



CHSER: A Dataset and Case Study on Generative Speech Error Correction for Child ASR

Natarajan Balaji Shankar¹, Zilai Wang¹, Kaiyuan Zhang¹, Mohan Shi¹, Abeer Alwan¹

¹Dept of Electrical and Computer Engineering, University of California, Los Angeles, USA

balaji11312, zilaiwang2001, kaiyuanzhang, shimohan@ucla.edu, alwan@ee.ucla.edu

Abstract

Automatic Speech Recognition (ASR) systems struggle with child speech due to its distinct acoustic and linguistic variability and limited availability of child speech datasets, leading to high transcription error rates. While ASR error correction (AEC) methods have improved adult speech transcription, their effectiveness on child speech remains largely unexplored. To address this, we introduce CHSER, a Generative Speech Error Correction (GenSEC) dataset for child speech, comprising 200K hypothesis-transcription pairs spanning diverse age groups and speaking styles. Results demonstrate that fine-tuning on the CHSER dataset achieves up to a 28.5% relative WER reduction in a zero-shot setting and a 13.3% reduction when applied to fine-tuned ASR systems. Additionally, our error analysis reveals that while GenSEC improves substitution and deletion errors, it struggles with insertions and child-specific disfluencies. These findings highlight the potential of GenSEC for improving child ASR.¹

Index Terms: Child Automatic Speech Recognition, Error Correction, Disfluency Detection

1. Introduction

Speech Foundation Models (SFMs) have advanced Automatic Speech Recognition (ASR), demonstrating state-of-the-art performance and strong zero-shot generalization [1, 2, 3, 4, 5]. Their success stems from large-scale pretraining and transformer-based architectures, yet they struggle with child speech, exhibiting significantly high error rates [6]. The performance gap between child and adult ASR arises from the distinct characteristics of child speech, including higher pitch, greater intra- and inter-speaker variability, and developmental changes [7, 8, 9, 10]. Additionally, child speech contains greater disfluencies (e.g., repetitions, restarts), grammatical inconsistencies, and non-standard word usage, which degrade ASR accuracy [11]. Variability across speaking styles, such as scripted vs. spontaneous speech, further complicates generalization, necessitating effective post-processing to enhance transcription accuracy.

Traditional ASR error correction (AEC) methods focus on rescoring and hypothesis selection, including n-gram-based rescoring [12], consensus decoding [13], and neural rescoring with RNNs and transformers [14, 15]. More recent AEC approaches leverage machine translation models for domain adaptation [16], fine-tuned RoBERTa models for readability improvements [17], and crossmodal methods incorporating both audio and text representations [18]. Confidence-based correction strategies also integrate utterance rejection for improved

robustness [19].

Generative speech error correction (GenSEC) has emerged as a promising approach, leveraging large language models (LLMs) to refine ASR outputs by learning error patterns from large-scale paired data. Datasets such as HyPoradise [20] and RobustHyPoradise [21] serve as benchmarks, demonstrating that fine-tuned generative models significantly enhance transcription accuracy. GenSEC exploits full N-best lists for improved hypothesis selection. LoRA-based [22] fine-tuning of Llama 2 [23] and Flan T5 [24] models enables efficient adaptation [20, 25] without requiring full retraining. Other approaches include language-space noise embeddings for robustness [21], integration of ASR embeddings with LLMs [26], lattice-constrained decoding [27], retrieval-augmented generation for out-of-domain correction [28], and multi-expert models for ASR and speech translation [29]. However, these approaches are trained on adult speech, limiting their effectiveness for child-specific disfluencies and linguistic patterns.

To address this gap, we introduce CHSER, a GenSEC dataset designed to enhance ASR transcription accuracy for child speech. Our dataset comprises 200k child speech ASR hypothesis pairs across diverse age groups (4–14 years) and speaking styles (scripted, spontaneous), enabling models to learn child-specific error patterns. We evaluate existing GenSEC methods to establish baselines and analyze their limitations in handling child speech variability. Furthermore, we assess the adaptability of child-specific GenSEC models in both zero-shot and fine-tuned settings, demonstrating their impact on improving child ASR performance across different domains.

Our contributions can be summarized as follows:

- We introduce the first large-scale dataset for child ASR Generative Speech Error Correction, comprising 200K hypothesis-transcription pairs spanning multiple age groups (4–14 years old) and speaking styles (scripted and spontaneous).
- Through systematic evaluation, we address four key research questions: **(RQ1)** The effectiveness of existing GenSEC methods on child speech, **(RQ2)** The robustness of child-specific error correction models across different domains, **(RQ3)** The applicability of GenSEC models to fine-tuned ASR systems of differing architectures, and **(RQ4)** Whether GenSEC alters the nature of ASR errors by analyzing error distribution shifts post correction.²

¹This work was supported in part by the NSF.

²Our dataset, code, and models are available at <https://github.com/balaji11312/CHSER>

2. CHSER Dataset

To construct a benchmark for child speech ASR error correction, we compile a dataset from multiple publicly available child speech corpora, ensuring diversity in age ranges and speaking styles. Our dataset consists of 200K ASR hypothesis-transcription pairs sourced from four major child speech datasets: MyST [30], CMU Kids [31], CSLU OGI Kids [32], and CHILDES English OCSC Corpus [33]. Similar to the procedure used for adult speech in [20], we generated N-best lists of sentence hypotheses using the Whisper-base.en model [4]. To improve hypothesis-transcription pair quality, we apply the following filtering steps: remove repeated utterances, retain only the top 5 hypotheses with the highest probabilities in the N-best list, exclude transcriptions shorter than three words, and discard utterances containing words outside the Whisper tokenizer vocabulary. The CHSER dataset in total contains over 200K ASR hypothesis-transcription pairs and is summarized in Table 1. All dataset divisions into train, dev, and test sets follow the original splits in their respective source datasets. Below, we provide an overview of each source dataset:

The My Science Tutor (MyST) corpus [30] is a 400-hour English speech dataset collected from 1,371 students in grades 3–5 (ages 8–11). It features spontaneous, unscripted conversations with a virtual science tutor on topics such as physics, geography, and biology. Of this, we include 64,187 utterances in the train set, 10,458 in the dev set, and 11,591 in the test set.

The CMU Kids Corpus [34] consists of read-aloud sentences from 6–11-year-old children. It includes 24 male and 52 female speakers with a total of 5,180 utterances recorded over 9 hours of speech. We incorporate 3,784 utterances in the train set, 642 in the dev set, and 754 in the test set.

The CSLU OGI Kids Corpus [32] contains speech from 1,100 children (ages 4–14) across both scripted and spontaneous speech conditions. The read speech section consists of simple words, sentences, and digit strings. The spontaneous speech section includes alphabet recitation and unscripted monologues. We utilize both sections in CHSER, with 18,121 utterances (scripted) and 3,730 utterances (spontaneous) in the train set, 1,971 and 451 in the dev set, and 5,795 and 457 in the test set, respectively.

The OCSC corpus contains 156 hours of speech from 303 primarily monolingual English speaking children aged 4–9. The dataset includes a mix of scripted and spontaneous speech, covering structured tasks such as reading passages, and letter and number identification, as well as open-ended tasks like picture descriptions and conversational interactions. As there is no official split, we incorporate 64,581 utterances in the train set, 7,875 in the dev set, and 8,090 in the test set ensuring no speaker overlap between the subsets.

Table 1: Statistics of the CHSER dataset split by source, speaking style, and age range. The CHSER dataset consists of hypotheses-transcription pairs generated using Whisper-base.en in a zero-shot beam search setting.

Source	Style	Age Range	Train # Pairs	Dev # Pairs	Test # Pairs
MyST	Spontaneous	8–11	64,187	10,458	11,591
CMU Kids	Scripted	6–11	3,784	642	754
CSLU OGI	Spontaneous	4–14	3,730	451	457
	Scripted	4–14	18,121	1,971	5,795
OCSC	Scripted/Spontaneous	4–9	64,581	7,875	8,090
All			154,403	21,397	26,687

3. Experimental Setup and Results

3.1. RQ1: Do existing GenSEC methods perform well on child speech?

Table 2: WER of Traditional Reranking Models (Trigram LM, Transformer LM) and oracle WER (N-best Oracle (o_{nb}), and Compositional Oracle (o_{cp})) on the CHSER dataset. Baseline refers to WER before error correction.

Dataset	Baseline	Trigram	Transformer	o_{nb}	o_{cp}
All	30.5	29.2	29.9	17.7	8.6
MyST	28.1	27.0	29.9	17.2	8.1
CMU Kids	18.7	13.0	12.2	10.3	7.9
OGI Spontaneous	40.6	36.4	39.6	25.6	10.4
OGI Scripted	26.2	16.5	15.7	11.2	7.9
OCSC	43.1	48.9	40.2	23.3	11.2

Table 3: WER of LLM-based GenSEC models pre-trained on the HyParadise adult speech dataset (Llama Hyp, T5 Hyp) and fine-tuned on CHSER (Llama FT, T5 FT) evaluated on the CHSER dataset. Baseline refers to WER before error correction. Bold face numbers indicate best results. Statistically significant results ($p < 0.05$) are indicated with *. Baseline definition and notation conventions apply to all subsequent WER tables

Dataset	Baseline	Llama Hyp	T5 Hyp	Llama FT	T5 FT
All	30.5	31.0	27.3	24.8	21.8*
MyST	28.1	27.2	26.4	22.5	20.7*
CMU Kids	18.7	16.6	15.5	13.6	12.9*
OGI Spontaneous	40.6	37.6	36.3	32.8	29.7*
OGI Scripted	26.2	30.4	26.1	16.2	13.3*
OCSC	43.1	47.9	40.7	40.6	32.2*

Table 4: WER of In-Context Learning with GPT-4o mini with different prompting examples ($n=0$, $n=1$, $n=5$) on the CHSER dataset.

Dataset	Baseline	$n=0$	$n=1$	$n=5$
All	30.5	42.1	27.3	27.4
MyST	28.1	34.1	25.5	25.0
CMU Kids	18.7	28.9	13.8*	16.4
OGI Spontaneous	40.6	55.5	38.3	35.7*
OGI Scripted	26.2	40.0	19.6*	23.0
OCSC	43.1	77.0	39.6*	40.2

To assess the effectiveness of Generative Speech Error Correction (GenSEC) methods on child speech, we evaluate commonly used approaches including Llama 2 [23], Flan T5 [24], and In-Context Learning (ICL) using baselines established in [35, 20]. We calculate WER and assess statistical significance using the sctk package from the NIST Scoring Toolkit [36] to ensure a fair comparison across methods.

• Traditional Reranking-Based Models:

We include two language model baselines that perform log-probability score-based error correction: (i) Trigram LM: A trigram language model trained on in-domain textual data from the CHSER train set, serving as a strong traditional baseline. (ii) Transformer LM: A transformer-based language model pre-trained on the LibriSpeech LM corpus [37], then fine-tuned on in-domain text from the CHSER train set. The effectiveness of these models in correcting errors is reported in Table 2. Additionally, we report two oracle WER values as theoretical performance upper bounds: (i) Oracle-NB (o_{nb}): The WER of the best candidate in the N-best hypothesis list. (ii) Oracle-CP (o_{cp}): The lowest possible WER achievable based on an optimal composition of tokens across multiple hypotheses.

- **Fine-Tuned GenSEC Models:**

To evaluate the generalization of GenSEC models to child speech, we assess Flan T5 large [24] and Llama 2 7B [23] models pre-trained on HyPoradise [20], a dataset designed for adult speech correction. Additionally, we instruction fine-tune Flan T5 large and Llama 2 7B models on CHSER using 8-bit quantization [38] with LoRA adapters [22] for 5 epochs. The results for these models are presented in Table 3.

- **In-Context Learning (ICL) with GPT-4o mini:**

We investigate In-Context Learning (ICL) [39] as an alternative to fine-tuning, evaluating GPT-4o mini [40] for error correction. Three prompting strategies are tested: (i) n=0 (Zero-shot prompting), where the model is provided only the ASR output and no examples; (ii) n=1 (One-shot prompting), where one correction example is provided; and (iii) n=5 (Few-shot prompting), where five examples are provided before generating corrections. The performance of GPT-4o mini across these strategies is shown in Table 4.

Our findings, summarized in Tables 2, 3, and 4, demonstrate key trends in child speech error correction, highlighting the strengths and limitations of different approaches.

As shown in Table 2, traditional reranking-based models (Trigram LM and Transformer LM) provide only minor WER improvements over the baseline ASR model. The Trigram LM performs slightly better than the Transformer LM across most datasets. However, these methods fail to generalize effectively, with limited impact on spontaneous speech datasets such as OGI Spontaneous and OCSC, where the WER remains significantly high. Oracle analyses suggest that optimal hypothesis selection and compositional error correction could yield substantial additional improvements, though no current method reaches these theoretical bounds.

Table 3 demonstrates that fine-tuned LLM-based GenSEC models (T5 FT, Llama FT) significantly outperform both methods in table 2, and pre-trained models. The Flan T5 model fine-tuned on the CHSER dataset (T5 FT) achieves the lowest WER across all datasets, with a relative WER reduction of 28.5% over the baseline ASR model. Notably, T5 FT reduces absolute WER by 12.9% on OGI Scripted, one of the largest improvements observed, demonstrating that structured read speech benefits the most from fine-tuning. Performance remains weaker on spontaneous datasets, such as OCSC, where WER remains at 32.2%, indicating that modeling child speech variability remains a challenge.

Table 4 presents the WER results of ICL-based correction using GPT-4o mini, tested with different prompting strategies (zero-shot, one-shot, few-shot). Zero-shot prompting (n=0) performs significantly worse than all other methods, often increasing WER rather than reducing it. One-shot and Few-shot prompting improves on the baseline, but are still weaker than fine-tuned models, demonstrating that ICL, while useful, is not a substitute for fine-tuning when domain adaptation is necessary.

3.2. RQ2: Are child error correction models robust across different child specific domains?

To assess the effectiveness of GenSEC models on child speech, we examine whether a model trained on one subset of the CHSER dataset can generalize to other domains. Specifically, we compare the WER of three systems in Table 5: A baseline ASR model, which provides the uncorrected transcriptions (Baseline), A T5 model fine-tuned on the entire CHSER dataset

Table 5: WER across T5 Fine-Tuned on CHSER (T5 FT CHSER), and T5 Fine-Tuned on MyST subset of CHSER (T5 FT MyST), on the CHSER dataset.

Dataset	Baseline	T5 FT CHSER	T5 FT MyST
All	30.5	21.8*	22.4
MyST	28.1	20.7	19.9*
CMU Kids	18.7	12.9*	16.2
OGI Spontaneous	40.6	29.7	29.2
OGI Scripted	26.2	13.3*	17.7
OCSC	43.1	32.2*	35.7

(T5 FT CHSER), A T5 model fine-tuned only on the MyST subset of CHSER (T5 FT MyST). All T5 models are instruction fine-tuned in settings similar to those in Section 3.1.

Our results indicate that T5 FT MyST performs well on its in-domain data (MyST) but struggles to generalize to other scripted datasets, such as CMU Kids and OGI Scripted, where its WER increases significantly. This suggests that models fine-tuned on a single child speech domain capture domain-specific characteristics effectively but lack the adaptability needed for diverse datasets.

In contrast, T5 FT CHSER consistently achieves lower WER across most domains, demonstrating that training on a diverse dataset improves generalization. These results highlight the importance of multi-domain training for GenSEC models, as fine-tuning on a single dataset may lead to overfitting and reduced effectiveness in unseen child speech domains.

3.3. RQ3: Do error correction models offer robustness for fine-tuned ASR models of different architectures?

Table 6: WER reduction with a T5 model fine-tuned on the MyST subset of CHSER, evaluated across different ASR models on the MyST test set. The ASR hypotheses are generated using versions of Whisper-tiny.en (Whisper Tiny), Whisper-small.en (Whisper Small), and WavLM-large (WavLM) fine-tuned on the MyST train set.

ASR Model	Baseline	T5 Corrected
WavLM	10.1	9.2*
Whisper Small	11.4	10.3*
Whisper Tiny	12.8	11.1*

Since the CHSER dataset is generated from beam search hypotheses of a zero-shot Whisper-base.en model, an important question is whether WER improvements from error correction generalize to different ASR models, including those fine-tuned on child speech.

To evaluate this, as an example we test a Flan T5 model fine-tuned on the MyST subset of CHSER (instruction fine-tuned in settings similar to those in Section 3.1) against transcriptions from three ASR models, as shown in Table 6: Whisper-tiny.en (attention encoder decoder (AED), 39M parameters, pretrained on 680k hours), Whisper-small.en (AED, 244M parameters, pretrained on 680k hours), and WavLM-large (CTC encoder-only model, 316M parameters, pretrained on 94k hours), all fine-tuned on the MyST dataset. We note that our reported WER for the beam search decoding of Whisper-based models is higher than the WER obtained via greedy decoding as reported in the literature [6, 41], likely due to the additional temperature setting applied during beam search to encourage more diverse output generation.

The results, presented in Table 6, demonstrate that the Flan T5 model fine-tuned on the MyST section of the CHSER dataset provides consistent WER (9-13.3%) reductions across all fine-tuned ASR models, demonstrating its robustness in improv-

ing transcription accuracy irrespective of the underlying ASR model.

3.4. RQ4: Are the errors present after error correction different from errors before correction?

While GenSEC improves transcription accuracy, it is equally important to examine whether the nature of transcription errors shifts post-correction. We categorize errors in the original CHSER dataset and those remaining after correction using the best performing model from Section 3.1 (T5 FT), and conduct the following analysis:

- **Error Type Distribution:**

We analyze errors using the insertion, substitution, and deletion (I/S/D) framework to determine whether the fine-tuned T5 model introduces biases toward specific types of corrections. Figure 1 shows that substitutions are the most frequent error type both before and after correction. While GenSEC significantly reduces substitutions and deletions, it struggles to correct insertion-based errors, suggesting that erroneous insertions in ASR outputs are more challenging to remove than missing or misrecognized words.

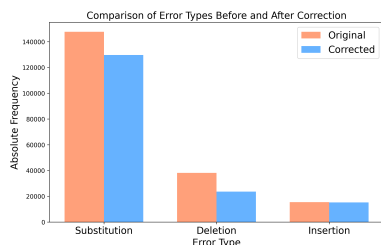


Figure 1: Comparison of error types (substitutions, deletions, insertions) before and after error correction. Errors are categorized by comparing ASR hypotheses to ground truth transcriptions.

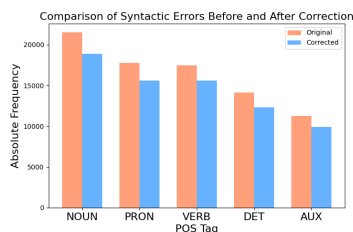


Figure 2: Syntactic errors categorized by part-of-speech (POS) tags (NOUN, PRON, VERB, DET, AUX) before and after error correction. Errors are identified by comparing ASR hypotheses to ground truth transcriptions.

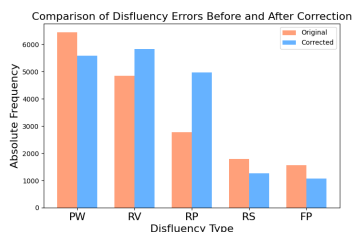


Figure 3: Disfluency errors - filled pauses (FP), partial words (PW), repetitions (RP), revisions (RV), and restarts (RS), before and after error correction. Disfluencies are identified in the ground truth transcriptions, ASR hypotheses, and corrected outputs to track correction rates.

- **Syntactic Errors:**

To evaluate whether the fine-tuned T5 model introduces syntactic shifts, we categorize substitution-based errors by part-

of-speech (POS) tags using spaCy [42]. This analysis is conducted by aligning the ground truth transcription to the hypothesis pre- and post-correction, and focuses specifically on words that were substituted. The results, shown in Figure 2, reveal a consistent reduction in syntactic errors across the top five most frequent POS categories.

- **Disfluency Errors:**

Disfluencies are a common feature of child speech, and we analyze their detection and correction by first identifying the number of disfluencies present in the ground truth transcripts. We then compare this with the number detected in the ASR hypotheses, tracking the detection error rate for each category: filled pauses (FP), partial words (PW), repetitions (RP), revisions (RV), and restarts (RS) using the procedure detailed in [43]. Finally, we evaluate how these errors change post-correction to determine whether GenSEC models mitigate or amplify disfluency related errors. Figure 3 shows that while the fine-tuned T5 model improves the detection rate for partial words, filled pauses and restarts, it struggles with repetitions and revisions, leading to an increase in their error rates after correction. This suggests that GenSEC models primarily focus on syntactic and lexical refinements but remain challenged by disfluency recognition, potentially misclassifying them or introducing new errors.

4. Conclusion

This study presents CHSER, a Generative Speech Error Correction (GenSEC) dataset comprising 200K hypothesis-transcription pairs spanning diverse age groups and speaking styles specifically designed to address the challenges of ASR transcription for child speech. Through a comprehensive evaluation, we systematically investigated four key research questions:

- (RQ1) Effectiveness of existing GenSEC methods: We found that pre-trained GenSEC models (LLaMA 2, Flan T5), originally developed for adult speech, struggle to generalize to child speech, highlighting the need for child-specific adaptation. Fine-tuning on CHSER significantly improved WER across multiple child speech datasets, with a fine-tuned Flan T5 model achieving a 28.5% relative WER reduction in a zero-shot setting, demonstrating better generalization across diverse domains.
- (RQ2) Robustness of child-specific correction models across domains: Models fine-tuned on a single dataset perform well in-domain but struggle to generalize, especially to datasets with different speaking styles.
- (RQ3) Applicability to fine-tuned ASR models: When applied to fine-tuned ASR systems with differing architectures, GenSEC remained effective, achieving a 13.3% relative WER reduction.
- (RQ4) Shifts in error distribution post-correction: Our error analysis revealed that while GenSEC significantly reduces substitutions and deletions, it struggles with insertion errors and child-specific disfluencies such as repetitions and revisions, underscoring the linguistic challenges unique to child ASR.

These findings underscore the potential of GenSEC to improve the performance of child ASR. Future directions include exploring hybrid approaches that integrate language modeling with prosodic and phonetic cues to improve GenSEC’s handling of disfluencies present in child speech.

5. References

- [1] S. Chen *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [2] W.-N. Hsu *et al.*, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [3] A. Baeovski *et al.*, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [4] A. Radford *et al.*, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning, ICML 2023*, vol. 202, 2023, pp. 28 492–28 518.
- [5] K. C. Puvvada *et al.*, “Less is more: Accurate speech recognition & translation without web-scale data,” in *Interspeech 2024*, 2024.
- [6] R. Fan *et al.*, “Benchmarking children’s asr with supervised and self-supervised speech foundation models,” in *Interspeech 2024*, 2024.
- [7] M. Gerosa *et al.*, “Analyzing children’s speech: An acoustic study of consonants and consonant-vowel transition,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006.
- [8] S. Lee *et al.*, “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters,” *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 03 1999.
- [9] G. Yeung and A. Alwan, “On the difficulties of automatic speech recognition for kindergarten-aged children,” in *Interspeech 2018*, 2018, pp. 1661–1665.
- [10] S. Dutta *et al.*, “Challenges remain in building ASR for spontaneous preschool children speech in naturalistic educational environments,” in *Interspeech 2022*, 2022, pp. 2706–2710.
- [11] C. T. Zackheim and E. G. Conture, “Childhood stuttering and speech disfluencies in relation to children’s mean length of utterance: A preliminary study,” *Journal of Fluency Disorders*, vol. 28, no. 2, pp. 115–142, 2003.
- [12] S. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 3, pp. 400–401, 2003.
- [13] L. Mangu *et al.*, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.
- [14] T. Mikolov *et al.*, “Recurrent neural network based language model,” in *Interspeech 2010*, vol. 2, no. 3, 2010, pp. 1045–1048.
- [15] T. Tanaka *et al.*, “Neural error corrective language models for automatic speech recognition,” in *Interspeech 2018*, 2018.
- [16] A. Mani *et al.*, “Asr error correction and domain adaptation using machine translation,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6344–6348.
- [17] J. Liao *et al.*, “Improving readability for automatic speech recognition transcription,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 5, May 2023.
- [18] B. Lin and L. Wang, “Multi-modal asr error correction with joint asr error detection,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.
- [19] J. Du *et al.*, “Cross-modal asr post-processing system for error correction and utterance rejection,” *arXiv preprint arXiv:2201.03313*, 2022.
- [20] C. Chen *et al.*, “Hyporadise: An open baseline for generative speech recognition with large language models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [21] Y. Hu *et al.*, “Large language models are efficient learners of noise-robust speech recognition,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [22] E. J. Hu *et al.*, “Lora: Low-rank adaptation of large language models,” *International Conference on Learning Representations*, 2021.
- [23] H. Touvron *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [24] H. W. Chung *et al.*, “Scaling instruction-finetuned language models,” *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
- [25] M. La Quatra *et al.*, “Flanec: Exploring flan-t5 for post-asr error correction,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 608–615.
- [26] S. Radhakrishnan *et al.*, “Whispering LLaMA: A cross-modal generative error correction framework for speech recognition,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Dec. 2023.
- [27] R. Ma *et al.*, “Asr error correction using large language models,” *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
- [28] S. Ghosh *et al.*, “Failing forward: Improving generative error correction for asr with synthetic data and retrieval augmentation,” *arXiv preprint arXiv:2410.13198*, 2024.
- [29] Y.-T. Lin *et al.*, “Neko: Toward post recognition generative correction large language models with task-oriented experts,” *arXiv preprint arXiv:2411.05945*, 2024.
- [30] S. Pradhan *et al.*, “My science tutor (MyST)—a large corpus of children’s conversational speech,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, May 2024, pp. 12 040–12 045.
- [31] M. Eskenazi *et al.*, “The cmu kids corpus,” *Linguistic Data Consortium*, vol. 11, 1997.
- [32] K. Shobaki *et al.*, “The ogi kids speech corpus and recognizers,” in *6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. vol. 4, 258–261.
- [33] L. Wagner *et al.*, “The ohio child speech corpus,” *Speech Communication*, vol. 170, p. 103206, 2025.
- [34] M. S. Eskenazi, “Kids: A database of children’s speech,” *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2759–2759, 1996.
- [35] C.-H. H. Yang *et al.*, “Generative speech recognition error correction with large language models and task-activating prompting,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.
- [36] J. Fiscus, “SCTK: The NIST Scoring Toolkit,” National Institute of Standards and Technology, 2007, [Software].
- [37] V. Panayotov *et al.*, “Librispeech: An ASR corpus based on public domain audio books,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.
- [38] B. Jacob *et al.*, “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.
- [39] S. M. Xie *et al.*, “An explanation of in-context learning as implicit bayesian inference,” in *International Conference on Learning Representations*, 2022.
- [40] A. Hurst *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.
- [41] A. A. Attia *et al.*, “Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 74–80.
- [42] M. Honnibal *et al.*, “spacy: Industrial-strength natural language processing in python,” 2020, available at <https://spacy.io/>.
- [43] A. Romana *et al.*, “Automatic disfluency detection from untranscribed speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.