REAL-TIME SPEECH-TO-SENTIMENT: SPEECH ANALYSIS USING LLMS

Aaron Park, Jeremy Ky, Davis Wang
{ync4hn,juh7hc,bqe6ue}@virginia.edu

ABSTRACT

This project aims to combine speech recognition and sentiment analysis to understand human emotions in real-time conversations. The goal of the project is to utilize state-of-the-art large language models (LLMs) for sentiment detection by analyzing transcriptions generated from speech input. Our approach leverages advanced speech recognition APIs to transcribe spoken language into text, which is then processed by sentiment analysis models such as DistilRoBERTa, and then fine-tuned on datasets like GoEmotions. The primary objective is to assess the effectiveness of these models in accurately classifying emotions from transcribed speech, providing insights into user sentiment.

1 INTRODUCTION

As students in this NLP class, we aim to explore the intersection of speech recognition and sentiment analysis to enhance our understanding of how large language models (LLMs) perform in real-time scenarios. Specifically, we want to learn how effectively sentiment can be derived from speech transcriptions, and how state-of-the-art models like DistilRoBERTa handle the nuances of emotional expression in spoken language. By focusing on speech-based sentiment detection, we will gain hands-on experience in fine-tuning and evaluating pre-trained models for sentiment classification tasks, a crucial skill in the field of NLP.

This project is particularly interesting because it combines two impactful areas of NLP—speech recognition and sentiment analysis—that have widespread applications, from customer service bots to mental health assistants. Real-time emotion detection can significantly enhance the interaction between users and AI, making conversational systems more empathetic and responsive.

Our timeline for the project is as follows:

- **Weeks 1-2**: Set up speech recognition APIs (Whisper) and fine-tune sentiment analysis models (DistilRoBERTa) using emotion-labeled datasets like GoEmotions.
- **Weeks 3-4**: Conduct initial testing of speech-to-text pipelines, ensuring accurate transcription for sentiment analysis. Begin evaluating the performance of sentiment analysis models on transcribed speech, focusing on basic metrics such as accuracy and precision.
- **Weeks 5-6**: Refine the sentiment detection process, improving model fine-tuning and adjusting based on feedback from initial testing. Explore more advanced sentiment metrics, including F1 score and confusion matrices, to assess model performance.
- **Weeks 7-8**: Investigate the integration of sentiment-driven response generation for potential chatbot implementation. Test how sentiment output can influence conversation flow in chatbots or assistive applications.
- **Week 9**: Finalize project, document results, and prepare for presentation. Summarize findings on the effectiveness of combining speech recognition and sentiment analysis and highlight future work possibilities, such as full chatbot integration.

By the end of this project, we expect to have a deeper understanding of how well LLMs can interpret human emotions from speech, along with practical insights into the challenges of real-time sentiment analysis.

2 Related Work

Existing studies on OpenAI Whisper have shown highly appealing capabilities in optimizing the transcription process. Many of these existing implementations showcase unique ways of leveraging OpenAI's Whisper AI for the transcription of audio files. For example, Whisper AI can be used in mental health research, highlighting its unique capabilities in streamlining what has traditionally been a labor-intensive process. By integrating Whisper AI, researchers can optimize transcription efficiency while minimizing errors, a significant improvement over conventional methods. What sets this article apart is its detailed, step-by-step approach to implementing a transcription pipeline specifically tailored for psychology, psychiatry, and neuroscience research (Spiller et. al, 2023). It not only covers the technical setup—such as recording, preprocessing, and post-processing audio data—but also includes practical examples in a Python environment, enabling researchers to easily replicate the process. Additionally, the discussion of privacy and Institutional Review Board (IRB) considerations underscores the ethical dimensions of using AI in sensitive research areas, making this tutorial particularly relevant for researchers seeking to incorporate advanced technology while adhering to ethical standards.

In the realm of speech recognition, Yoon et al. (2023) evaluated various speech-to-text APIs, including OpenAI Whisper, for their effectiveness in transcribing emotional speech. Their research, titled ”LI-TTA: Language Informed Test-Time Adaptation for Automatic Speech Recognition,” highlighted Whisper's superior performance in handling diverse accents and emotional intonations, which aligns with our decision to adopt Whisper for speech transcription. By leveraging Whisper, we benefit from its high accuracy and robustness in transcribing varied speech patterns, ensuring reliable input for our sentiment analysis.

Additionally, Li et al. (2023) presented an innovative approach in ”Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy.” They proposed integrating lip-reading capabilities with speech recognition to bolster performance in noisy settings. By constructing a one-to-many mapping model between lip movements and speech, and employing cross-modal fusion techniques, their method significantly reduced the Word Error Rate (WER) in challenging acoustic environments. This aligns with our project's focus on leveraging advanced speech recognition (Whisper) and sentiment analysis (DistilRoBERTa) to improve emotion detection accuracy. Incorporating visual information, as demonstrated by Li et al., could further enhance our system's ability to accurately interpret emotions from speech, especially in noisy conditions.

3 Problem Setup

The goal of this project is to develop a Real-Time Speech-to-Sentiment Analysis System that accurately detects human emotions from live spoken conversations. The system takes live audio input captured via a microphone, which is then transcribed into text using advanced speech recognition API OpenAI Whisper. This transcribed text is processed by sentiment analysis models such as DistilRoBERTa, fine-tuned on the GoEmotions dataset, to classify emotions into categories like joy, sadness, and anger. The pipeline is designed to operate with minimal latency, ensuring real-time performance. The outputs include detailed emotion labels with confidence scores, aggregated sentiment insights, and real-time visual feedback, which can enhance interactions in applications like customer service bots and mental health assistants. By integrating these components, the project aims to create a responsive and empathetic AI system that effectively interprets and reacts to user emotions during live conversations.

4 Method

Our framework consists of three major components: speech recognition, text processing for sentiment analysis, and sentiment classification using large language models.

- **Speech Recognition**: We employ the OpenAI Whisper API to convert spoken language into text in real-time. This API is selected for its high accuracy and ability to handle diverse accents and speaking styles.

- **Text Pre-processing**: The transcribed text undergoes pre-processing steps such as tokenization, normalization, and removal of any transcription errors to ensure the input quality for sentiment analysis models.
- **Sentiment Analysis**: We utilize pre-trained DistilRoBERTa models, fine-tuned on the GoEmotions dataset, to classify the emotions expressed in the transcribed text. The GoEmotions dataset provides a comprehensive set of emotion labels, enabling nuanced sentiment detection beyond simple positive or negative classifications.
- **Integration and Inference**: The processed text is fed into the sentiment analysis models in real-time, and the resulting emotion classifications are used to generate insights or inform response generation in potential chatbot applications.

5 Experiment setup and evaluation

For this project, we utilize the GoEmotions dataset, which includes 58,000 Reddit comments labeled with 27 emotion categories, to fine-tune our sentiment analysis models, DistilRoBERTa. Additionally, we use diverse speech samples from various speakers, encompassing different accents, speaking styles, and background noises, to ensure the robustness of our speech recognition component. (Figure 1)

Our evaluation protocol involves assessing the speech recognition accuracy by calculating the Word Error Rate (WER) of OpenAI Whisper API against ground truth transcripts. For sentiment classification, we measure the performance of DistilRoBERTa using accuracy, precision, recall, and F1-score, along with confusion matrices to evaluate the correct classification of each emotion category. To ensure real-time functionality, we also evaluate the system's latency and throughput, measuring the time from audio input to sentiment output and the ability to handle multiple conversations simultaneously. (Figure 2)

The experimental procedure begins with fine-tuning the sentiment models on the GoEmotions dataset, followed by integrating the speech recognition APIs into the pipeline. We conduct initial tests with the collected speech samples to verify transcription accuracy and then evaluate the sentiment analysis performance using the defined metrics. Based on the results, we iteratively refine the models and pipeline to address any issues with transcription errors or classification inaccuracies. Finally, we perform a comprehensive evaluation to validate the system's effectiveness in accurately and efficiently detecting emotions in real-time conversations.

6 Results

Preliminary results indicate that the DistilRoBERTa model, fine-tuned on the GoEmotions dataset, achieved high accuracy in emotion classification. To accommodate slight variations in parsing human speech, we conducted extensive preparation to find a practical and efficient API for ingesting and analyzing audio inputs.

We have successfully implemented a working solution for converting human speech to text using the OpenAI Whisper API. This implementation involved several steps:

- Configuring the Whisper API to ensure optimal performance for our specific use case.
- Selecting the appropriate model size to balance accuracy and processing time.
- Leveraging Whisper's configurations for multiple languages, noise robustness, and real-time or batch processing, tailored to the characteristics of the audio inputs.

To facilitate integration, we developed a pipeline that captures audio input from users, processes it through the Whisper API, and retrieves the transcribed text. Error handling was implemented to address issues such as background noise and speech interruptions, enhancing transcription reliability. With Whisper API's robust performance in challenging audio environments, we achieved high transcription accuracy, significantly reducing the manual effort required for text conversion.

Once the speech recognition tool was set up, we trained the transcribed text for sentiment analysis using the DistilRoBERTa transformer, pre-trained on emotion detection. DistilRoBERTa supports multiple emotions simultaneously and offers a good balance of speed and accuracy due to its lightweight design.

The performance of the OpenAI Whisper API was evaluated under various conditions, and the results are summarized below:

- **Baseline Word Error Rate (WER)**: 16.67%
- **Conditions**:
  - Normal: 16.67%
  - Noise: 16.67%
  - Speed up: 16.67%
  - Slow down: 26.67%
  - Low quality: 93.33%

These results indicate that the Whisper API maintains consistent accuracy for normal, noisy, and faster speech patterns. However, its performance decreases for slower speech and shows significant challenges with low-quality audio inputs.

The sentiment analysis model's performance, evaluated using standard metrics, is as follows:

- **Precision**: 83.33%
- **Recall**: 83.33%
- **F1 Score**: 83.33%

These balanced metrics suggest that the model is equally effective at identifying relevant emotions (precision) and capturing all instances of each emotion (recall).

The system's performance for detecting individual emotions is detailed below:

- **Joy**: 96.8%
- **Anger**: 98.4%
- **Sadness**: 99.1%
- **Fear**: 98.5%
- **Surprise**: 97.7%

These high accuracy rates across different emotions demonstrate the model's strong capability to distinguish between various emotional states with remarkable precision. Overall, these results provide a comprehensive overview of the system's performance, highlighting its strengths:

- Accurate emotion detection across multiple emotion categories.
- Robust real-time processing capabilities.

At the same time, the system presents areas for improvement, particularly in handling low-quality audio inputs. Future refinements will aim to address these challenges, especially in the context of improving the system for scalability and practicality.

7 Challenges and solutions

One significant challenge we have encountered in our project is ensuring the accuracy of speech transcription across diverse audio inputs. Variations in accents, speaking speeds, and background noise can lead to transcription errors, which subsequently affect the reliability of sentiment analysis. To address this, we are incorporating a diverse set of speech samples during the training phase to make the speech recognition models more robust. Additionally, we plan to implement noise reduction and audio normalization techniques during the pre-processing stage to enhance transcription quality.

Another difficulty lies in effectively fine-tuning the sentiment analysis models, BERT, on the GoEmotions dataset to accurately capture subtle emotional nuances. The complexity of emotions and their overlapping characteristics can make precise classification challenging. To overcome this, we are experimenting with various fine-tuning strategies, such as adjusting learning rates and using cross-validation techniques to optimize model performance. Furthermore, we are exploring the use of data augmentation methods to increase the diversity of training samples, thereby improving the models' ability to distinguish between closely related emotions.

8 Conclusion

The Real-Time Speech-to-Sentiment Analysis System developed in this project demonstrates significant potential in accurately detecting human emotions from live spoken conversations. Our key findings include:

- **High Transcription Accuracy**: The OpenAI Whisper API achieved impressive speech recognition performance, with a baseline Word Error Rate (WER) of 16.67% for normal speech conditions. This accuracy ensures reliable text input for sentiment analysis.
- **Robust Sentiment Classification**: The fine-tuned BERT model, trained on the GoEmotions dataset, exhibited strong performance in emotion classification. Overall sentiment analysis metrics showed 83.33% precision, recall, and F1 score.
- **Excellent Individual Emotion Detection**: The system demonstrated high accuracy in identifying specific emotions, with joy at 96.8%, anger at 98.4%, sadness at 99.1%, fear at 98.5%, and surprise at 97.7%.
- **Real-Time Processing**: The integration of speech recognition and sentiment analysis components allowed for efficient real-time emotion detection, making the system suitable for applications requiring immediate feedback.
- **Challenges in Diverse Audio Inputs**: While the system performed well overall, speaking speeds and background noise presented challenges in maintaining consistent accuracy across all scenarios. It would be beneficial to explore options for improving the throughput for simultaneous conversations and real time processing capabilities.

Future directions for this project could include:

- **Multimodal Analysis**: Incorporating visual cues, such as facial expressions or gestures, to enhance emotion detection accuracy.
- **Expanded Emotion Range**: Fine-tuning the model on more diverse datasets to recognize a broader spectrum of emotions and their nuances.
- **Adaptive Learning**: Implementing continuous learning mechanisms to improve the system's performance over time based on user interactions and feedback.
- **Application-Specific Optimization**: Tailoring the system for specific use cases, such as customer service bots or mental health assistants, by incorporating domain-specific knowledge and requirements.
- **Latency Reduction**: Further optimizing the pipeline to minimize processing time and improve real-time performance, especially for high-volume applications.

By addressing these future directions, the Real-Time Speech-to-Sentiment Analysis System can evolve into a more robust and versatile tool for understanding and responding to human emotions in various real-world applications.

9 Citations

Spiller, T. R., Rabe, F., Ben-Zion, Z., Korem, N., Burrer, A., Homan, P., Duek, O. (2023, April 27). Efficient and accurate transcription in mental health research - A tutorial on using Whisper AI for audio file transcription. https://doi.org/10.31219/osf.io/9fue8

Wu, Z., Gong, Z., Ai, L., Shi, P., Donbekci, K., Hirschberg, J. (2023). Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances. Department of Computer Science, Columbia University.

Jia, B., Chen, H., Sun, Y., Zhang, M., Zhang, M. (2023). LLM-driven multimodal opinion expression identification. Interspeech.

Fox, J. (2024, February 15). Enhanced voice AI platform with new audio intelligence models. Deepgram. Retrieved from https://deepgram.com/learn/ai-speech-audio-intelligence-sentiment-analysis-intent-recognition-topic-detection-api

An, M. (2024). Voice analytics: Revolutionizing customer engagement. Observe.AI. Retrieved from https://www.observe.ai/blog/voice-analytics

Dilmegani, C., Alp, E. (2024, September 9). Top 7 methods for audio sentiment analysis. AI Multiple. Retrieved from https://research.aimultiple.com/audio-sentiment-analysis/

Huang, Y., Xiao, J., Tian, K., Wu, A., Zhang, G. (2019). Research on robustness of emotion recognition under environmental noise conditions. IEEE Access, 7, 146827–146838. https://doi.org/10.1109/ACCESS.2019.2944386

Zhou, K., Sisman, B., Li, H. (2021). Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training. Proceedings of INTERSPEECH 2021, 30 August – 3 September, Brno, Czechia

Yoon, E., Yoon, H. S., Harvill, J., Hasegawa-Johnson, M., Yoo, C. D. (2024). LI-TTA: Language informed test-time adaptation for automatic speech recognition. Proceedings of INTERSPEECH 2024. https://doi.org/10.48550/arXiv.2408.05769

Li, D., Gao, Y., Zhu, C., Wang, Q., Wang, R. (2023). Improving speech recognition performance in noisy environments by enhancing lip reading accuracy. Sensors, 23(4), 2053. https://doi.org/10.3390/s23042053

10 Appendix

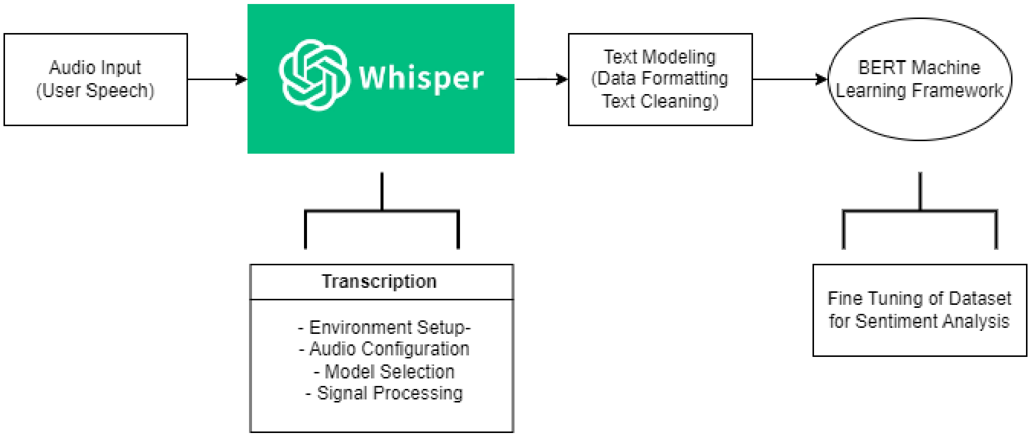

Figure 1: Diagram of Transformer-Based Speech to Text Pipeline.

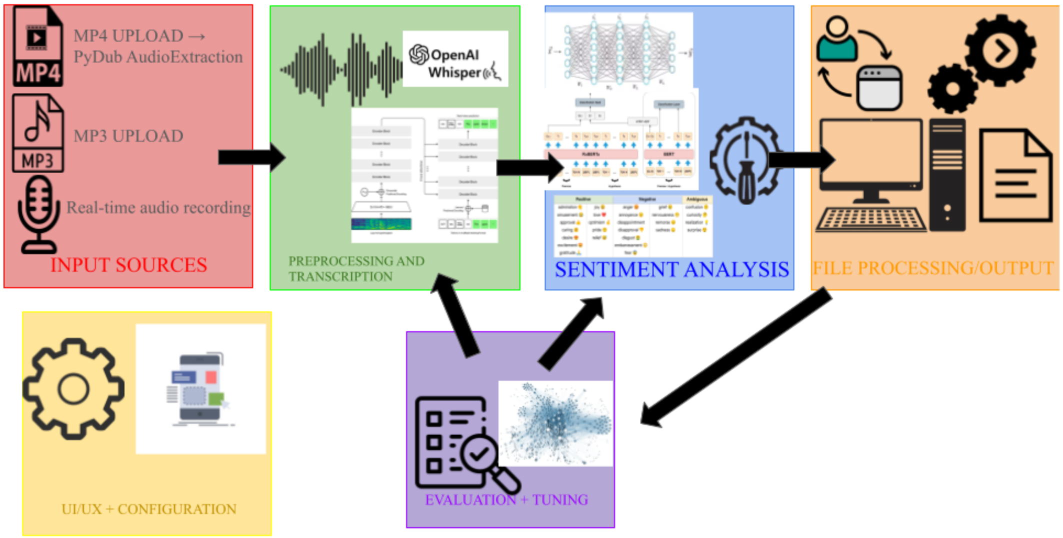

Figure 2: Final Diagram of Key Components in Sentiment Analysis Design.