SpeakForm: An AI-Powered Voice Assistant for Real-Time Clinical Form Generation

1st Lakshmi Priya Armugan
Data Analytics for Business
St.Clair College
Windsor, Canada
lakshmipriya1794@gmail.com

2nd Kanchan Devi
Data Analytics for Business
St.Clair College
Windsor, Canada
kanchan1809786786@gmail.com

3rd Karthikeyan Baskaran
Data Analytics for Business
St.Clair College
Windsor, Canada
karthikeyanbaskarca@gmail.com

4th Sachin Ramasubramanian
Data Analytics for Business
St. Clair College
Windsor, Canada
sachinr2911@gmail.com

5th Akash Balu
Data Analytics for Business
St. Clair College
Windsor, Canada
akaahklk@gmail.com

***Abstract*—In modern healthcare, manual documentation during consultations reduces direct patient interaction and introduces errors. This paper presents Speak Form, an AI-powered voice assistant that transcribes patient conversations using the Google Cloud Speech-to-Text API and preprocesses the data through Gemini API. The system extracts and structures key patient details such as name, age, gender, and symptoms, which are then displayed dynamically through an intuitive user interface. This real-time form generation system enhances clinician efficiency by automating clinical documentation while ensuring high transcription accuracy and effective symptom classification. Developed with Python and modern front-end tools, Speak Form improves clinician usability and reduces documentation errors. Future work includes multilingual support, integration with Electronic Health Records (EHR), and expanding the system's capabilities for broader healthcare applications.**

***Index Terms*—Speak Form, Voice Assistant, Clinical Documentation, Speech-to-Text, Google Cloud Speech-to-Text, Gemini API**

I. Introduction

In modern healthcare settings, physicians often conduct consultations while simultaneously recording patient information into structured forms or Electronic Health Record (EHR) systems. This manual data entry process is time consuming, prone to transcription errors, and results in workflow inefficiencies. Consequently, healthcare professionals face increased administrative burdens, reduced time for direct patient interaction, and a higher risk of documentation inaccuracies.

To address these challenges, there is a growing interest in applying artificial intelligence (AI) technologies especially speech recognition and natural language processing (NLP) to automate documentation and enhance clinical workflows. Voice assistants powered by AI offer the potential to transform the patient-provider interaction by enabling real-time

transcription, intelligent information extraction, and automatic population of clinical systems.

This paper presents Speak Form, an AI-driven voice assistant system specifically designed to transcribe and process real- time patient conversations. The system utilizes Google Cloud Speech-to-Text API for accurate transcription of voice records. The transcribed text is then processed using the Gemini API, which extracts essential patient details such as name, age, gender, and symptoms. This cleaned and structured data is displayed dynamically through a user-friendly interface, presenting the information in the form of a digital clinical form, which facilitates seamless review by healthcare professionals.

The primary objectives of this work are to:

- Develop a speech recognition pipeline using Google Cloud Speech-to-Text API to transcribe consultation audio into text.
- Use the Gemini API to extract and preprocess relevant clinical entities from the transcribed text.
- Render the cleaned data in a structured format on a front-end interface for healthcare professionals.

The current scope is limited to English-language consultations in a simulated environment, with an emphasis on basic patient intake and symptom collection. Future work may explore integration with real-time EHR systems, support for multilingual interactions, and deployment in real-world clinical settings.

II. Literature Review

Voice assistant technologies have gained significant attention for their potential to reduce the administrative burden on healthcare providers. Research by Baker et al. [3] highlights how Automated Speech Recognition (ASR) systems can streamline documentation and improve the efficiency of clinical workflows. These tools enable healthcare practitioners to transcribe spoken interactions more quickly and accurately than manual entry, which can reduce the likelihood of transcription errors. The integration of speech-to-text engines into

Electronic Health Records (EHRs) has been the subject of considerable research. Bharti et al. [5] emphasizes the benefits of voice interfaces in mitigating clinician burnout and enhancing the overall usability of EHR systems. In addition, Bartle et al.
[4] conducted evaluations of various ASR models, concluding that domain-specific and context-aware models offer superior performance in clinical scenarios, particularly when handling medical jargon and symptom-specific terminology. Among the leading tools in this space are Google Cloud Speech-to-Text API and other advanced speech recognition technologies. These systems have demonstrated high accuracy in various acoustic environments and are increasingly being utilized for healthcare applications such as medical transcription. The Google Cloud Speech-to-Text API has shown impressive results in real-time transcription with minimal latency, making it a suitable backbone for conversational AI in healthcare settings [6], [7]. However, many existing solutions focus primarily on improving transcription accuracy, while neglecting to address the complete workflow from transcription to data preprocessing and the final structured presentation of the information. Often, the critical steps of organizing the transcribed data and extracting relevant patient information (e.g., name, age, symptoms) into usable formats for clinicians are overlooked [6]. This project addresses this gap by proposing an integrated pipeline that incorporates:

- **G**oogle Cloud Speech-to-Text API for transcription of clinical conversations.
- **G**emini API for intelligent data preprocessing and entity extraction (e.g., name, age, gender, symptoms).
- A front-end user interface to display the cleaned data in a structured and easily accessible format for healthcare professionals.

By combining transcription, NLP-driven data extraction, and seamless UI integration into a unified system, this work advances the field beyond transcription to facilitate real-world implementation in healthcare environments.

III. Methodology

A. Data Collection

To develop and evaluate the voice assistant tool, a dataset comprising simulated healthcare conversations, appointment requests, and symptom descriptions were utilized. The dataset was sourced from Hugging Face website. Real patient data was not incorporated due to stringent privacy and ethical constraints. The use of synthetic data ensures compliance with healthcare data regulations while providing a controlled environment for system development and testing.

B. Data Preprocessing by Prompt-Based NLP Using Gemini API

Preprocessing the collected data was essential to enhance model performance and ensure accurate interpretation. The preprocessing steps included:

- **Text Cleaning:** Removal of filler words, extraneous punctuation, and irrelevant linguistic artifacts.
- **Speech-to-Text Conversion:** Real-time audio data was transcribed using Google's Speech Recognition API.

Once transcribed, the text is processed through the Gemini API using prompt engineering techniques to extract relevant information such as patient name, age, symptoms, diagnosis, and treatment recommendations. The model uses a Large Language Model structure to ensure consistent extraction, even across varied speech patterns.

C. Data Storage and CSV Structuring

The structured data from the Gemini API is cleaned and saved in a comma-separated values (.csv) format. This file acts as the master dataset for integration.

D. Form Population and PDF Generation

Using the CSV file, the system automatically populates a digital patient form template. The completed form is saved into a secure backend database for long-term record keeping. A corresponding PDF document is generated for each patient entry, providing a printable and shareable report for healthcare providers.

E. Tools and Technological Stack

The system was implemented using Python as the primary programming language. Key tools and libraries included Google Speech Recognition for converting spoken input into textual format. For natural language processing and extraction tasks, the Gemini API was employed. Development and testing were conducted using Jupyter Notebook.

F. System Integration and Impact

This end-to-end process eliminates manual transcription, reduces error rates, and enhances data uniformity. It also supports clinicians by providing them with structured, searchable, and up-to-date patient records in real time.

G. Abbreviations and Acronyms

The following abbreviations and acronyms are used throughout this paper:

- **AI** – Artificial Intelligence
- **API** – Application Programming Interface
- **CSV** – Comma-Separated Values
- **EHR** – Electronic Health Record
- **NLP** – Natural Language Processing
- **LLM** – Large Language Model

All abbreviations are defined at their first mention in the text and are used consistently throughout the manuscript.

IV. System Architecture

The system architecture of the proposed voice-enabled medical assistant is illustrated in Figure 1. The overall workflow is divided into three primary components: transcription, data preprocessing, and front-end UI generation.

- **Transcription Module:** This module utilizes the Google Cloud Speech-to-Text API to transcribe audio recordings of doctor-patient consultations into text.

- **Preprocessing and Data Extraction:** The transcribed text is processed by the Gemini API, which extracts key information, such as patient name, age, gender, and symptoms.
- **UI Generation:** The cleaned and structured data is rendered through a dynamic user interface, presenting the extracted patient information in a standardized clinical form.

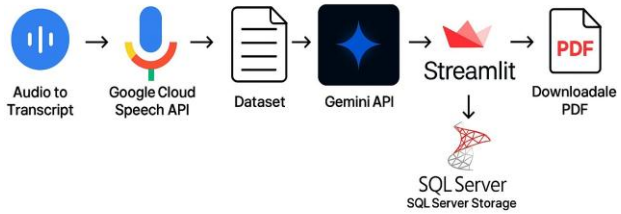

Fig. 1. System Architecture of the Voice-Enabled Medical Assistant.

FUTURE SCOPES

Based on the current findings and the potential of voice-enabled medical assistants, the following future directions are recommended to enhance system capabilities and practical adoption:

- **Deployment in Real-world Healthcare Environments:** Future iterations should be tested and deployed across diverse clinical settings such as hospitals, outpatient clinics, and elder care centers. Real-world deployment will help assess practical usability, system robustness, and its effectiveness in supporting medical staff during patient interactions. This will also provide actionable feedback to further refine system performance and user experience.
- **Multilingual Support:** To improve accessibility and inclusivity, especially in linguistically diverse regions, future versions should incorporate multilingual capabilities. Support for various languages and dialects will ensure that the system can cater to a broader patient demographic, including non-native speakers, thereby expanding its usability in global healthcare environments.
- **Integration with EHR Systems:** Seamless integration with existing EHR platforms is critical for real-time population of patient data. This would streamline documentation workflows, reduce manual entry errors, and enhance data consistency, ultimately improving operational efficiency for healthcare providers.
- **Support for Additional Data Modalities:** Future versions should incorporate the ability to handle additional data types such as medical images and laboratory results, and sensor data. A multimodal approach would provide a more holistic patient profile, enabling more accurate diagnosis and personalized treatment recommendations.
- **Policy and Regulatory Compliance:** Strict adherence to healthcare regulations such as the Health Insurance Portability and Accountability Act (HIPAA) is essential. Future implementations must incorporate secure data management practices, including user access control, encryption, and compliance auditing to maintain patient confidentiality and trust.
- **Continuous System Improvement and User Feedback:** Ongoing enhancement of the system should be driven by direct feedback from healthcare professionals and advancements in speech recognition and natural language processing technologies. This iterative development approach will ensure that the system remains aligned with evolving clinical needs and technological standards.

CONCLUSION

This paper presented the design and implementation of a voice-enabled medical assistant that leverages automatic speech recognition (ASR) and natural language processing (NLP) techniques to facilitate seamless interaction between users and healthcare systems. The system enables hands-free access to medical information, appointment scheduling, and symptom inquiry, thereby enhancing the efficiency and accessibility of healthcare services. Experimental evaluations demonstrated the effectiveness of the assistant in understanding domain- specific queries and responding with contextually relevant information. Future work will focus on expanding the medical knowledge base, integrating multilingual support, and improving the system's robustness in noisy clinical environments. The proposed solution has the potential to serve as a valuable tool in modern healthcare infrastructure, particularly in remote patient monitoring.

ACKNOWLEDGEMENT

The authors would like to thank the faculty and staff of the Department of Computer Science at [St. Clair College] for their valuable guidance and support throughout this research. Special thanks to [Sutharsan Sivagnanam], whose expertise and encouragement were instrumental in completing this work. The authors also acknowledge the use of open-source tools, which significantly contributed to the development and testing of the voice-enabled medical assistant system.

REFERENCES

[1] M. Al Shamsi, A. Alkass, and H. Alnaqbi, “Understanding key drivers affecting students’ use of artificial intelligence-based voice assistants,” 2022.

[2] R. Alagha and D. Helbing, “Evaluating the quality of voice assistants’ responses to consumer health questions about vaccines,” 2019.

[3] A. Baker, D. Patel, J. Sun, and K. Hughes, “AI vs human doctors for triage and diagnosis,” 2020.

[4] N. Bartle, J. Z. Wang, and J. E. Morris, “A second voice: Opportunities for voice assistants to support home health aides,” 2022.

[5] A. Bharti, S. R. Malhotra, and R. Sinha, “Medbot: Conversational AI for telehealth after COVID-19,” 2020.

[6] T. Bickmore, R. Trinh, T. Asadi, and M. Paasche-Orlow, “Patient and consumer safety risks when using conversational assistants for medical information,” *NPJ Digit. Med.*, vol. 1, no. 13, 2018.

[7] R. Brewer, L. McIntosh, and S. McLaughlin, “Older adults’ use of voice assistants for health information,” *Proc. ACM Hum. -Compute. Interact.*, vol. 6, CSCW2, Article 128, 2022.

[8] M. Buinhas, P. Lima, F. Silva, and R. P. Rocha, “Virtual assistant for type 2 diabetes self-care,” in *Proc. 13th Iberian Conf. Inf. Syst. Technol. (CISTI)*, 2018.