The First Step Towards Voice-Interactive Surgical LLMs

Jiahao Xu[*1] JIAHAOXU@WHU.EDU.CN
Jax Luo[*2] JLUO5@MGH.HARVARD.EDU
Nazim Haouchine[3] NHAOUCHINE@BWH.HARVARD.EDU
Scott Raymond[2] RAYMONS3@CCF.ORG
[1] *Department of Computer Science, Wuhan University, Hubei, China*
[2] *Neurological Institute, Cleveland Clinic, OH, USA*
[3] *Brigham and Women's Hospital, Harvard Medical School, MA, USA*

Abstract

Large language models (LLMs) have rapidly advanced healthcare applications such as disease diagnosis; however, their integration into surgical practice remains largely unexplored. One key barrier is the physical constraint inherent in the surgical environment—surgeons' hands are typically occupied during procedures, rendering traditional input modalities such as keyboards or touch interfaces impractical. In this study, we investigate methods for enabling LLMs to process spoken input and generate verbal responses, thereby facilitating hands-free interaction. We further developed a web-based, code-free prototype of a voice-interactive surgical LLM, accessible to any user with an internet connection. This work establishes a foundational step toward the broader goal of developing operating room–ready surgical AI systems.

Keywords: Surgical AI, Large Language Models, Voice-interactive.

1. Introduction

Large language models (LLMs) have rapidly transformed how we interact with technology by enabling human-like text generation and understanding. In healthcare, these models have demonstrated considerable potential in supporting clinical tasks such as disease diagnosis and treatment planning(Pan et al., 2025; Liu et al., 2023; Li et al., 2024; Tanno et al., 2024; abd Matthias Keicher et al., 2025). Despite these advances, their direct application in surgery remains largely unexplored. The high-stakes environment of surgery demands absolute accuracy and real-time responsiveness, which LLMs cannot yet guarantee. Additionally, surgeons' hands are almost always occupied during procedures, making it impractical to use keyboards or other traditional input devices for interacting with LLMs. This is a major practical barrier to the direct application of LLMs within the environment of an operating room.

Voice-based interaction offers hands-free communication that enables surgeons to issue commands or ask questions without interrupting their surgical tasks. This approach could allow seamless access to AI by interfacing with LLMs through spoken language. As such, developing a voice-interactive interface represents the first step toward making LLMs usable in surgical settings. In this study, we investigate methods for enabling LLMs to process spoken input and deliver verbal responses. We further developed a prototype web-based voice-interactive surgical LLM designed for experimentation and evaluation by researchers and clinicians.

[*] Contributed equally

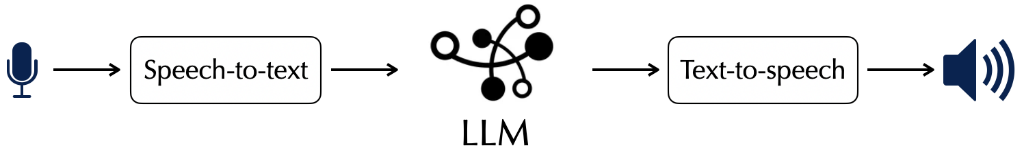


Figure 1: A concise overview of the components comprising the prototype voice-interactive surgical LLM.

2. Method

Two primary architectures have been proposed to enable voice-based interaction with LLMs: (1) the speech-to-speech (S2S) architecture and (2) the chained, or cascaded, architecture. The S2S architecture is capable of directly processing audio inputs and generating audio outputs, making it particularly well-suited for low-latency, highly interactive conversational applications. However, their application in surgical contexts is limited by the necessity of extensive domain-specific audio training data, which is currently lacking in the surgical field.

In comparison, the chained architecture represents a more practical and modular approach. It processes audio input sequentially—first transcribing speech to text, then generating responses using an LLM, and finally synthesizing speech from the generated text. Through the integration of a microphone, a speech-to-text module, and a text-to-speech module, this architecture enables the incorporation of any existing LLM into a voice-interactive system.

To develop a code-free application for surgeons to play with, we employed Gradio[1] to construct a web-based interface. The implementation details of the prototype voice-interactive surgical LLM are as follows:

1. Microphone: Integrated Gradio microphone component;

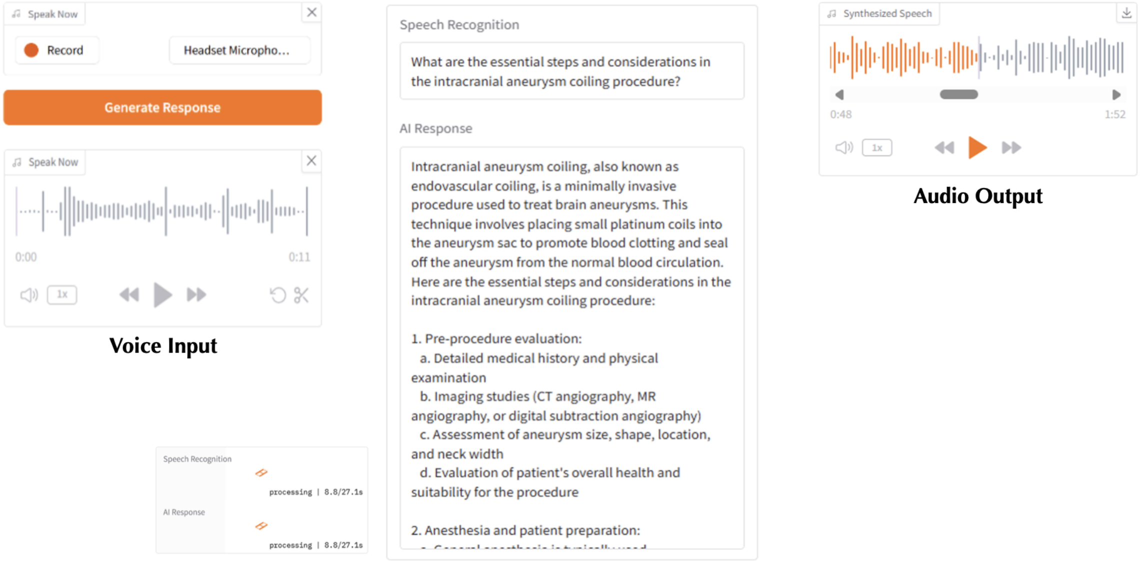
2. Speech-to-text module: Python `speech recognition` package from PyPI[2];

3. LLM: Claude 3.5 Sonnet[3], tested as the core LLM;

4. Text-to-speech module: Google *text-to-speech* for converting LLM-generated text into an `.mp3` audio file, which is rendered using Gradio's audio playback component.

Fig.1 provides a concise overview of the components comprising the prototype surgical LLM, implemented using the chained architecture.

[1] https://www.gradio.app
[2] https://pypi.org/project/SpeechRecognition
[3] https://www.anthropic.com/news/claude-3-5-sonnet



Speech Recognition

What are the essential steps and considerations in the intracranial aneurysm coiling procedure?

AI Response

Intracranial aneurysm coiling, also known as endovascular coiling, is a minimally invasive procedure used to treat brain aneurysms. This technique involves placing small platinum coils into the aneurysm sac to promote blood clotting and seal off the aneurysm from the normal blood circulation. Here are the essential steps and considerations in the intracranial aneurysm coiling procedure:

1. Pre-procedure evaluation:
   a. Detailed medical history and physical examination
   b. Imaging studies (CT angiography, MR angiography, or digital subtraction angiography)
   c. Assessment of aneurysm size, shape, location, and neck width
   d. Evaluation of patient's overall health and suitability for the procedure

2. Anesthesia and patient preparation:
   a. General anesthesia is typically used

Speech Recognition

```
processing | 8.8/27.1s
```

AI Response

```
processing | 8.8/27.1s
```

Figure 2: A demo of the proposed voice-interactive surgical LLM.

3. Demo

The prototype surgical LLM is a web-based application accessible to any user with an internet connection. As shown in Fig.2, users can initiate interaction by simply pressing the 'Record' button and speaking to the system. The application captures the audio input, transcribes it into text, and forwards it to the LLM. The model then generates a response, which is subsequently converted to speech and played back through the integrated audio output.

4. Discussion

In this work, we explored approaches for developing voice-interactive surgical LLMs and implemented a code-free, web-based prototype accessible to surgeons via any internet-connected laptop. This represents an initial step toward the long-term goal of deploying LLMs in operative settings. As a next phase, we plan to evaluate the model's speech-to-text performance using word error rate as a benchmark. Future efforts will involve testing the system in more complex and task-specific scenarios, including the accurate recognition of surgical terminology and procedural commands. Additionally, we aim to fine-tune domain-specific surgical LLMs using real-world surgical data to enhance contextual understanding and task relevance.

References

Lukas Buess abd Matthias Keicher, Nassir Navab, Andreas Maier, and Soroosh Tayebi Arasteh. From large language models to multimodal ai: A scoping review on the potential of generative ai in medicine, 2025.

Binxu Li, Tiankai Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. Mmedagent: Learning to use medical tools with multi-modal agent, 2024.

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. Radiology-llama2: Best-in-class large language model for radiology, 2023.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning, 2025.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, Karan Singhal, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Sara Mahdavi, Zahra Ahmed, Yossi Matias, Joelle Barral, SM Ali Eslami, Danielle Belgrave, Yun Liu, Sreenivasa Raju Kalidindi, Shravya Shetty, Vivek Natarajan, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, and Ira Ktena. Collaboration between clinicians and vision–language models in radiology report generation, 2024.