International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

MILO: Mimicking and Interactive Learning with Oration

Prof. Shubhangi Bhagat[1], Priti Panhale[2], Tanmaya Panchariya[3], Vedanti Kshirsagar[4], Vishakha Deshmukh[5]

[1]Assistant Professor, Department of Computer Engineering, TSSM's BSCOER, Narhe, Pune, Maharashtra, India
[2,3,4,5]Student, Department of Computer Engineering, TSSM's BSCOER, Narhe, Pune, Maharashtra, India

---

Abstract - *MILO (Mimicking and Interactive Learning with Oration) is a voice-based learning assistant with multilingual fallback support designed for children in the age group of 4-10 years. The voice assistants that are presently used are designed for adults and have boring interfaces for children to be used. It is also observed that these assistants have difficulty in understanding child speech. MILO solves these issues by introducing a natural interactive voice assistant specifically designed for children with child friendly interfaces which help them in learning in a playful manner. The system integrates offline Speech-to-Text (STT), Semantic Regeneration, Zero-Shot AI Voice Cloning, and Digital Signal Processing (DSP) within an interactive, 3D setting. It uses Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) to convert input speech into meaningful questions and gives responses accordingly. It has two modes, an entertainment mode for curiosity, interaction and playful activities and an educational mode where children can learn with the help of question answering and basic quizzes.*

Keywords: Conversational AI, Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Voice User Interface (VUI), Child-Centric Learning, Educational Technology, Digital Signal Processing (DSP), Text-to-Speech (TTS).

1.INTRODUCTION

As technology is evolving rapidly the way humans and computers interact have changed over years. The voice assistants that are used today help us in our day-to-day life making our work easier, they also provide us with information and services. But these voice assistants are not suitable for children as they are not specifically designed for them and also fail to understand child speech frequently [1]. Children need a voice assistant that can understand them and help them in learning by promoting interaction, curiosity and participation. This introduces a need for child-centric intelligent assistant that can facilitate educational engagement via natural speech. MILO (Mimicking and Interactive Learning with Oration) aims to tackle this issue by offering a voice-based platform that can help children in learning and education in a playful manner with security and engagement.[2], [3].

2. LITERATURE SURVEY

Research indicates that children see voice-based assistants as social entities and expect conversational interactions that are natural and aware of context. However, the current systems struggle to understand them properly and do not provide uniform and context-related interactions [1], [2].

From a speech processing point of view, Automatic Speech Recognition (ASR) systems face major difficulties in processing children's speech due to variations in the way of speaking, pitch differences and articulation [8]. Studies suggest that well designed systems are required to increase efficiency in speech recognition for child users [4], [10].

A significant difference is observed in children's results (in learning and education) due to potential of voice-based learning assistants and conversational agents. This is because of their interactive techniques such as storytelling, questioning and feedback driven learning [2], [5]. However, the majority of these systems are not specifically designed for early childhood education and lack structured educational coordination.

Another significant constraint noted in current studies is the lack of strong safety and content filtering in the systems. Research suggests that the systems designed for children should have a filtering system and so as to ensure safety and content related reactions and replies rather than unrelated and off-topic interactions [6].

In summary, the literature highlights progress in voice interaction and educational AI technologies. However, a completely integrated, child-centric, secure, and adaptable voice-driven learning assistant continues to be an unexplored research opportunity that MILO seeks to fill.

3.PROPOSED SYSTEM/METHODOLOGY

The proposed system employs a hybrid system, a dual processing framework made up of two separate operational pipelines: the Entertainment & Mimicry Pipeline and the Education & Learning Pipeline. This design allows MILO to merge engaging voice interaction with flexible learning features within an integrated system.

The complete system combines Automatic Speech Recognition (ASR), Digital Signal Processing (DSP) Natural

Language Processing (NLP), AI-driven voice generation, and gamified reward-based interaction strategies to establish an engaging and secure interaction environment [1], [2], [9].

3.1 Flow 1: Entertainment & Mimicry Pipeline

The Entertainment & Mimicry Pipeline is utilized in entertainment-centric settings like the Kitchen, Bathroom, Bedroom, and Shopping modules.

The interaction starts when the child gives voice input via microphone. The Vosk Automatic Speech Recognition (ASR) engine instantly transforms speech into text [7]. The produced text is then sent to the AI Voice Transformation module, where celebrity-like voice cloning and multilingual Text-to-Speech synthesis are utilized through ElevenLabs and Google TTS.

The audio response produced is aligned with the avatar animation using Lottie animation and animated text kit, which helps in producing an engaging and interactive environment for children promoting communication and participation.

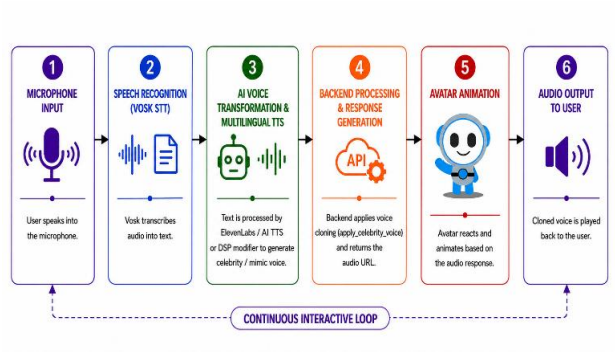

Fig-1: Flow1:- Entertainment & Mimicry Pipeline

3.2 Flow 2: Cognitive learning pipeline (Educational Pipeline)

The Cognitive Learning Pipeline is intended for engaging educational activities in the Smart Classroom (School room) setting.

The child interacts with educational prompts via voice input, the Vosk Speech Recognition engine converts this speech into text. The acknowledged text is then analyzed by the NLP evaluation system to assess if the target for learning is achieved or not, which includes recognizing numbers, words, or pronunciation trends.

According to the assessment results, the system generates responses. Suppose if the answer is correct then appreciate or if it is wrong then give corrective advice. These techniques used in AI based assistants and personalized learning helps in improvement of engagement of children in studies and improves their knowledge [5].

To boost motivation, MILO provides a gamified reward system that offers virtual currency (coins) or achievement points for correct answers. Ultimately, the produced reply is transformed into speech via the TTS engine and presented with animated avatar engagement.

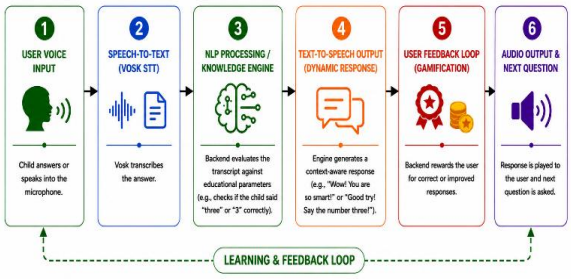

Fig-2: Flow2 :- Cognitive learning pipeline

4.IMPLEMENTATION DETAILS

The implementation of MILO (Mimicking and Interactive Learning with Oration) utilizes a modular client–server framework/architecture for immediate, child focused voice engagement. The system combines a Flutter-based frontend, a FastAPI inference backend, and dedicated speech processing and AI synthesis components to provide entertainment and educational features.

4.1 System Architecture Overview

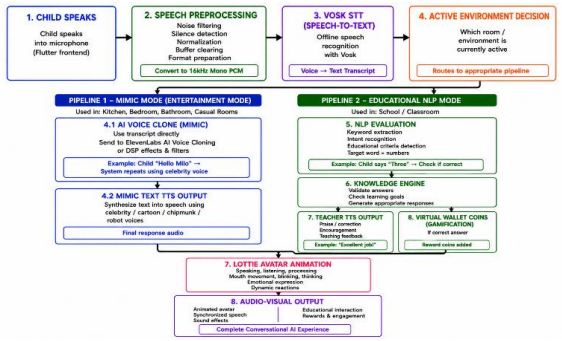

Fig-3: Hybrid Architecture Flow

i) Frontend Component (Flutter App)

The frontend is built using Flutter (Dart) to ensure deployment across various platforms. It manages voice recording with the help of microphone access and controls audio playback through the audioplayers package. The additional feature used is avatar animation with the help of Lottie animation and Animated Text Kit for attractive visuals and animation.

ii) Inference Layer for Backend (FastAPI)

The backend utilizes Python FastAPI to facilitate asynchronous handling of multipart audio requests. It serves as the main controller that manages speech processing, NLP analysis, and response creation in real time.

iii) Voice Recognition Component (Vosk ASR)

The system employs Vosk (offline ASR based on Kaldi) to transform speech into text [8]. It is enhanced for low-resource and real-time uses, making it ideal for processing child speech and for deployment at the edge [10].

iv) Audio Processing (FFmpeg DSP)

Before transcription, audio signals undergo processing through FFmpeg-based DSP pipelines that carry out noise reduction, resampling, and format normalization. This enhances the precision and reliability of speech recognition.

v) NLP and Decision-Making Engine

The transcribed text undergoes analysis through a streamlined NLP module that identifies the purpose and guides to either mode:

a} Entertainment Mode

This promotes confidence, curiosity and spontaneous communication in children.

b} Cognitive Learning (Education) Mode

This ensures that children learn in a playful way using gamified reward system.

vi) Synthesis of AI Voices (ElevenLabs TTS)

The reply is accordingly created through ElevenLabs AI voice synthesis, delivering high quality, expressive, and natural sounding speech output. The reply is subsequently transmitted to the frontend for playback and avatar alignment.

4.2 Implementation Algorithm

The system relies on two critical algorithmic flows to ensure uninterrupted operation:

4.2.1 Algorithm 1: Continuous Autonomous Acoustic Loop

```
Function StartAutonomousLoop():

  Initialize MicrophoneStream with 100ms sampling rate
  State.isRecording = True

  While State.isRecording:

    amplitude = MicrophoneStream.getCurrentDecibels()

    // Strict silence threshold logic
    If amplitude < -20.0 dB:

      If SilenceTimer is not running:
        Start SilenceTimer(duration = 2.0 seconds)

      Else:
        Cancel SilenceTimer

    // Failsafe for WebAudio anomalies
    If ExecutionTime > 8.0 seconds:
      Force Trigger StopRecording()

  On SilenceTimer Completion:

    Payload = StopRecording()
    State.isProcessing = True

    ResponseAudio = Await Backend.Process(Payload)

    PlayAudio(ResponseAudio)

  On PlayAudio Completion:

    If State.isAwake:
      Restart StartAutonomousLoop()
      // Recursive infinite loop
```

4.2.2 Algorithm 2: Fault-Tolerant Synthesis Routing

```
Function ProcessAudioRequest(Payload, VoiceID, RoomState):

  // 1. Transcoding
  PCM_Audio = FFmpeg.Convert(Payload, "16kHz Mono")

  // 2. Transcription
  Transcript = KaldiModel.Recognize(PCM_Audio)

  // 3. Semantic Regeneration
  If RoomState == "School":

    If ExtractKeywords(Transcript) matches TargetWord:

      RegeneratedText = "Correct! Good job."
      RewardTokens = 1

    Else:

      RegeneratedText = "Try again! Say the word."
      RewardTokens = 0

  Else:

    RegeneratedText = Transcript
    RewardTokens = 0

  // 4. Voice Generation
```

```
Try:

  FinalAudio = DeepLearningTTS.Synthesize(
    RegeneratedText,
    VoiceID
  )

Catch APIQuotaError or NetworkTimeout:

  // Graceful fallback to DSP
  FinalAudio = FFmpeg.ApplyFilter(
    Payload,
    effect = "Chipmunk"
  )

Return HTTPResponse(FinalAudio, RewardTokens)
```

5. RESULT & ANALYSIS

The MILO system was analyzed as a working prototype to examine its effectiveness in real-time voice communication in both Entertainment and Educational modes. The analysis concentrated on the quality of responses, system latency, performance in speech recognition, and level of user engagement.

5.1 System Performance Evaluation

Table 1- Quantitative Prototype Evaluation

Sr.No.	Parameter	Result
1.	Average Response Time	~ 2.0–3.0 seconds
2.	Supported Languages	Multiple Languages Supported
3.	Noise Tolerance	Moderate
4.	Speech Recognition Accuracy	~87% accuracy in low-noise environments
5.	Audio Sampling Rate	16 kHz Mono PCM
6.	System Processing Mode	Real-time asynchronous processing

Table 2- Qualitative Performance Observation

Sr.No.	Parameter	Result
1.	TTS Output Quality	Natural and expressive speech synthesis generated using ElevenLabs API
2.	User Engagement	Increased engagement observed due to avatar-based interaction, gamified learning, and animated feedback
3.	Mode Transition Efficiency	Smooth switching between entertainment and educational interaction modes
4.	System Reliability	Stable operation observed during continuous prototype evaluation
5.	Audio Processing Efficiency	Effective noise reduction and audio normalization achieved using FFmpeg DSP pipelines
6.	ASR Processing Capability	Efficient offline speech recognition enabled using Vosk ASR engine

5.2 System Output Screenshots

Fig-4: Home-screen/Login Page

Fig-5: School Room Interface

Fig-6: Shopping Interface

Fig-7: Kitchen Interface

Fig-8: Bathroom Interface

5.3 LIMITATIONS

However, the effectiveness relies on the quality of the audio input, as decreased accuracy is noted in noisy settings or with indistinct pronunciation, highlighting a known drawback of existing automatic speech recognition technologies for real-time use.

1]The accuracy of speech recognition reduces in noisy settings.
2]Reliance on online connectivity for AI voice generation (ElevenLabs API).
3]NLP based on rules restricts sophisticated conversational comprehension.
4]Prototype stage system lacking extensive user evaluation.
5]Latency can still be improved

6.CONCLUSION

MILO (Mimicking and Interactive Learning with Oration) offers, a voice-based platform that supports multilingual fallback, it combines speech recognition, natural language processing, digital signal processing and AI-generated voice synthesis to provide engaging learning and entertainment experiences for children. The suggested dual pipeline design balances learning and playful activities by differentiating educational engagement from playful imitation, allowing for context-related (based on keywords) and adaptive communication with minimal latency.

The system incorporates lightweight NLP-based semantic evaluation mechanism, enabling it to handle various linguistic inputs and improving accessibility for a broader audience. The combination of Flutter, FastAPI, Vosk, FFmpeg, and ElevenLabs creates a scalable and effective framework for real-time voice interaction.

Prototype assessment shows consistent speech processing, highly synthesized output, and strong user involvement, further improved with avatar-based interaction and gamified rewards. In summary, MILO shows the feasibility of creating a child-centric intelligent assistant with multilingual fallback support that integrates education and entertainment in a unified AI-based voice assistant system.

7.FUTURE SCOPE

MILO can be extended by integrating large language models (LLMs) to facilitate more natural, context-sensitive, and personalized interactive learning. This would improve the system's capacity to manage complex queries and offer personalized tutoring based on the history of user interactions.

Future enhancements might concentrate on enhancing latency and multilingual features by enabling dynamic language changes with better speech recognition and synthesis precision for various accents. This will increase the scope of users in upcoming times. It can also add elements such as storytelling and poem recitation for educational and entertainment uses.

Furthermore, adding emotion and sentiment analysis from voice input can assist the system in designing responses according to the child's emotional condition, enhancing interaction and engagement quality. The system can additionally be expanded into a cloud-based platform featuring learning analytics and curriculum oriented adaptive content delivery for organized educational assistance.

ACKNOWLEDGEMENT

We would like to extend our sincere gratitude to our Head of Department (HOD) and guide for their valuable guidance, suggestions and continuous support throughout the development of this project. These played an important role in successful completion of MILO (Mimicking and Interactive Learning with Oration).

We also extend our appreciation to the Department of Computer Engineering and our institution for providing

the necessary resources and a supportive atmosphere for carrying out this project.

REFERENCES

[1] S. B. Lovato and A. M. Piper, "Young Children and Voice Search: What We Know From Human-Computer Interaction Research," Frontiers in Psychology, vol. 10, 2019, doi: 10.3389/fpsyg.2019.00008.

[2] C. Oranç and A. Ruggeri, "Alexa, let me ask you something different: Children's adaptive information search with voice assistants," Human Behavior and Emerging Technologies, vol. 3, no. 7, 2021, doi: 10.1002/hbe2.270.

[3] Y. Tong, F. Wang, and W. Wang, "Fairies in the Box: Children's Perception and Interaction towards Voice Assistants," Human Behavior and Emerging Technologies, 2022, doi: 10.1155/2022/1273814.

[4] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition," arXiv preprint arXiv:1805.03322, 2018.

[5] S. Sajja et al., "AI-enabled intelligent assistant for personalized learning," arXiv preprint arXiv:2309.10892, 2023.

[6] M. S. Gupta et al., "SkillBot: Identifying risky content for children in voice assistants," arXiv preprint arXiv:2102.03382, 2021.

[7] A. Veysov, "Toward open-source automatic speech recognition for real-time applications using Vosk," 2020.

[8] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.

[9] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., 2023.

[10] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, 2022.