Enhanced Speech-to-Text System Documentation

# Introduction

This project focuses on enhancing the accuracy and reliability of speech-to-text systems by reducing ghost results (false positives or fabricated transcriptions) and improving transcription accuracy. The techniques implemented in this project aim to provide a robust solution for handling noisy environments and diverse speakers, with a particular focus on fine-tuning, noise reduction, confidence scoring, and language model integration.

# Techniques Implemented

## 1. Noise Reduction

Goal: Filter out background noise and improve the clarity of the speech signal before it is fed into the speech recognition model.  
Method: We use harmonic-percussive separation and noisereduce to remove non-speech components from the audio, resulting in cleaner input.

## 2. Acoustic Model Enhancement

Goal: Improve the speech recognition model’s accuracy by fine-tuning it on a dataset that reflects real-world conditions (accents, noise, etc.).  
Method: We fine-tune the Wav2Vec2 model, which is pre-trained on large speech corpora. This makes the model more robust to various acoustic conditions.

## 3. Language Model Integration

Goal: Refine the transcription by providing better contextual understanding.  
Method: We use GPT-2 to refine the transcriptions generated by the acoustic model. This ensures that the transcriptions are coherent and contextually accurate.

## 4. Confidence Scoring

Goal: Filter out low-confidence transcriptions that are likely to be ghost results.  
Method: We assign confidence scores to each transcription segment based on the acoustic model’s softmax output. Low-confidence segments are either flagged for review or removed entirely.

# System Overview

The enhanced speech-to-text system follows this pipeline:  
1. Audio Input: An audio file is uploaded and preprocessed.  
2. Noise Reduction: Background noise is filtered out to clean the audio signal.  
3. Transcription: The cleaned audio is transcribed using a fine-tuned Wav2Vec2 model.  
4. Confidence Scoring: The transcription is evaluated for confidence, and low-confidence results are removed.  
5. Language Model Refinement: The transcription is refined using a language model (GPT-2) to improve contextual accuracy.  
6. Final Output: The final transcription and confidence score are provided to the user.

# Setup and Installation

## Prerequisites

Make sure you have the following installed:  
- Python 3.7 or higher  
- pip for package management

## Installation

1. Clone the repository:  
 ```bash  
 git clone https://github.com/your-username/speech-to-text-enhanced.git  
 cd speech-to-text-enhanced  
 ```  
2. Install the required dependencies:  
 ```bash  
 pip install -r requirements.txt  
 ```

# Running the Application

To run the application using Streamlit, use the following command:  
```bash  
streamlit run enhanced\_streamlit\_speech\_to\_text.py  
```  
This will launch the app in your browser, where you can upload an audio file and see the enhanced transcription process in action.

# Docker Deployment

To deploy the application using Docker, follow these steps:  
### Step 1: Create Docker Image  
1. Ensure you have Docker installed on your system.  
2. Build the Docker image using the following command:  
 ```bash  
 docker build -t speech-to-text-app .  
 ```  
### Step 2: Run Docker Container  
3. Once the image is built, run the container with the following command:  
 ```bash  
 docker run -p 8501:8501 speech-to-text-app  
 ```  
This will start the Streamlit app inside the Docker container and expose it on port `8501`. You can access the app by visiting `http://localhost:8501` in your browser.

# Performance Metrics

To measure the system's effectiveness, we use the \*\*Word Error Rate (WER)\*\* as our primary metric. WER quantifies the transcription accuracy by comparing the system's output with the reference (ground truth) transcription.

WER is calculated as:  
```  
WER = (S + D + I) / N  
```  
Where:  
- \*\*S\*\*: Substitutions  
- \*\*D\*\*: Deletions  
- \*\*I\*\*: Insertions  
- \*\*N\*\*: Total words in the reference.

# Demonstration and Results

After implementing the techniques outlined above, we observed a \*\*significant reduction in ghost results\*\* and an improvement in transcription accuracy.

Video Demonstration: [Loom Video Link](https://www.loom.com/share/f85de820e90d41a3a1243fa84d5a0b7a)