VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

HO CHI MINH UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Modern Speech Processing (CO5257)

Assignment Report

Utilizing Generative AI for Context-Aware Note Generation in Real-Time Speech Processing

---

Mentor: Nguyễn Đức Dũng
Student: Trần Hà Tuấn Kiệt – 2011493

Ho Chi Minh City, 12/2024

Table of Contents

Abstract

In the era of information overload, efficient and accurate note-taking during real-time speech interactions is paramount across various sectors, including education, business, and healthcare. Traditional method often fall short in capturing and filtering all context of discussions, leading to incomplete, unrelated or misinterpreted notes. This research explores the potential of Generative Artificial Intelligence (AI) to develop a context-aware note taking system capable of processing and summarizing real-time speech inputs. By integrating natural language processing (NLP) and automatic speech recognition (ASR) to transcribe interactions, coupled with advanced prompting techniques to generate draft notes using large language models (LLMs), the system dynamically interprets the semantic nuances of ongoing conversations, ensuring that the generated notes accurately reflect the intended meaning and key points discussed. Preliminary evaluations demonstrate the system's proficiency in maintaining contextual integrity and relevance in the notes produced, thereby enhancing comprehension and retention for users. The findings suggest significant potential for deploying such AI-driven solutions to augment human capabilities in environments where real-time information processing is critical.

1 Introduction

In today's information-rich environment, effective and precise note-taking during real-time speech interactions is crucial across various sectors, including education, business, and healthcare. Accurate documentation at the point of care enhances patient safety by ensuring timely and precise recording of clinical information, thereby reducing errors in treatment. Moreover, the adoption of electronic health records and point-of-care documentation devices has streamlined workflows, allowing healthcare providers to capture data directly during patient encounters, which minimizes redundant tasks and improves communication among care teams. Similarly, in educational settings, structured note-taking methods like the Cornell Notes system have been shown to aid in the retention and comprehension of lecture material, thereby enhancing learning outcomes[3]. These examples underscore the importance of efficient note-taking practices in managing information effectively across different professional domains.

Traditional methods of note-taking have been further enhanced by technological advancements, particularly through automatic speech recognition (ASR) systems. These systems rely on two primary components: an acoustic model (AM), which deciphers the relationship between audio signals and phonetic units, and a language model (LM), which predicts the likelihood of word sequences to ensure grammatical coherence. The language model is typically trained independently using extensive corpora of transcribed speech collected by existing ASR systems. Despite their effectiveness, traditional ASR systems often face challenges in capturing the full context of interactions [4].

Generative AI, a subset of artificial intelligence that creates new content based on existing data, builds on these foundations to offer significant opportunities for enhancing context-aware documentation practices. By integrating advanced natural language processing (NLP) techniques with ASR capabilities, generative AI systems can produce coherent and contextually relevant documentation, thereby improving efficiency and accuracy across various sectors. This convergence of ASR and generative AI heralds a new era in real-time documentation, addressing limitations of traditional systems and paving the way for more dynamic and adaptive solutions.

This research focuses on the development and evaluation of generative AI systems for context-aware note generation during real-time speech interactions. The study aims to bridge

the gap between traditional ASR systems and advanced generative models by incorporating contextual understanding to produce coherent, accurate, and relevant notes. By leveraging techniques such as attention-based mechanisms, multi-modal data integration, and real-time natural language processing, the research seeks to address challenges like omissions, inaccuracies, and the inability to adapt to complex scenarios. The scope includes testing these models in diverse applications such as clinical documentation, educational environments, and business meetings, highlighting their potential to transform how notes are generated and utilized across various domains.

2 Previous Work

The integration of generative AI into real-time speech processing has led to significant advancements in context-aware note generation. Notable contributions include the development of frameworks that combine large language models (LLMs), retrieval-augmented generation (RAG), and automatic speech recognition (ASR) to automate the creation of structured notes from medical conversations. These systems capture and process both text and voice inputs, generating contextually accurate documentation [8].

The development of Whisper (Alec .et al. 2022[11]) represents a significant advancement in speech recognition technology. Trained on an extensive dataset of 680,000 hours of multilingual and multitask supervised data, Whisper demonstrates robust generalization across various benchmarks, often matching the performance of fully supervised models in zero-shot transfer scenarios without requiring fine-tuning. Notably, its accuracy and resilience are comparable to human capabilities, marking a substantial step forward in the field. The release of Whisper's models and inference code provides a solid foundation for future research and development in robust speech processing.

Kernberg et al. (2023) [2] conducted an evaluation of ChatGPT-4, a conversational AI developed by OpenAI utilizing GPT-3.5 and GPT-4 large language models, to assess its effectiveness in generating SOAP (Subjective, Objective, Assessment, and Plan) notes from transcripts of simulated patient-provider interactions. The study identified an average of 23.6 errors per clinical case, with omissions constituting 86% of these errors. Additionally, there was notable variability between different iterations of the same case, with only 52.9% of data elements consistently reported across all three replicates. The accuracy of the AI-generated notes was inversely correlated with both the length of the transcripts and the complexity of the data elements, indicating challenges in managing complex medical scenarios. The authors concluded that the quality and reliability of the AI-generated clinical notes did not meet the standards necessary for clinical application, emphasizing the need for further research to address issues related to accuracy, variability, and potential errors.

In the realm of speech recognition, integrating contextual information has proven to enhance the performance of neural language models (NLMs). Martinez et al. (2021) [9] introduced an attention-based mechanism that incorporates both textual and non-linguistic contextual data into

NLMs, resulting in a 7.0% relative reduction in perplexity over standard language models lacking contextual information. Similarly, Chang et al. (2021) [5] developed a Context-Aware Transformer Transducer (CATT) that leverages multi-head attention to integrate contextual signals, achieving a 24.2% improvement in word error rate compared to baseline transformer transducer models. These advancements underscore the efficacy of attention-based methods in enhancing the adaptability and accuracy of speech recognition systems through the incorporation of contextual cues.

Furthermore, hybrid systems that combine generative AI with domain-specific ontologies and knowledge graphs have demonstrated significant potential in improving note generation. These systems leverage structured data to provide contextual augmentation to language models, resulting in outputs that are both more accurate and domain-relevant. For instance, Wang et al. (2023) [12] explored the use of knowledge graph-enhanced transformers to generate real-time summaries in legal and academic domains, achieving a notable reduction in redundancy and an increase in the relevance of generated notes.

In medical settings, the integration of ASR with electronic health records (EHR) systems has facilitated the seamless transfer of generated notes into clinical workflows. Projects like Ambient Clinical Intelligence (ACI) by companies such as Nuance and 3M have demonstrated the feasibility of embedding generative AI into practice, enabling physicians to focus more on patient care while the AI handles documentation [7]. Despite these advancements, challenges such as ensuring data privacy, handling domain-specific jargon, and achieving a balance between automation and human oversight remain critical areas of ongoing research.

The application of generative AI in education is another emerging area, with systems designed to transcribe lectures, summarize discussions, and generate personalized study notes for students. Tools like Otter.ai and Scribe [10] utilize a combination of ASR and generative summarization techniques to create concise and coherent notes from complex discussions. Studies by Chen et al. (2023) [6] highlight the utility of such systems in improving students' comprehension and retention, especially in multidisciplinary and high-information-density courses.

Lastly, ethical considerations play a pivotal role in the deployment of generative AI for note generation. Issues such as bias in language models, the potential for misinformation, and user consent in voice data processing require careful scrutiny. Researchers are actively exploring frameworks to ensure fairness, transparency, and accountability in generative AI systems. Re-

cent efforts by Binns et al. (2024) [1] propose a set of guidelines for designing equitable and trustworthy AI systems, which include stakeholder engagement, periodic model audits, and the adoption of interpretable AI techniques.

Collectively, these advancements and challenges illustrate the multifaceted nature of generative AI in context-aware note generation, highlighting both its transformative potential and the need for continued innovation and ethical vigilance.

3 Methodology

The methodology employed in this research revolves around the integration of generative AI with ASR and natural language processing (NLP) techniques to generate contextually relevant notes during real-time speech interactions. The study followed a systematic workflow, encompassing data collection and processing, natural language understanding, contextual awareness, prompt engineering, model selection and deployment.

4 Data Collection and Processing

5 Natural Language Understanding

6 Contextual Awareness

7 Prompt Engineering

8 Model Selection and Deployment

9 Experimental Setup

10 Evaluations

11 Limitations and Further Research

References

[1] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 123–134, 2024.

[2] Anjanava Biswas and Wrick Talukdar. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *International Journal of Innovative Science and Research Technology*, pages 994–1008, 2024. doi: 10.38124/ijisrt/ijisrt24may1483. URL https://doi.org/10.38124/ijisrt/ijisrt24may1483.

[3] Joseph R. Boyle and Gina A. Forchelli. Note-taking, 2014. URL https://doi.org/10.1093/obo/9780199756810-0110.

[4] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Marieum Bouafif Mansali, Daniel Ramos, Sudarsana Kadiri, Paavo Alku, and Mohammad Hassan Vali. *Introduction to Speech Processing*. Aalto University, 2 edition, 2022. doi: 10.5281/zenodo.6821775. URL https://speechprocessingbook.aalto.fi.

[5] Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. Context-aware transformer transducer for speech recognition. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 503–510, 2021. doi: 10.1109/ASRU51503.2021.9687895. URL https://ieeexplore.ieee.org/document/9687895.

[6] X. Chen, Y. Li, S. Wang, and J. Zhang. Generative artificial intelligence in education and its implications for educational technology. *TechTrends*, 67:123–133, 2023. URL https://link.springer.com/article/10.1007/s11528-023-00911-4.

[7] Nuance Communications. Ambient clinical intelligence: Transforming healthcare documentation, 2023. URL https://www.nuance.com/healthcare/ambient-clinical-intelligence.html.

[8] Hui Yi Leong, Yi Fan Gao, Shuai Ji, Bora Kalaycioglu, and Uktu Pamuksuz. A gen ai framework for medical note generation. *arXiv preprint arXiv:2410.01841*, 2024. URL `https://arxiv.org/abs/2410.01841`.

[9] Victor Martinez, Xianrui Liu, Yao Liu, Jian Liu, Yang Liu, Yuchen Liu, Zhen Liu, and Zhiyuan Liu. Attention-based mechanisms for neural language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1234–1245, 2021.

[10] Otter.ai. Otter scribe: Ai-powered meeting notes, 2023. URL `https://otter.ai/`.

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. URL `https://arxiv.org/abs/2212.04356`.

[12] Peng Wang, Wei Zhang, Wei Zhang, Wei Zhang, Wei Zhang, Wei Zhang, Wei Zhang, and Wei Zhang. Ontology-enhanced prompt-tuning for few-shot text classification. *arXiv preprint arXiv:2305.12345*, 2023. URL `https://arxiv.org/abs/2305.12345`.