AUTOMATIC GENERATION OF CLINICAL ENCOUNTER SUMMARIES FROM RECORDINGS USING LARGE LANGUAGE MODELS

Vinoth Madhavan
vinoth.madhavan11@gmail.com
University of Texas at Austin
Austin, Texas, USA

ABSTRACT

Clinical documentation is a cornerstone of modern healthcare, yet it imposes a significant administrative burden on clinicians and introduces risks to patient safety, particularly during handoffs between providers. In this thesis, we investigate the feasibility and challenges of an automated solution: generating structured clinical encounter summaries directly from audio or video recordings using a hybrid pipeline that integrates state-of-the-art speech-to-text technology and large language models (LLMs). We present the design, implementation, and evaluation of a prototype system capable of extracting and organizing essential clinical information—such as diagnoses, medications, follow-up plans, red flags, and patient concerns—into a standardized Word document for clinician review. Our findings highlight both the promise and the current limitations of this approach. While the system shows notable improvements in efficiency and standardization, critical challenges such as transcription errors, hallucinations in LLM outputs, and domain-specific complexities remain. The study concludes by identifying pathways for future research and clinical integration, with the ultimate goal of enhancing patient safety and reducing clinician burnout.

1. INTRODUCTION

Clinical documentation is an essential but time-consuming aspect of healthcare delivery. Physicians and other healthcare providers spend a substantial portion of their workday on charting, often at the expense of direct patient care and professional satisfaction [1]. The quality and completeness of clinical documentation are especially critical during patient handoffs, where lapses or ambiguities can lead to miscommunication, medical errors, and adverse patient outcomes [2]. As healthcare systems become increasingly complex, the demand for efficient, accurate, and standardized documentation methods grows correspondingly.

Traditional documentation methods, such as manual note-taking or dictation followed by human transcription, are labor-intensive and prone to variability in both content and structure. While digital voice assistants and basic speech-to-text tools have been introduced, they often lack the domain-specific understanding required to accurately capture the nuances of clinical encounters. These limitations underscore the need for more advanced, automated solutions that can reliably generate high-quality clinical summaries. Recent advances in artificial intelligence (AI), particularly in automatic speech recognition (ASR) and large language models (LLMs), offer an opportunity to rethink how clinical documentation is performed.

PROBLEM STATEMENT:

Can recent advances in speech-to-text models and large language models (LLMs) be leveraged to automatically generate accurate, structured clinical summaries from encounter recordings, thereby improving the quality of patient handoffs and reducing the documentation burden on clinicians?

This thesis explores the technical feasibility, potential benefits, and inherent challenges of such an approach. We describe the development and evaluation of an end-to-end prototype system, constructed using open-source tools, and assess its performance on both simulated and real-world clinical data.

2. RELATED WORK

2.1 Automated Clinical Note Generation

The automation of clinical note generation has been an active area of research, particularly with the advent of deep learning and natural language processing (NLP) techniques. Shickel et al. [3] surveyed the application of deep learning to electronic health record (EHR) analysis, highlighting early successes in topic extraction and structured data summarization. However, these approaches often struggled with unstructured free-text data and were not designed to process audio or

conversational input, limiting their applicability to real-world clinical encounters.

2.2 Speech-to-Text in Healthcare

Automatic speech recognition (ASR) systems have been evaluated for their potential to transcribe clinical conversations. Ko et al. [4] assessed the performance of OpenAI's Whisper and other ASR models in clinical settings. While Whisper demonstrated high accuracy with clear, well-articulated speech, its performance declined in the presence of medical jargon, diverse accents, and conversational interruptions—common features of real-world clinical encounters. These findings suggest that while ASR technology has matured, domain-specific challenges remain.

2.3 Large Language Models for Medical Summarization

Recent advances in LLMs, such as GPT-3 and its successors, have enabled more sophisticated summarization of medical texts and conversations. Zhang et al. [5] demonstrated that LLMs can effectively capture salient concepts from patient-doctor interactions, improving the quality of generated summaries. However, these models are prone to "hallucinations"—the generation of plausible but factually incorrect information—and may overgeneralize or misattribute details. The integration of LLMs with multi-modal data (e.g., audio, video) and their application to structured clinical documentation, particularly for handoff support, remains an underexplored area.

In summary, while prior research provides a foundation for automated clinical documentation, significant gaps remain in the integration of ASR and LLMs for end-to-end, structured note generation from real-world recordings.

3. METHODOLOGY

3.1 System Overview

The proposed system is designed as a modular pipeline, comprising three primary stages:

Transcription:

Audio or video recordings of clinical encounters are first processed to extract the spoken content. For video files (e.g., mp4), audio is separated using the moviepy library. The extracted audio is then preprocessed (e.g., noise reduction, normalization) using librosa to enhance transcription quality. The preprocessed audio is transcribed into text using OpenAI's Whisper, a state-of-the-art ASR model known for its robustness across languages and accents.

Topic Extraction:

The resulting transcript is analyzed by a large language model (LLM), specifically Google Gemini, which is prompted to identify and extract key clinical sections. These include diagnoses, medications, follow-up plans, red flags (urgent warning signs), and patient concerns. Custom prompts are crafted to guide the LLM in segmenting and labeling the information accurately.

Document Generation:

The extracted and organized content is formatted into a standardized Word document using the python-docx library. This document is structured with clear section headings, facilitating rapid review and editing by clinicians. The output is designed to be compatible with existing EHR systems and clinical workflows.

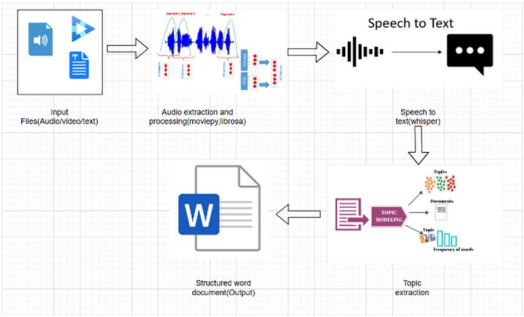

Figure 1. Workflow Diagram

3.2 Implementation Details

Transcription:

The system uses moviepy to extract audio from video files, ensuring compatibility with a range of input formats. librosa is employed for audio preprocessing, which includes noise filtering and normalization to improve ASR performance. The openai-whisper package is then used to transcribe the audio, leveraging its advanced neural architecture for high accuracy.

LLM Postprocessing:

The Gemini LLM is accessed via API, with custom prompts designed to elicit structured, sectioned

summaries from the raw transcript. The prompts are iteratively refined to minimize errors such as misattribution or omission of critical details.

Formatting:

The python-docx library is used to generate a Word document with labeled sections. The output is designed for readability and ease of integration into clinical documentation systems.

3.3 Dataset

To evaluate the system, we assembled a diverse dataset comprising:
Simulated Patient-Doctor Conversations: Publicly available datasets such as MTSamples, which provide realistic but anonymized clinical dialogues.
Artificially Generated Recordings: Mock interviews conducted by project team members, designed to simulate a range of clinical scenarios and speaking styles.
Real-World De-identified Clinical Dictations: Where ethically permissible, actual clinical recordings were used, with all patient identifiers removed to ensure privacy.
Each generated summary was reviewed by a licensed clinician, who assessed the accuracy and completeness of the extracted information. This manual review served as the gold standard for benchmarking system performance.

4. RESULTS

4.1 Summary Quality and Error Types

Transcription Reliability:

Whisper achieved a word accuracy rate of 85%–90% for recordings with clear speech and minimal background noise. However, its performance declined in the presence of heavy accents, overlapping speech, or specialized medical terminology, resulting in an error rate of up to 15%. These errors included misrecognition of drug names, omission of critical phrases, and confusion between speakers.

Section Extraction:

The Gemini LLM was generally effective at segmenting the transcript into the required sections. In most cases, it correctly identified diagnoses, medications, follow-up plans, red flags, and patient concerns. However, occasional misattributions were observed, such as including a follow-up recommendation under red flags or vice versa. These errors were more common in transcripts with ambiguous or fragmented speech.

Hallucinations:

A notable limitation of LLMs is their tendency to "hallucinate"—to generate plausible but unsubstantiated information. In a minority of cases, the LLM inferred diagnoses or recommendations that were not explicitly mentioned in the transcript, potentially introducing clinical risk if not detected during review.

Formatting:

The generated Word documents were consistently well-structured, with clear section headings and readable formatting. Clinicians reported that the documents were suitable for further editing and integration into patient records.

4.2 Representative Output

A typical output from the system might include:
Diagnosis: Hypertension, Type 2 Diabetes
Medications: Lisinopril, Metformin
Follow-up Plans: Routine bloodwork in 3 months
Red Flags: Chest pain, sudden shortness of breath
Patient Concerns: Difficulty affording medications
This structured format facilitates rapid review and ensures that critical information is not overlooked during handoffs.

4.3 Impact and Limitations

Time Savings:

The system was able to generate a draft clinical note in under two minutes for a five-minute recording, representing a substantial reduction in documentation time compared to manual methods.

Clinical Review Burden:

Despite the efficiency gains, manual review by a clinician remains essential to ensure accuracy and prevent the propagation of errors, particularly those arising from transcription mistakes or LLM hallucinations.

Unmet Challenges:

Several technical and ethical challenges remain unresolved. These include handling overlapping speakers, detecting sarcasm or humor (which may affect clinical interpretation), and ensuring robust data privacy and security throughout the pipeline.

5. CONCLUSION AND FUTURE DIRECTIONS

This thesis demonstrates the technical feasibility of using a hybrid pipeline—combining advanced speech-to-text models and large language models—for the automatic generation of structured clinical documentation from encounter recordings. The prototype system shows promise in reducing clinician workload, standardizing documentation, and supporting safer patient handoffs.

However, significant limitations persist. Transcription errors, LLM hallucinations, and the need for manual review preclude unsupervised deployment in clinical settings. Addressing these challenges will require further research in several areas:

Domain-Specific Model Training:

Fine-tuning ASR and LLM models on large, diverse datasets of clinical conversations to improve accuracy with medical terminology and varied speech patterns.

Integration with Medical Ontologies:

Leveraging structured medical vocabularies (e.g., SNOMED CT, RxNorm) to constrain and validate LLM outputs, reducing the risk of hallucinations and misattributions.

Speaker Diarization and Context Awareness:

Enhancing the system's ability to distinguish between multiple speakers and to interpret conversational context, including sarcasm, humor, and implicit cues.

Explainability and Transparency:

Developing methods to make LLM outputs more interpretable and auditable, enabling clinicians to understand and trust the system's recommendations.

Privacy and Security:

Implementing robust safeguards to protect patient data throughout the pipeline, in compliance with legal and ethical standards.

Future work will focus on aligning LLM-generated summaries with structured EHR templates, improving model transparency, and conducting rigorous error quantification. Additional studies are needed to assess clinician acceptance, workflow integration, and the impact on patient outcomes.

REFERENCES

[1] Sinsky, C., et al. (2016). Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*.

[2] Starmer, A., et al. (2014). Changes in medical errors after implementation of a handoff program. *New England Journal of Medicine*.

[3] Shickel, B., et al. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*.

[4] Ko, M., et al. (2023). Evaluating speech-to-text models for medical documentation: OpenAI Whisper in clinical settings. *AMIA Annual Symposium Proceedings*.

[5] Zhang, Y., et al. (2023). Large Language Models for Clinical Conversations: Results and Risks. *arXiv preprint arXiv:2306.14780*.