RQ3: Does Context Improve LLM Error Correction? prompt.

> You are a helpful assistant that helps to correct transcriptions from a child in a tutoring session.
> Here are the previous {num context} sentences in the conversation:
> {prev sentences}
>
> Here is the current ASR output:
> {speaker}: {best}
> Here are other hypotheses from the ASR:
> {others}
>
> Please output the corrected transcription for {speaker}.

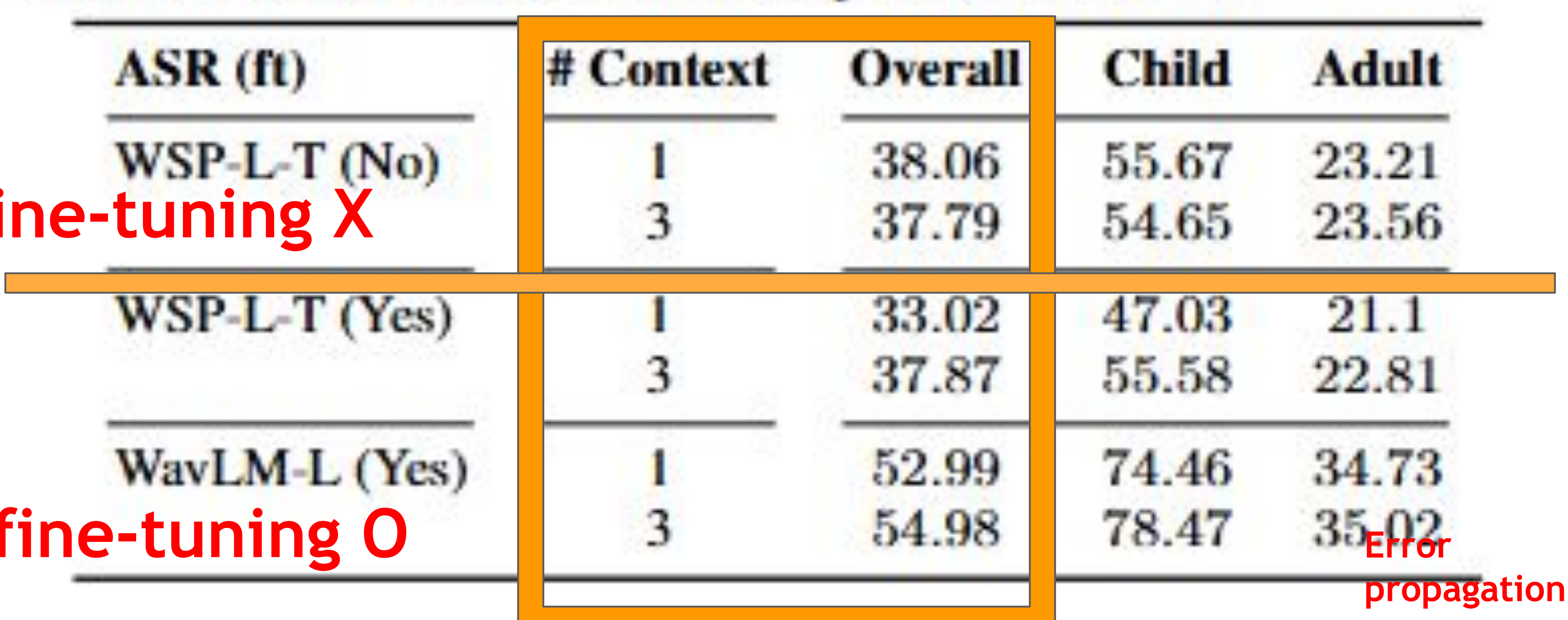


WSP-L-T	Unused	40.77
	1B	39.11
	8B	**37.09**



ASR	LLaMA3	Overall
WSP-L-T	Unused	**32.11**
	1B	[illegible]
	8B	32.92
WavLM-L	Unused	66.33
	1B	56.82
	8B	**50.58**

Table 3: *WER comparison with LLaMA 3.1-8*
correction using context. ADOS-Mod3 datase
dicates whether the ASR model is fine-tuned or not.



ASR (ft)	# Context	Overall	Child	Adult
WSP-L-T (No)	1	38.06	55.67	23.21
	3	37.79	54.65	23.56
WSP-L-T (Yes)	1	33.02	47.03	21.1
	3	37.87	55.58	22.81
WavLM-L (Yes)	1	52.99	74.46	34.73
	3	54.98	78.47	35.02

fine-tuning X

fine-tuning O

Error propagation

Experiments : Results

RQ4: Analysis on utterance length. zero-shot

- LLM correction most effective for single-word utterances
- Reason: Whisper often produces phonetically similar but contextually incorrect words
- LLM helps refine these into more plausible utterances

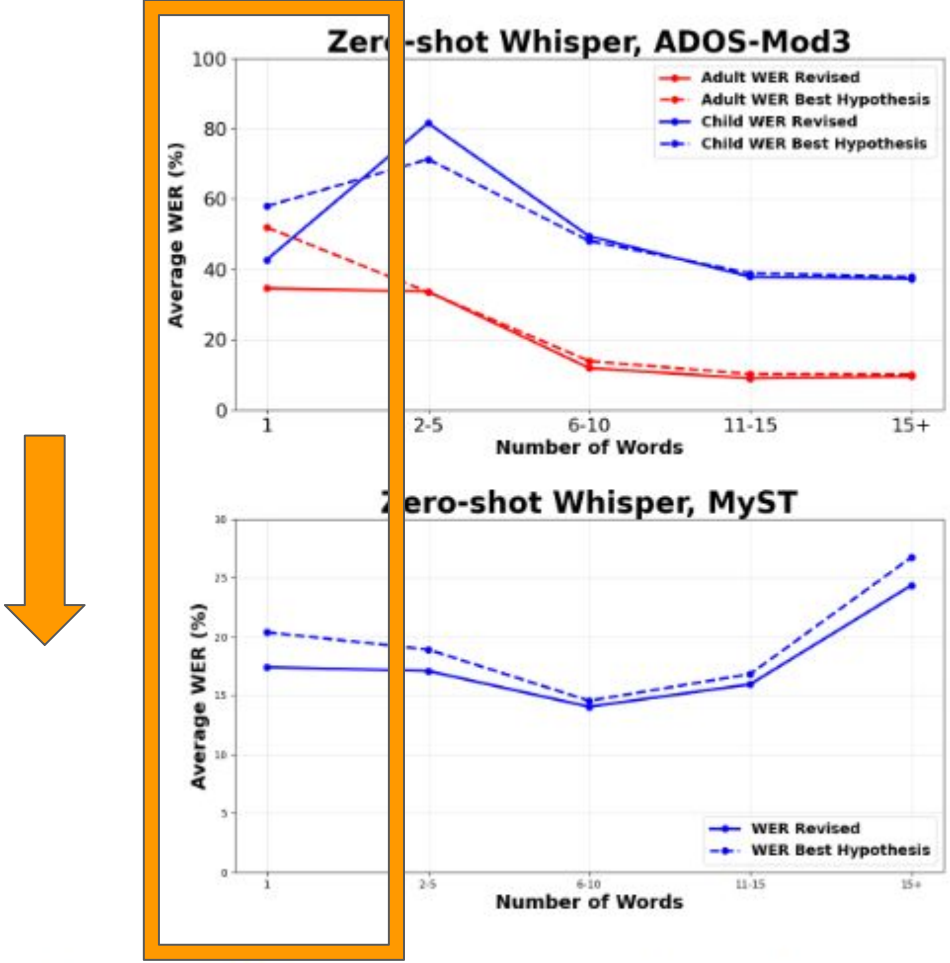


Figure 4: *WERs by utterance lengths with zero-shot Whisper ASR (WSP-L-T). Results from both datasets.*

Experiments : Results

RQ4: Analysis on utterance length. fine-tuning

- Whisper: little improvement except for single-word child utterances but contextually incorrect words
- WavLM: consistent improvements across utterance lengths

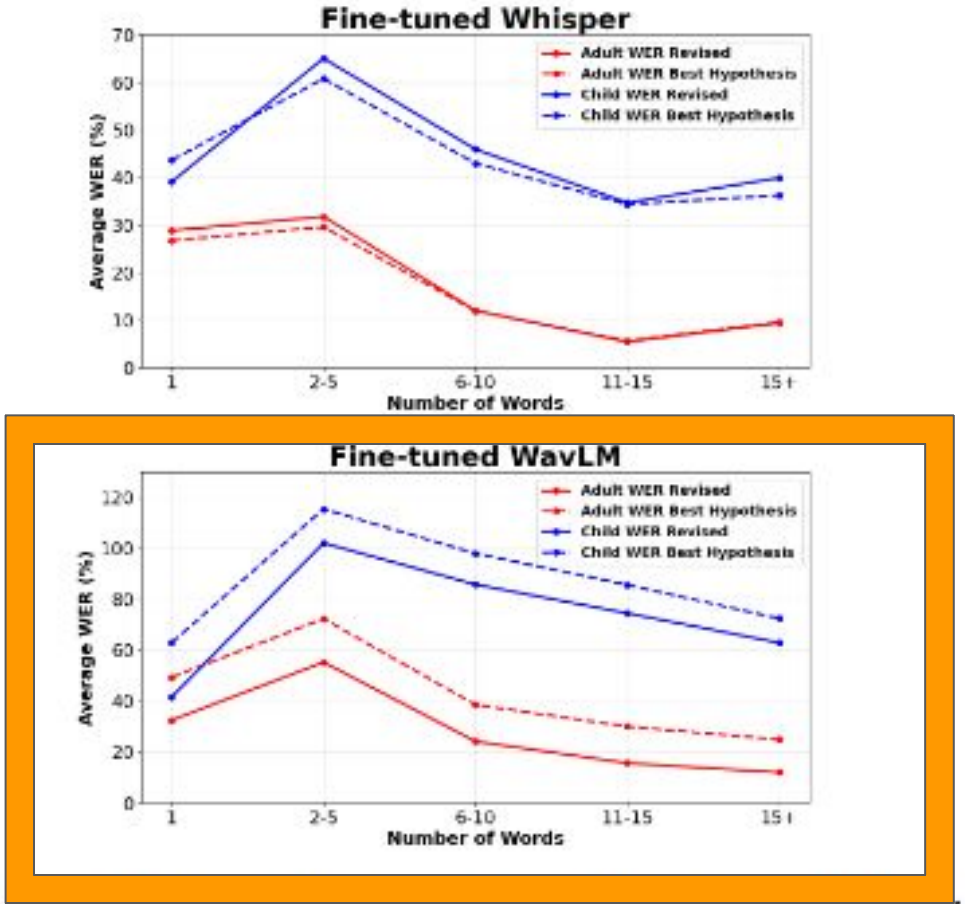


Figure 5: *WERs by utterance lengths with fine-tuned ASR models (WSP-L-T, WavLM-L), using the ADOS-Mod3 dataset.*

Limitation & Future Directions

- Effectiveness depends heavily on LLM size
- Restricted by the scarcity and diversity of child conversational speech datasets
- Develop robust methods to integrate conversational context without error propagation
- Current approach separates ASR and LLM; tighter integration may be needed for better performance.

Conclusion

- **Larger** LLMs consistently improve zero-shot ASR (Whisper)
- For fine-tuned ASR:
    - Strong improvements for **CTC-based models** (WavLM) via **spelling correction**
    - Minimal gains for autoregressive models (Whisper)
- Incorporating conversational context degrades performance due to error propagation
- Improvements mainly for short utterances, limited or negative for longer ones

Proposal

- Integration with Acoustic Features
  - ex) AudioLM, SpeechGPT, SpeechChain
- Robustness to Child-Specific Variability
  - Analyzing vocal and prosodic features of autistic children’s speech
- Exploring Decoding Strategies in ASR Models

Appendix

- Character Error Rate (CER) : 문자(character) 단위의 오류율을 측정
  - 언어적 특성: 한국어, 일본어와 같이 단어 경계가 모호하거나 음절 단위가 중요한 언어, 또는 합성어(agglutinative language)가 많은 언어에서 WER보다 더 정확한 오류를 반영
  - 철자 오류: ASR 시스템이 단어를 잘못 인식하기보다는 철자를 틀리는 경우, WER은 100% 오류로 보지만 CER은 더 작은 오류율을 보여줄 수 있음
- llm의 문맥 정보 활용 능력을 올리기 위한 방법들
  - 컨텍스트 정보의 신뢰도 관리
    - ASR 신뢰도 점수 활용 : 이전 발화의 ASR 결과에 대한 신뢰도 점수(confidence score)를 함께 LLM에 입력으로 제공
    - 컨텍스트 선택적 적용 : 임계값 기준으로 사용
  - 프롬프트 엔지니어링 강화
    - LLM 프롬프트에 이전 컨텍스트의 잠재적 오류 가능성을 명시하고, 현재 발화의 ASR 후보군에 더 집중하도록 지시

Appendix

제목	저자	학회	요약
Who Said What? An Automated Approach to Analyzing Speech in Preschool Classrooms	T. Merritt, J. VanDam, et al.		치원 교실에서 녹음된 아이들과 교사의 발화를 자동으로 분석하는 시스템 제안
Using Data Augmentation and Time-Scale Modification to Improve ASR of Children's Speech in Noisy Environments	F. S. A. Rauf, R. A. Rasheed, et al.	Applied Sciences (MDPI), 2021	잡음 환경에서 아동 음성 ASR 성능을 개선하기 위해 데이터 증강과 시간 스케일 변형 기법을 적용
Activity Focused Speech Recognition of Preschool Children	M. Seidl, K. Evanini, et al.	BEA Workshop @ ACL (2022)	다양한 교실 맥락(놀이, 대화 등)에서 발생하는 발화 특징을 반영한 음성 인식 실험을 통해 자연스러운 교실 환경에서의 적용 가능성 탐구