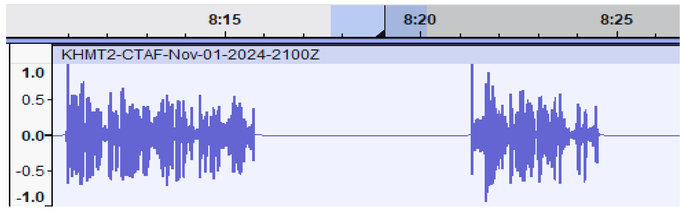
AI Speech to Text for Military Communications

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. Air communications, in particular, are hindered by wind, radio static, and engine noise, degrading transcription quality.
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, struggling with proper nouns, and omitting sentences under high noise conditions. Our fine-tuning aims to overcome these issues, improving transcription accuracy in noisy, mission-critical settings.

Current Design Flow

Audio Collection

- 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.

Data Processing

- Silence is removed from the audio
  - Silence can confuse the whisper model and cause it to hallucinate
- The tool we use is FFmpeg

FFmpeg

Open AI Whisper Model

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

OpenAI/Whisper

Text File Output

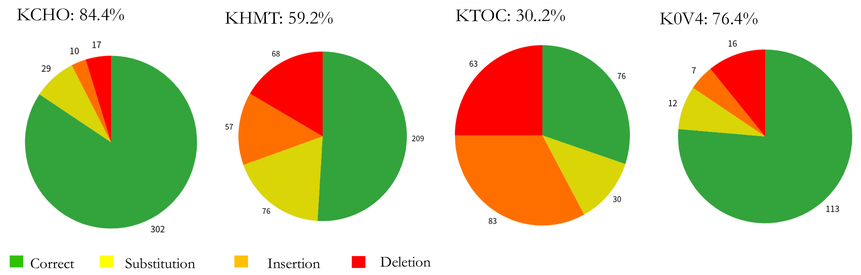
- Audio is captioned and output as a .txt file
- The file can then be examined for inaccuracies
  - These inaccuracies can help identify what the model struggles with so during fine-tuning, specific issues can be targeted

What is Whisper?

Whisper is an open-source, general-purpose speech recognition model developed by OpenAI, introduced in the December 2022 paper "Robust Speech Recognition via Large-Scale Weak Supervision." 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. This flexibility allows users to select models suited to specific needs, from real-time, lightweight processing to high-accuracy transcription. Whisper's versatility and robust performance make it a powerful tool for a range of speech and language applications.

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. These recordings represent diverse real-world scenarios from multiple pilots, aircraft, and radio equipment. The audio was collected from the airports KCHO (Charlottesville Albemarle Airport, Virginia), KHMT (Hemet-Ryan Airport, California), K0V4 (Brookneal-Campbell County Airport, Virginia), and KTOC (Toccoa Airport, Georgia). This provides a robust foundation for evaluating the Whisper model.



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. This precisely labeled dataset will help the model filter out cockpit and radio interference while improving speech detection in noisy environments. The fine-tuning process will also enhance recognition of proper nouns and the phonetic alphabet. Additionally, the model may learn to leverage contextual audio signatures to aid in aircraft identification and improve transcription accuracy.

Fine-Tune Targets

In its current state, Whisper struggles with identifying proper noun, significantly hurting its accuracy. It was unable to recognize "Hemet" (the name of the airport), incorrectly labeling every instance as "have it." In every Hemet-Ryan airport transmission, the messages start and end with "Hemet traffic." 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

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

Python

Fine Tuned Whisper Model

- The robust Fine-Tuned Model will process the audio in real time
- Trained on aviation-specific data
- Matches real-world radio conditions
- Enhanced comprehension abilities, with better understanding of aviation terminology
- Capabilities to identify specific aircraft based on contextual audio signatures

Captioned Output

- Live transcription generates captions that are efficiently archived in text files for future reference and integration with other aviation services
- Automated extraction of routine information reduces cognitive load on pilots, allowing them to focus on critical tasks
- Minimizes human error in aircraft identification and communication, particularly in high-traffic scenarios

VCU College of Engineering