Context-Aware AI Stenography: Enhancing Media Accessibility for the Deaf and Hard of Hearing

CS4991 Capstone Report, 2025

Tony Chang
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia, USA
jmb5jh@virginia.edu

ABSTRACT

The deaf and hard-of-hearing community faces significant challenges in accessing audio-based media, as current closed captioning systems and human stenography often fail to provide accurate, timely, and contextually coherent captions. I propose a context-aware AI stenography model leveraging natural language processing (NLP) and machine learning (ML) to generate highly accurate and contextually appropriate captions. The model will be designed to analyze grammatical structures and contextual cues, enabling it to produce readable text that preserves the intended meaning of speech. It will be trained on diverse datasets, including variations in accents, speech patterns, and multi-speaker environments, with real-time functionality ensuring minimal latency for live events. The expected outcome is a robust, inclusive captioning system that surpasses traditional methods by offering more accurate, coherent, and timely captions, significantly enhancing media accessibility for the deaf and hard-of-hearing community. Future work will involve extensive testing in real-world scenarios to evaluate performance, identify potential bugs or inaccuracies, and optimize the model further. Additionally, future phases could expand the system's capabilities to include support for more languages and integration with emerging media technologies.

1. INTRODUCTION

Access to audio-based media remains a significant challenge for deaf and hard-of-hearing individuals, despite advancements in closed captioning and speech recognition technology. Traditional closed captions, often generated by simple speech-to-text models, struggle with accuracy, frequently misinterpreting words due to a lack of contextual awareness. While human stenographers offer a more precise alternative, their real-time transcription is limited by typing speed and the potential for human error.

These shortcomings create barriers to understanding complex content, especially in fast-paced or multi-speaker environments like lectures, meetings, and live broadcasts. To address these challenges, this project proposes an advanced AI-driven captioning system that leverages natural language processing and machine learning to produce more accurate, coherent and context-aware transcriptions.

2. RELATED WORKS

AI-driven text generation and steganography have evolved from rule-based and statistical models to advanced NLP techniques leveraging RNNs and LSTMs for more realistic and secure message embedding (Gurunath, et al 2021). However, traditional

methods often lacked automation, accuracy, and resilience against security threats.

Recent advancements in LLMs have introduced new challenges, as their black-box nature restricts access to training parameters essential for traditional steganographic mapping (Wu, et. al., 2024). To address this, LLM-Stega utilizes encrypted mapping and reject sampling optimization to embed secret messages via LLM user interfaces, ensuring both accuracy and fluency. These developments highlight the expanding role of NLP in both accessibility and security, making AI-generated text a crucial area for further research.

3. PROPOSAL DESIGN

This project proposes an AI-driven captioning system designed to address the limitations of traditional closed captioning methods. By leveraging machine learning (ML) and natural language processing (NLP), the system will generate more accurate, contextually aware, and real-time captions for various audio-based media, including internet videos, lectures, and live meetings. The following subsections detail the system's core design components, including data collection, model architecture, speaker differentiation, and latency optimization.

3.1 Data Collection and Training

To improve caption accuracy and adaptability, the AI model will be trained on diverse speech datasets incorporating variations in accent, pronunciation, and speech patterns. This ensures the system can recognize and process different linguistic styles, making captions more accessible for users with distinct speech characteristics. Additionally, specialized datasets featuring overlapping speech from multiple speakers will be collected to enhance the model's ability to differentiate between voices in multi-speaker environments.

3.2 Model Architecture and Processing

The system will employ a deep learning-based NLP model, likely built on transformers or an LSTM-based recurrent network, to analyze spoken language. Unlike traditional speech-to-text models that transcribe words verbatim, this AI will construct sentences that align with natural grammar rules and contextual meaning. The model will also reference surrounding dialogue cues to improve coherence, reducing common errors found in conventional closed captions.

3.3 Speaker Differentiation

A crucial feature of this system is real-time speaker differentiation, enabling accurate attribution of dialogue in group settings. By training on multi-speaker datasets, the model will learn to distinguish between different vocal patterns, assigning captions accordingly. This feature is particularly useful for meetings, lectures, and discussions, where distinguishing between speakers enhances clarity and comprehension.

3.4 Latency Optimization for Real-Time Use

To ensure minimal delay between spoken words and displayed captions, the AI will integrate latency optimization techniques, such as parallel processing and lightweight model compression. This allows for instantaneous caption generation, making the system suitable for live settings where real-time accuracy is critical.

By addressing these challenges—caption accuracy, speaker differentiation, and real-time processing—this AI-powered system will significantly enhance media accessibility for deaf and hard-of-hearing

individuals, as well as improve caption quality for a wider audience.

4. ANTICIPATED RESULTS

This project is expected to result in a highly accurate, real-time, AI-powered captioning system that surpasses traditional closed captioning methods in accuracy, contextual understanding, and speaker differentiation. By leveraging machine learning and natural language processing, the system will produce captions that are grammatically correct, contextually aware, and semantically meaningful rather than simple word-for-word transcriptions. This will significantly reduce errors caused by misinterpretations of accents, pronunciation variations, and overlapping speech, making captions more readable and reliable for users.

Additionally, the real-time speaker differentiation feature will enhance group discussions, lectures, and meetings by ensuring that each speaker's contributions are accurately attributed. This is particularly beneficial for academic and professional environments, where clear communication is essential. The system's low-latency performance will also make it suitable for live applications, improving accessibility for deaf and hard-of-hearing individuals while also benefiting language learners, people in noisy environments, and professionals relying on captions for clarity. By addressing these challenges, this project will contribute to the broader effort of making digital media more inclusive and accessible to diverse audiences.

5. CONCLUSION

This project presents an AI-powered captioning system designed to improve accessibility for deaf and hard-of-hearing individuals while also enhancing the overall captioning experience for a broader audience. By addressing the limitations of traditional closed captioning and human stenography, this system introduces contextually aware, real-time captions that ensure greater accuracy, readability and speaker differentiation. Its ability to process varied speech patterns, accents and multi-speaker conversations makes it a valuable tool for education, professional communication and media consumption.

Beyond its primary use case, the project has the potential to improve AI-driven speech recognition and natural language understanding, contributing to advancements in accessibility technology. The anticipated benefits include greater inclusivity, improved comprehension for language learners and enhanced usability in challenging listening environments. By developing an innovative approach to captioning, this project reinforces the importance of equitable access to information in digital and live communication settings.

6. FUTURE WORK

The next steps for this project involve further training and refinement of the AI model to enhance captioning accuracy and real-time performance. This includes expanding the dataset with more diverse speech samples, testing the system in various real-world scenarios, and fine-tuning speaker differentiation algorithms to ensure reliability in multi-speaker environments. Additionally, user testing and feedback will be essential to refine the model's performance and address potential challenges in deployment.

Future expansions of this project could explore integrating AI-driven captioning into video conferencing platforms, streaming services and smart devices. Enhancing support for multiple languages

and real-time translation would further broaden the system's impact, making it a valuable tool for global communication and accessibility. Ultimately, this project serves as a foundation for advancing AI in assistive technology, ensuring that audio-based content remains inclusive and accessible for all users.

REFERENCES

Gurunath, R., (2021) A novel approach for linguistic steganography evaluation based on artificial neural networks. In IEEE Access, vol. 9, pp. 120869-120879, 2021, doi: 10.1109/ACCESS.2021.3108183

Wu, J., (2024). Generative text steganography with large language model. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24). Association for Computing Machinery, New York, NY, USA, 10345–10353. https://doi.org/10.1145/3664647.3680562