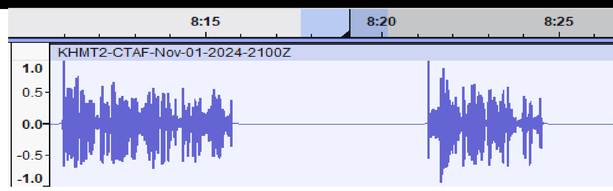
CAPSTONE DESIGN EXPO | **Computer** Science | **CS** 25-346

AI Speech to Text for Military Communications

Team members: Nathan Devore, Nate Eldering, Connor Kohout, Allen Lee | **Faculty adviser:** Tamer Nadeem, Ph.D. | **Sponsor:** DoD | **Mentor:** Clinton Farrell



Problem Statement

Military operations occur in noisy environments and involve specific terminology that current speech-to-text models cannot handle effectively. Our project addresses this challenge by fine-tuning OpenAI's Whisper model, a robust speech-to-text system, for non-civilian environments.
Initial testing with real air traffic communications shows the Whisper model performs inconsistently. Our fine-tuning aims to overcome these issues, improving transcription accuracy in noisy, mission-critical settings.

What is Whisper?

Whisper is an open-source, general-purpose speech recognition model developed by OpenAI. Trained on a diverse dataset of multilingual and multi-accented audio, Whisper offers robust performance across various scenarios, excelling at tasks like transcription, translation, and language identification.
Designed as a multitasking system, Whisper supports multilingual speech recognition, seamless speech translation, and can accurately detect the spoken language in an audio input. It comes in six model sizes, four of which are English-only, each offering a tradeoff between speed and accuracy. Whisper's versatility and robust performance make it a powerful tool for a range of speech and language applications.

Current Design Flow

Audio Collection → Data Processing → Open AI Whisper Model → Text File Output

- Currently, real ATC audio is collected from LiveATC.net archives
  - LiveATC has archived audio from air traffic communications

LiveATC.net
Live Air Traffic — From Their Headsets to You.

- Silence can confuse the whisper model and cause it to hallucinate, so all silence is removes using FFmpeg

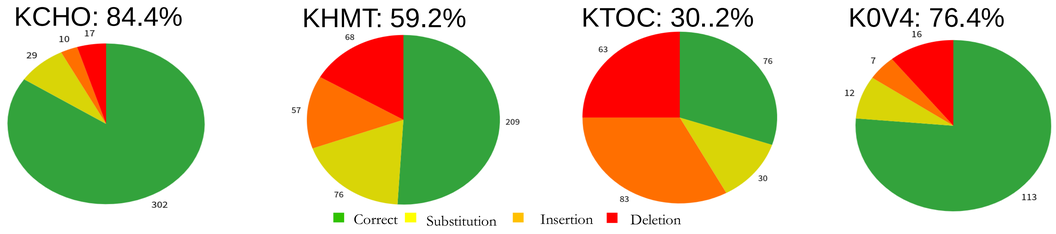
FFmpeg

- Whisper is an open-source audio captioning tool developed by open-AI
  - We have found the medium model transcribes the audio the best
- The current Whisper model is inconsistent and not very accurate

OpenAI/Whisper

- Audio is captioned and output as a .txt file
- The file can then be examined for inaccuracies
  - These inaccuracies can help identify what the model struggles with so specific issues can be targeted

Test Results



To establish a baseline for transcription accuracy, we manually transcribed ATC audio and compared the results to outputs from the Whisper model. The dataset comprises four audio files of varying clarity, sourced from LiveATC.net archives. This provides a robust foundation for evaluating the Whisper model.

The audio was collected from the following airports:

- KCHO (Charlottesville Albemarle Airport, Virginia),
- KHMT (Hemet-Ryan Airport, California),
- K0V4 (Brookneal-Campbell County Airport, Virginia),
- KTOC (Toccoa Airport, Georgia).

Fine-Tuning

Whisper's baseline performance can be enhanced through targeted fine-tuning using aviation-specific data. By having pilots read scripted communications during flights and collecting the audio through LiveATC archives, we ensure the training data matches real-world radio conditions.

Fine-Tune Targets

In its current state, Whisper struggles with identifying proper nouns, significantly hurting its accuracy. It was unable to recognize "Hemet" (the name of an airport), incorrectly labeling every instance as "have it." When we fine-tune the model, we will target proper noun recognition; this includes airport names, company names, aircraft manufacturers, and types.

Ideal Design Flow

Input Audio → Fine Tuned Whisper Model → Captioned Output

- Audio will be collected live from inside the cockpit or ground communications

Python

- The robust Fine-Tuned Model will process the audio in real time
- Trained on aviation-specific data
- Enhanced comprehension abilities, with better understanding of aviation terminology

- Live transcription generates captions that are efficiently archived in text files for future reference
- Automated extraction of routine information reduces cognitive load on pilots, allowing them to focus on critical tasks
- Minimizes human error in aircraft communication

VCU College of Engineering