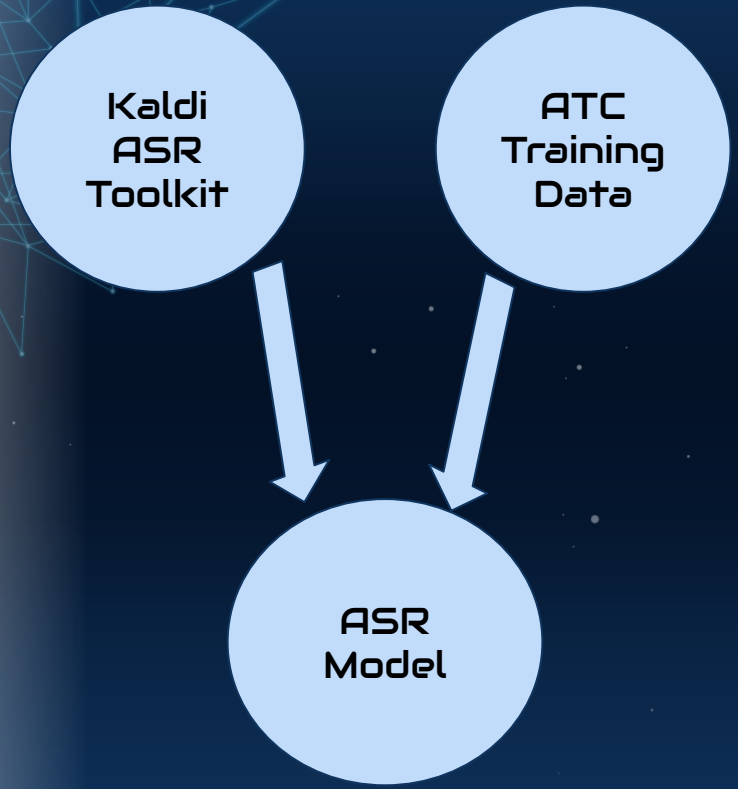
Kaldi ASR Research Team

Team: Tabitha O'Malley, Tamina Tisha, David Serfaty, Milan Haruyama, Adam Gallub
Customer: Dr. Jianhua Liu

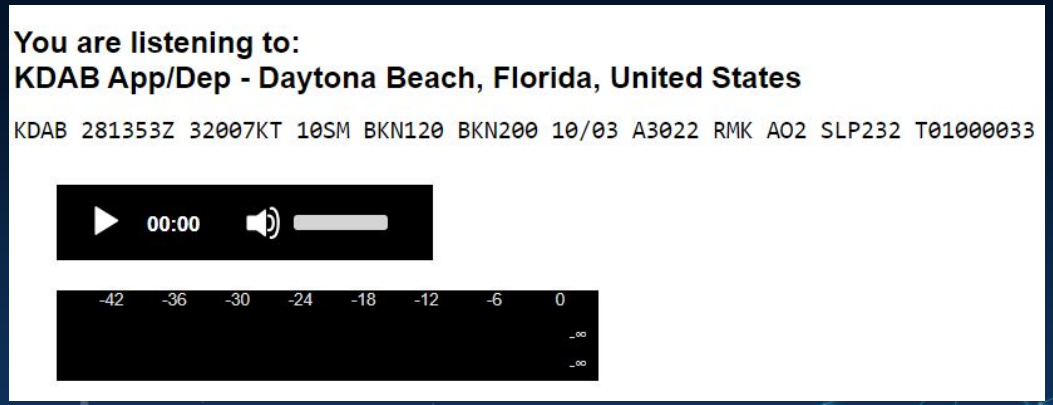


What is the Kaldi ASR Toolkit?

- **Kaldi ASR Toolkit**
  - Builds automatic speech recognition models
- **Automatic Speech Recognition**
  - Conversion of speech to text
- **Training Data**
  - 30-hour ATC dataset
- **ASR Model**
  - Speech recognition for ATC applications

Background

- Complexities in ATC communication requires extensive training
  - Instructor time
  - Lack of existing resources for students
- RTube Web Application
  - Transcribe live ATC transmissions in real time



Customer Needs

In Progress

1. Understanding current sample ASR model provided by Kaldi ASR Toolkit

Future Iterations

1. Adjust the sample ASR model using the 30-hour ATC dataset
2. Compare the performance of the trained models
3. Experiment with other models
4. Apply the best trained model
5. Experiment with callsign and frequency identification

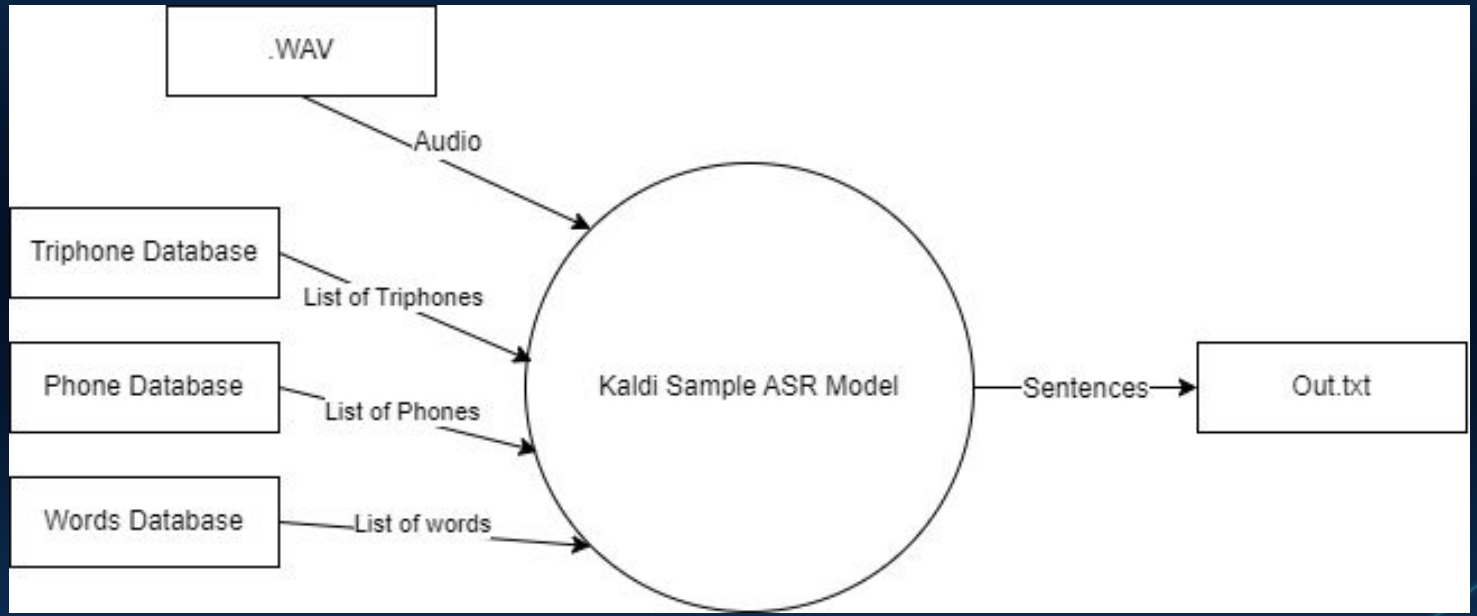
Major Requirements

1. **Computer**
   a. Windows Subsystem for Linux (WSL)
   b. Kaldi ASR Toolkit
   c. Sufficiently powerful GPU
   d. Sufficient amount of memory

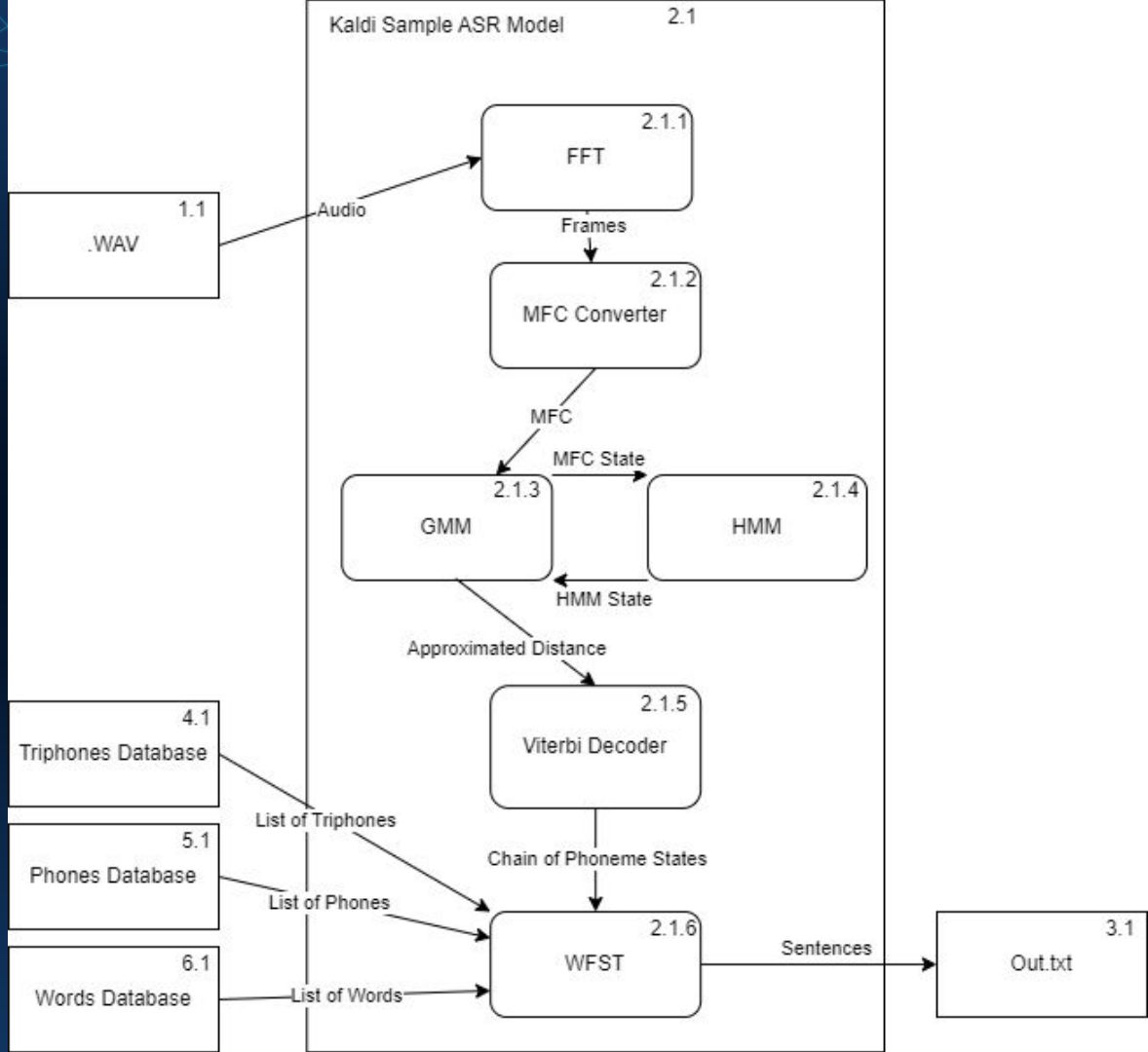
2. **WAV file as input**
   a. Contains spoken speech
   b. Convert non-WAV files using FFMPEG utility

3. **TXT file as output**
   a. Contains words transcribed from the WAV file

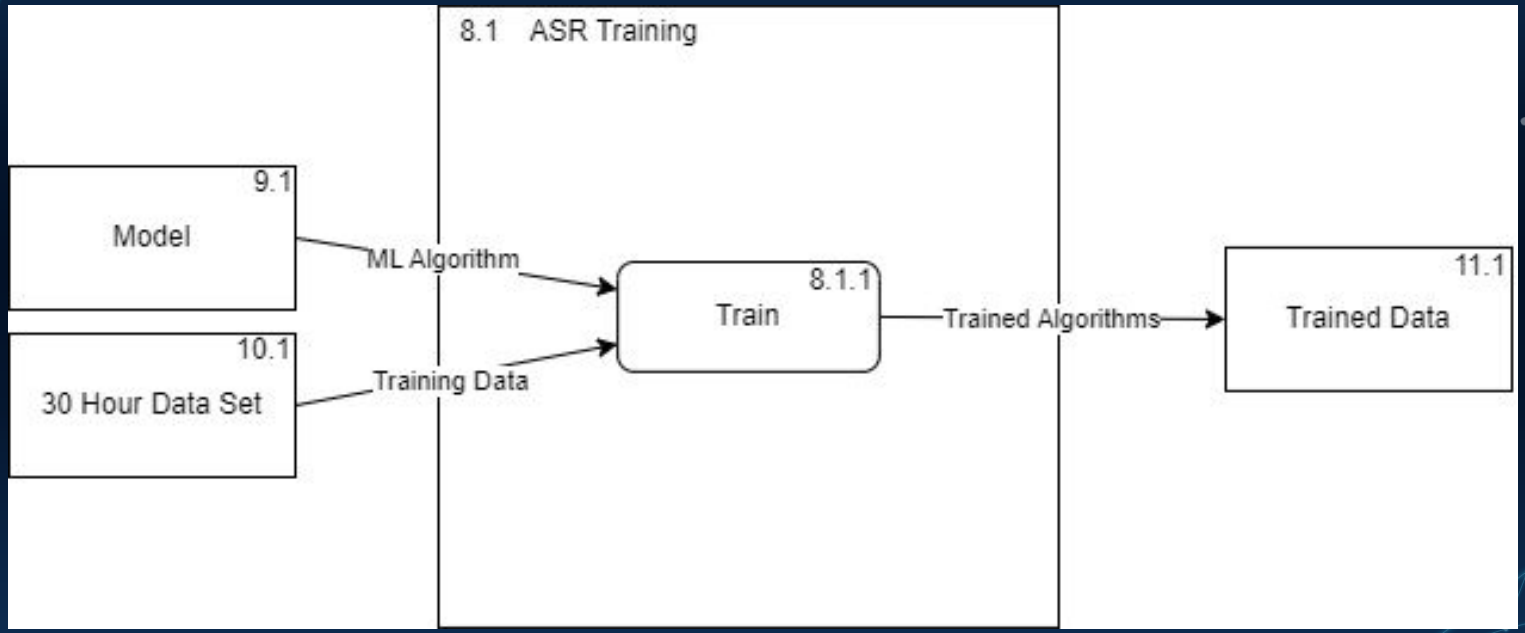
DFD Level 0: Context Diagram



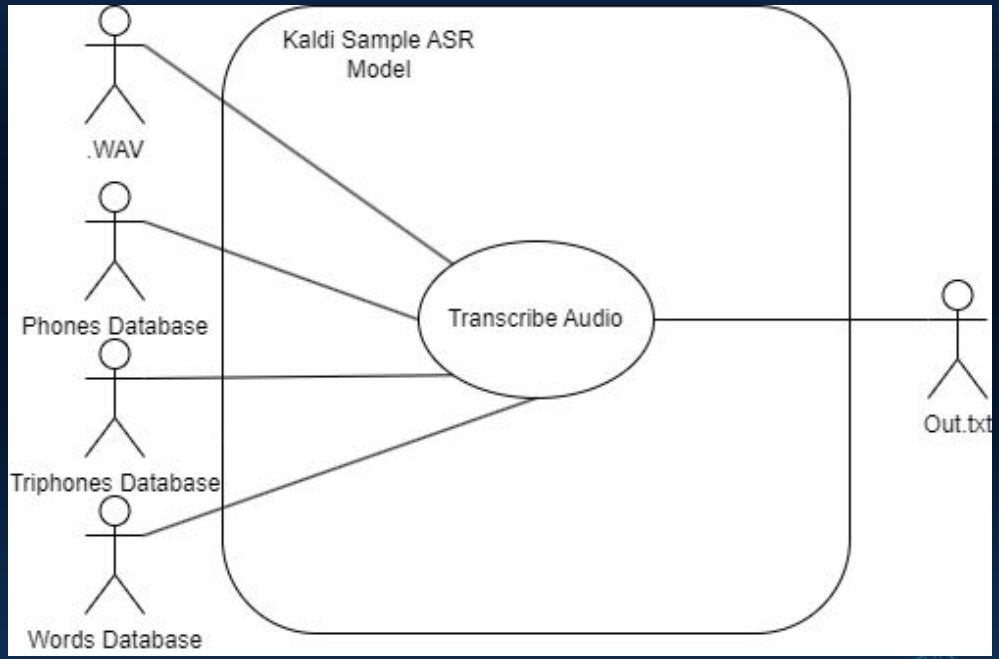
DFD Level 1 Sample ASR Model



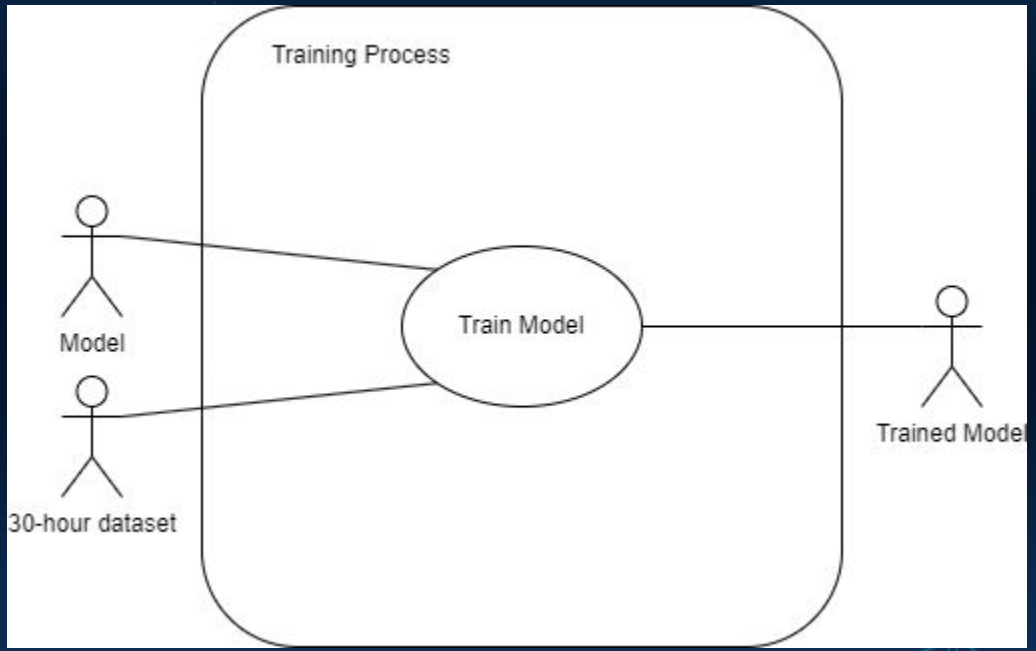
DFD Level 1 ASR Training



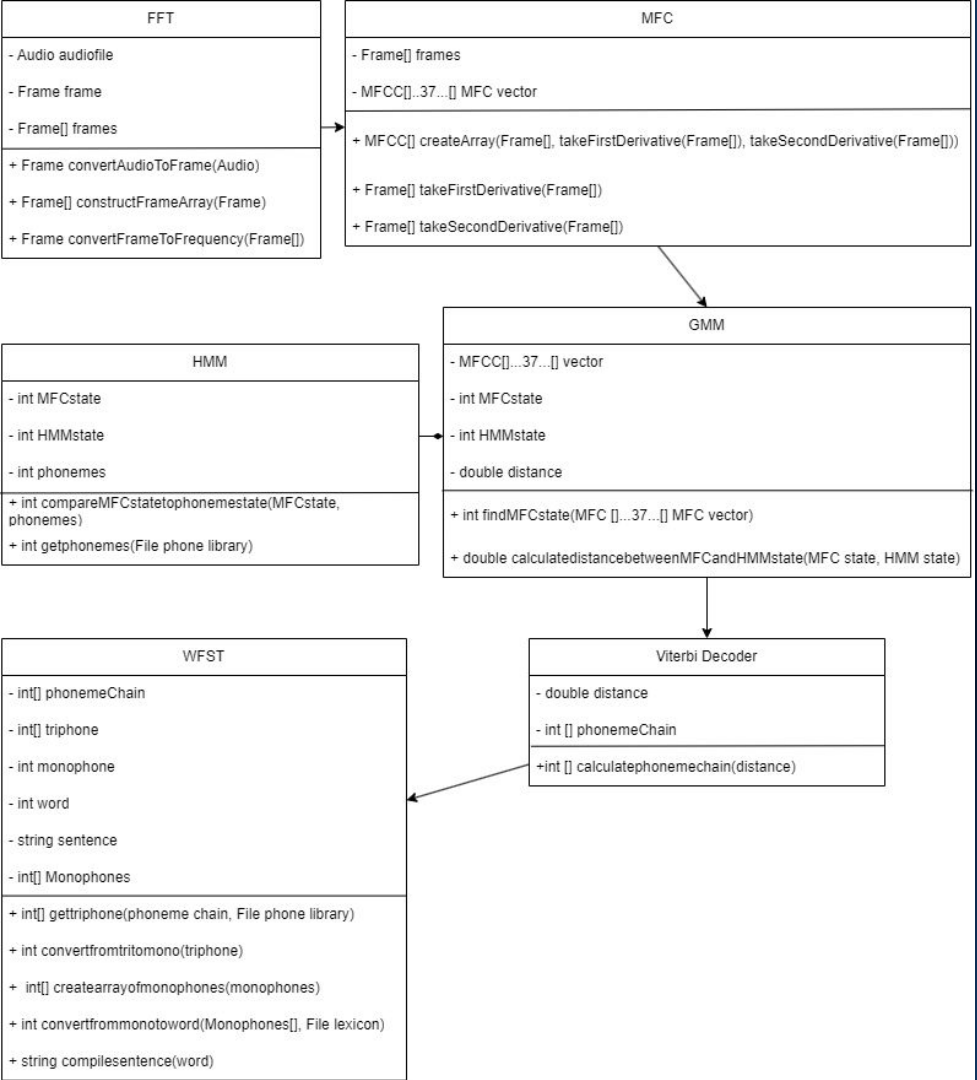
Use Case Diagram: Sample ASR



Use Case Diagram: Training Process



Class Diagram



Future Testing

Input Testing

- “The system shall return exceptions/errors if the input is a non-WAV file.”
- “The system shall return exceptions/errors if there are multiple inputs.”

Accuracy Testing

- Assess transcription accuracy of ASR model using predefined benchmarks (e.g., a minimum transcription accuracy of 80%)

Runtime Testing

- Evaluate efficiency of ASR model based on processing time
- Measure time taken by ASR model to process standard WAV file and produce output text file

Lessons Learned

Complexity of Speech Recognition

- Black box that became slightly more translucent
  - Spaghetti code
  - Bad documentation
- Nuances in human speech
  - Pronunciation (e.g., regional accents)
  - Articulation
  - Volume
  - Pace

Hardware Limitations

- High performance requirements
  - Training speed dependent on GPU speed

Next steps

Fall 2023	Spring 2024	202X	202X	202X	202X
Understand current ASR model (on going process)	**Adjust the model using the 30-hour ATC dataset**	**Compare the performance of trained models**	**Experiment with other models**	**Apply the best trained model**	**Experiment with callsign and frequency identification**

Questions?