VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

HO CHI MINH UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Speech Processing

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[2]. 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 [3].

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 [5].

The development of Whisper (Alec .et al. 2022[7]) 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) [1] 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) [6] 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) [4] 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.

These developments underscore the potential of generative AI to enhance the efficiency and accuracy of note-taking across various professional domains.

3 Approach

3.1 Data Processing

3.2 Model

3.3 Multitask Format

4 Contextual Background

5 Challenges and Further Research

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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. doi: 10.48550/arXiv.2410.01841. URL https://arxiv.org/abs/2410.01841.

[6] Richard Diehl Martinez, Scott Novotney, Ivan Bulyko, Ariya Rastrow, Andreas Stolcke, and Ankur Gandhe. Attention-based contextual language model adaptation for speech recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1994–2003, 2021. doi: 10.18653/v1/2021.findings-acl.175. URL https://aclanthology.org/2021.findings-acl.175.

[7] 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.