Designing an AI-Assisted Speech-Based Interactive System for Home-Based Sentence Practice for Children

Elizabeth Wiafe[1], **Frank Rudzicz**[2], **Lizbeth Escobedo**[3]

[1]Faculty of Computer Science – Dalhousie University
Halifax, Canada

el751593@dal.ca, frank@dal.ca, lizbeth.escobedo@dal.ca

Abstract. *Expressive sentence construction is foundational to children's academic development, yet opportunities for structured practice and receiving feedback outside clinical environments remain limited. This work-in-progress paper presents an AI-assisted speech-based interactive prototype designed to support home-based sentence formulation through multimodal prompts and automated feedback. The system integrates visual stimuli, guided target words, local speech recognition, and structured grammar-aware analysis that evaluates sentence completeness, target word use, complexity level, and contextual relevance through a conjunctive decision rule to provide supportive, non-diagnostic feedback. We describe the system architecture and outline a proposed pilot study with children aged 6-12, complemented by expert input from speech-language pathologists, to evaluate feasibility, usability, and perceived educational value.*

1. Introduction

The construction of expressive sentences is critically important for academic success, literacy acquisition, and active involvement in class activities by children. Constructing complete sentences requires the complex integration of meaning, grammar, and context and draws on active working memory and linguistic awareness [Thompson and Shapiro 2007]. As sentence formulation becomes increasingly complex, it becomes necessary for children to deal with grammatical markers, connect clauses in an appropriate manner, and modify language according to communicative contexts.

Recent developments in artificial intelligence (AI) have expanded possibilities for the design and deployment of technology-based educational tools [Tan et al. 2025]. Speech-based technology can help people formulate sentences in daily settings. The development of speech technology, supported by Natural Language Processing (NLP), has created opportunities for interactive learning tools. NLP refers to computational methods for processing and analyzing human languages, including syntactic parsing and semantic analysis [Mathew et al. 2021]. Speech technology, combining automatic speech recognition (ASR) and NLP, has been used in research-based educational systems to support children's language development [Bai et al. 2023].

This paper outlines the design and development of the AI-assisted, speech-based interactive prototype that enables children to practice sentence formulation. The system provides children with multimodal prompts that vary in sentence-structure complexity, including simple, compound, and complex sentence types. The research makes three significant contributions: the design of an interactive system for structured speech, where AI assists in the formulation of sentences, the design of automated feedback for assessing

the completeness of sentences and complexity level, while maintaining a friendly interaction and third, it integrates sentence formulation tasks within a speech-based environment specifically designed for children's expressive language practice.

2. Related Work: AI-Assisted Educational Systems

Recent research reveals an increasing focus on AI-based educational technology, particularly in personalized learning and the provision of instant feedback for learners [Kamalov et al. 2023]. Systematic reviews show a wide range of AI-learning tools, that personalize instructions, monitor student progress, and provide instant feedback on learners' queries and activities [Sun et al. 2026].

The Reading Tutor system of Project LISTEN was developed to improve children's oral fluency by using ASR and NLP techniques to detect reading errors and provide immediate corrective feedback[Molenaar et al. 2023]. The system also tracked learner performance across sessions, enabling progress monitoring over time. Similarly, AI-based tutoring systems such as EBS AI Peng Talk integrate end-to-end ASR with automatic proficiency evaluation to provide learners with pronunciation and fluency scores [Kang et al. 2024]. The system extracts word and phoneme-level fluency features and predicts pronunciation proficiency scores using regression and neural models. These systems contribute evidence that ASR can be integrated into instructional platforms to support performance tracking and immediate response generation. However, their evaluation mechanisms primarily target pronunciation accuracy, word recognition, or reading fluency rather than structured analysis of sentence-level grammatical complexity or clause organization.

While prior systems have incorporated speech recognition for pronunciation and fluency assessment, fewer systems explicitly evaluate structured sentence completeness, connector usage, and syntactic progression in children's spoken responses within a unified framework. The proposed system addresses this gap by analyzing children's spoken responses at the sentence level and providing structured feedback that supports progressive sentence development.

3. System Design and Methodology

3.1. System Architecture

The system is implemented as a speech-based, modular interactive system designed to support the construction of structured sentences in children. As illustrated in Figure 1, the architecture consists of four primary modules: User Interaction, the ASR Module, the NLP & Feedback Module, and Response to User, supported by internal resources including Target Words, Grammar Rules, and Feedback History

The interaction flow proceeds as follows. The child engages with the Prompt Interface under User Interaction. The spoken response is processed by the ASR Module, which converts speech into text. The transcript is then evaluated within the NLP & Feedback Module. Based on this evaluation, feedback is generated and delivered through the Response to User module.

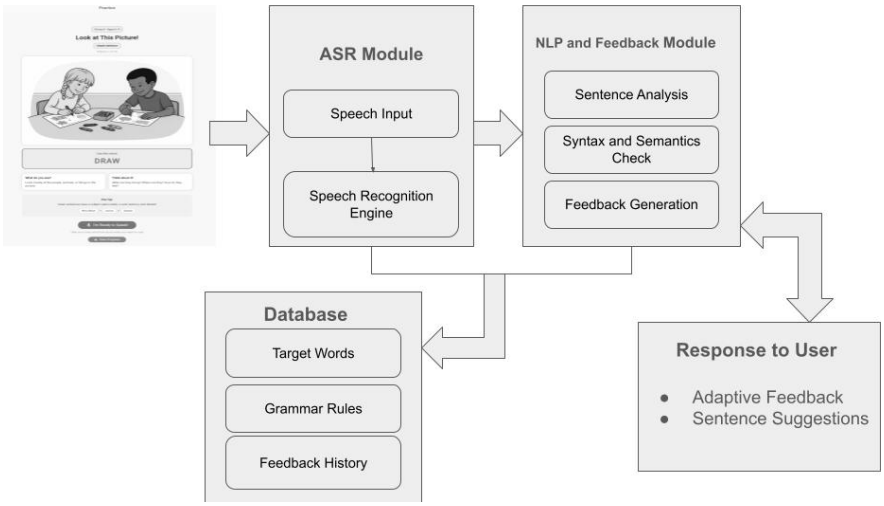


Figure 1. System Architecture of the Proposed AI-Assisted Speech-Based Sentence Practice System.

3.2. User Interface and Prompt Interface

The Prompt Interface presents a multimodal learning task consisting of a visual scene and a target word that must be incorporated into a sentence. The system design draws inspiration from traditional speech and language therapy assessment tools such as the Clinical Evaluation of Language Fundamentals (CELF-5) [Usha and Alex 2023], the Expressive Vocabulary Test (EVT-2), and the Goldman-Fristoe Test of Articulation (GFTA-3), etc.

Target words typically include conjunctions such as *because*, *if*, or *when*, which encourage causal or conditional sentence constructions. Tasks progressively increase in structural complexity, beginning with simple sentences and advancing to compound and complex forms. This is a type of scaffolded learning that gradually increases in difficulty.

3.3. ASR Module

The ASR Module receives speech input from the child and processes it using a Speech Recognition Engine powered by Whisper. The audio signal is transcribed into text, which becomes the input to the NLP & Feedback Module. This separation ensures that speech processing and linguistic evaluation remain modular and independently adaptable.

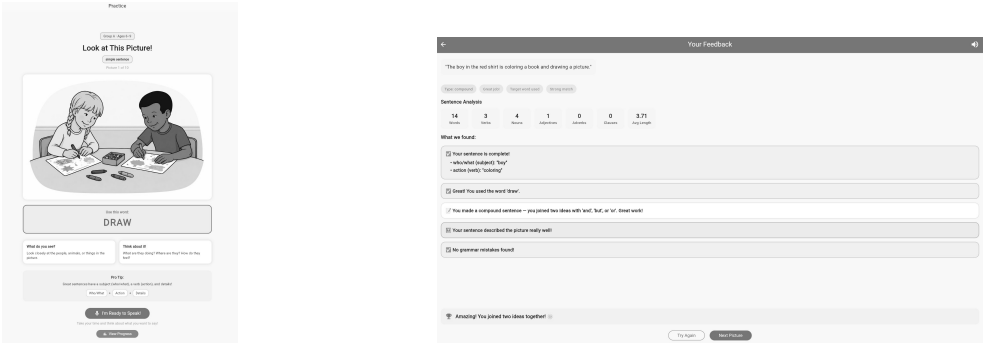
3.4. NLP & Feedback Module

The NLP & Feedback Module performs structured linguistic evaluation rather than full sentence rewriting. It consists of three internal components: Sentence Analysis, Syntax & Semantics Check, and Feedback Generation. Using tokenization, part-of-speech tagging, and dependency parsing (e.g., via spaCy), the Sentence Analysis component evaluates structural properties of the transcript. Specifically, the system checks: Target word presence, using lemma matching so inflected forms (e.g., *drawing*, *drew*) are correctly recognised; Correct connector usage, verifying that conjunctions appropriately link clauses; Sentence completeness, detected through subject (`nsubj` dependency) and finite verb (VERB or AUX POS tag) identification; Sentence complexity classification via connector presence: compound (and/but/or) and complex (because/if/when/although/since/while), verified against the expected complexity level for each prompt; Context match with the

image prompt, using lemma overlap between the child's response and a predefined set of scene-relevant keywords associated with the visual stimulus; Grammatical correctness, referencing predefined grammar rules; and a minimum word count threshold, where responses of fewer than three words are automatically rejected.

3.5. Response to User

The Feedback Generation component produces adaptive feedback that is delivered through the Response to User module. All six conditions apply a conjunctive decision rule: all must pass for an accept; a single failure triggers a retry with targeted feedback identifying the specific condition that was not met. If structural requirements are satisfied, positive feedback is provided. If required elements are missing or misused, sentence suggestions and guided prompts are generated to support revision while maintaining a supportive tone. Feedback explicitly names identified elements (e.g., subject: *child*; verb: *is*) to remain accessible across ages. "Extra information" refers to descriptive content beyond the required subject–verb structure. Figure 2 illustrates the Prompt Interface and the adaptive feedback response presented to the user.



(a) Prompt screen **(b)** Feedback screen

Figure 2. Interface screens of the proposed system: (a) example prompt screen showing a visual scene and target word; (b) example feedback screen showing supportive response after sentence analysis.

4. Conclusions, Limitations, and Future Work

The purpose of this work is to present a prototype that shows how speech based interaction can aid in the practice of building structured sentences outside a clinical setting. Unlike other educational tools that focus on vocabulary and pronunciation, this tool focuses on sentence building. It transforms elements of structured language therapy into an interactive tool for children to use independently. The key design aspect is how it encourages children to incorporate target words into sentences using visual scene prompts, where the image provides scene-relevant cues that guide lemma-based keyword matching without prescribing a single correct response. For example, using conjunctions and conditional markers encourages children to build sentences that express cause and condition. The second key point is that feedback is framed educationally rather than diagnostically; the system highlights structural patterns to guide revision rather than to assess clinical need. Instead of the system providing scores, the prototype offers prompts that encourage children to improve their responses.

One such aspect that needs to be enhanced in the future is how accurately it can identify speech uttered by children. Although it has been observed that advanced speech recognition technology like Whisper works effectively in most cases, speech uttered by children may have pronunciation difficulties, shorter speech segments, and background noise, which may lead to incorrect transcription and subsequently lead to incorrect grammatical analysis.

Our future work includes an evaluation study to help us with early evidence on how children interact with this speech-based tool for sentence construction practice. We will be able to determine whether the sentences constructed from the prompts are getting more and more complex, and whether the feedback system is easy to understand for children across varying ages and grammatical familiarity. Looking forward, there are plans to make the transcription and analysis process for children's speech more robust. There are also plans to develop adaptive feedback systems that adjust both difficulty and feedback presentation style based on a child's performance. Future developments may also incorporate session-level pattern tracking to identify persistent structural gaps, enabling adaptive scaffolding without crossing into diagnostic assessment. There are also plans to embed the system within broader educational or therapeutic contexts to support expressive language practice outside of a therapy session.

References

Bai, Y., García, C., Hubers, F., Cucchiarini, C., and Strik, H. (2023). An asr-based tutor for learning to read: How to optimize feedback to first graders. *ArXiv*, abs/2306.04190.

Kamalov, F., Calonge, D. S., and Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability*.

Kang, B., Jeon, H.-B., and Lee, Y. (2024). Ai-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation. *ETRI Journal*, 46:48 – 58.

Mathew, A., Paulose, J., and Info, A. (2021). Nlp-based personal learning assistant for school education. *International Journal of Electrical and Computer Engineering*, 11:4522–4530.

Molenaar, B., García, C., Strik, H., and Cucchiarini, C. (2023). Automatic assessment of oral reading accuracy for reading diagnostics. In *Interspeech*.

Sun, L., Xu, W., and Gao, Z. (2026). A Human-Centered Privacy (HCP) Approach to AI. In *Handbook of Human-Centered Artificial Intelligence*, pages 1–47. Springer.

Tan, L. Y., Hu, S., Yeo, D. J., and Cheong, K. H. (2025). Artificial Intelligence-Enabled Adaptive Learning Platforms: A Review. *Computers and Education: Artificial Intelligence*, 9:100429.

Thompson, C. K. and Shapiro, L. P. (2007). Complexity in Treatment of Syntactic Deficits. *Complexity*.

Usha, G. P. and Alex, J. (2023). Speech assessment tool methods for speech impaired children: A systematic literature review on the state-of-the-art in speech impairment analysis. *Multimedia Tools and Applications*, pages 1 – 38.