Voice-to-Text Converter Documentation

Overview

This document provides a comprehensive guide to the Voice-to-Text Converter, covering its features, functionalities, and usage. This tool is designed to convert audio files into text and evaluate the accuracy of the transcription using various metrics.

Key Features

- **Audio File Support:** Supports MP3 and WAV audio file formats.
- **Speech Transcription:** Utilizes the `faster_whisper` library for efficient and accurate speech-to-text conversion.
- **Accuracy Evaluation:** Calculates Word Error Rate, Match Error Rate, and other key metrics using the `jiwer` library.
- **Interactive Visualizations:** Provides interactive bar charts and visualizations using `Plotly` and `Matplotlib` for error analysis.
- **HTML Report Generation:** Generates a downloadable HTML report with color-coded transcription comparisons.

1. Getting Started

1.1. Installation and Setup

To use the Voice-to-Text Converter, ensure you have the following libraries installed:

- `faster_whisper`
- `jiwer`
- `pydub`
- `matplotlib`
- `plotly`

You can install these libraries using pip:

pip install faster_whisper jiwer pydub matplotlib plotly

1.2. Input Requirements

- **Audio Files:** The tool accepts both MP3 and WAV audio files. For MP3 files, the tool converts them to WAV format using `pydub` to ensure compatibility with the Whisper transcription model.
- **Ground Truth Input:** Users need to manually enter the reference transcription (ground truth) for WER evaluation.

2. Main Workings and Features

2.1. Audio File Upload and Preprocessing

- **Feature:** Supports uploading audio files in MP3 or WAV format.
- **Process:** Converts MP3 files to WAV format using `pydub`.
- **Purpose:** Ensures compatibility for Whisper transcription.

2.2. Speech Transcription using `faster-whisper`

- **Feature:** Transcribes uploaded audio files using the `faster_whisper` library.
- **Process:**
    - Loads the Whisper model (base or other sizes) with optional quantization for speed.
    - Transcribes the audio file segment-wise.
    - Stores the final transcribed text as a clean string.
- **Advantages:**
    - Faster transcription compared to the original Whisper implementation.
    - Efficient for large audio files.
    - Works directly in platforms like Google Colab.

2.3. Ground Truth Input for WER Evaluation

- **Feature:** Allows users to manually enter the reference transcription (ground truth).
- **Purpose:** Provides a basis for comparing the transcribed text and evaluating accuracy.

2.4. WER & Metrics Calculation using `jiwer`

- **Feature:** Calculates various metrics to evaluate transcription quality using the `jiwer` library.
- **Metrics Computed:**
    - **WER:** Overall transcription accuracy.
    - **Substitutions:** Incorrectly replaced words.
    - **Insertions:** Extra words added.
    - **Deletions:** Missing words.
    - **Hits:** Correctly matched words.
- **Benefits:**

    - Objectively evaluates transcription quality.
    - Highlights strengths and weaknesses of the model output.

2.5. Interactive Visualizations

- **Feature:** Generates interactive bar charts and visualizations for error analysis.
- **Visualization:**
    - Bar chart for Substitutions, Insertions, Deletions, and Hits.
    - Uses `Plotly` for interactivity and better visuals.
    - Optional: Pie or other charts to visualize proportions.
- **Insights:**
    - Provides a quick overview of major sources of error.
    - Offers a visual cue for model performance.

2.6. HTML Report Generation with Highlights

- **Feature:** Generates an HTML report with color-coded transcription comparison.
- **Highlights:**
    - Substitutions: Red
    - Insertions: Blue
    - Deletions: Gray
- **Benefits:**
    - Easy to interpret results.
    - Ideal for reporting and presentations.

2.7. Downloadable Report as HTML

- **Feature:** Saves the HTML output with a unique name and generates a downloadable link.
- **Process:**
    - Saves the HTML output with a unique name.
    - Generates a base64-encoded link for downloading directly.

2.8. Final Output Summary

- **Includes:**
    - Display of original vs. transcribed text.
    - Metrics summary.
    - HTML comparison view.

3. Usage

1. **Upload Audio File:** Upload an audio file in MP3 or WAV format.
2. **Enter Ground Truth:** Manually enter the reference transcription (ground truth).

3. **Run Transcription:** Execute the speech-to-text transcription process.
4. **Evaluate Results:** Review the metrics, visualizations, and HTML report to evaluate the transcription quality.
5. **Download Report:** Download the HTML report for detailed analysis and presentation.

4. Advantages

- Complete package for testing and evaluating speech-to-text models.
- Adaptable for academic, commercial, or research applications.

5. Troubleshooting

- **Issue:** If the transcription is not accurate, consider using a different Whisper model size or improving the audio quality.
- **Issue:** If the HTML report is not displaying correctly, ensure that all required libraries are installed and that the HTML file is opened in a compatible browser.