**Speech-to-Text Recognition Tool**

**Project Explainer & User Guide**

**1. Introduction**

The Speech-to-Text Recognition Tool is a modern web application that allows anyone to convert spoken words into written text. Powered by OpenAI’s Whisper model, it supports both uploading audio files and recording speech directly through your browser. It also features a built-in accuracy checker, letting you compare the transcription with the original text and analyze errors in detail.

Whether you’re a student, researcher, developer, or someone interested in speech technology, this tool is designed to be intuitive, informative, and highly accurate.

**2. What Does This Tool Do?**

* **Transcribes Speech:** Converts audio (your voice or files) into written text.
* **Supports Multiple Audio Formats:** Works with .wav, .mp3, and .m4a files.
* **Records Directly from Microphone:** No need for extra software-record and transcribe instantly.
* **Analyzes Transcription Accuracy:** Compares the transcription with the actual (ground truth) text and highlights differences.
* **Visualizes Errors:** Clearly shows where the transcription differs from the original, making it easy to spot mistakes.
* **Provides Detailed Statistics:** Shows error rates and types of mistakes for deeper analysis.

**3. Who Is This For?**

* **Students & Educators:** For transcribing lectures, interviews, or language learning.
* **Researchers:** To analyze speech recognition accuracy.
* **Developers:** As a starting point for building more advanced speech applications.
* **Anyone:** Who wants to convert speech to text quickly and accurately.

**4. Key Features in Detail**

**A. Audio Input**

* **File Upload:**  
  Upload any .wav, .mp3, or .m4a audio file. The tool will process and transcribe it automatically.
* **Microphone Recording:**  
  Click the microphone button to record your voice in real time. The tool gives you visual feedback while recording, making the process interactive and user-friendly.

**B. Transcription Engine**

* **OpenAI Whisper Model:**  
  One of the most advanced speech-to-text models available, offering high accuracy across multiple languages and accents.
* **Editable Transcription:**  
  The transcribed text appears in a clean text area, which you can edit if needed.

**C. Transcription Accuracy & Comparison**

* **Word Error Rate (WER):**  
  The tool calculates how many words in the transcription are different from the ground truth text. This is a standard measure for evaluating speech recognition systems.
* **Visual Highlighting:**
  + **Deletions (Red):** Words missing from the transcription.
  + **Insertions (Green):** Extra words added by the transcription.
  + **Substitutions (Yellow):** Words that were transcribed incorrectly.
* **Detailed Statistics:**  
  See the overall WER percentage, and counts of substitutions, deletions, and insertions. Also, view a list of specific words that were substituted.

**D. User Interface**

* **Modern, Responsive Design:**  
  The app works smoothly on desktops, laptops, and tablets.
* **Simple Controls:**  
  Clearly labeled buttons for uploading, recording, transcribing, and comparing.

**5. How Does It Work? (Technical Overview)**

1. **Frontend (User Interface):**
   * Built with HTML5, CSS3, and JavaScript.
   * Handles file uploads and microphone access.
   * Sends audio data to the backend for processing.
2. **Backend (Server Side):**
   * **Flask (Python):** Manages web requests and responses.
   * **Whisper Model:** Converts audio to text.
   * **jiwer Library:** Calculates Word Error Rate and differences.
   * **Audio Processing:** Uses pydub and soundfile to handle various audio formats.
   * **FFmpeg:** Required for converting and processing audio files.
3. **Workflow:**
   * User uploads or records audio.
   * Audio is sent to the server, transcribed, and the text is returned.
   * User can enter the ground truth text for comparison.
   * The app highlights differences and displays statistics.

**6. Getting Started (Installation & Setup)**

**Prerequisites**

* **Python 3.10 (recommended) or Python 3.9**
* **FFmpeg** (for audio processing)
* **Windows 10 or later, macOS, or Linux**

**Step-by-Step Installation**

1. **Create a Virtual Environment:**
   * **Windows:**

text

python -m venv venv

.\venv\Scripts\Activate.ps1

* + **macOS/Linux:**

text

python -m venv venv

source venv/bin/activate

1. **Install Dependencies:**

text

python -m pip install --upgrade pip

pip install -r requirements.txt

1. **Install FFmpeg:**
   * **Windows:** See SETUP.md for detailed steps.
   * **macOS:** brew install ffmpeg
   * **Linux:** sudo apt-get install ffmpeg
2. **Start the Application:**

text

python app.py

* + Open your browser and go to [http://localhost:5000](http://localhost:5000/)

**7. How to Use the Application**

1. **Upload or Record Audio:**
   * Click the “Upload” button to select an audio file, or
   * Click the microphone button to record your voice.
2. **Transcribe:**
   * The tool will automatically process the audio and display the transcribed text.
3. **Compare with Ground Truth:**
   * Enter the correct (ground truth) text in the provided area.
   * Click “Compare” to analyze the differences.
4. **View Results:**
   * See highlighted differences in the transcription.
   * Review detailed error statistics to understand accuracy.

**8. Troubleshooting**

* **Python Version:** Make sure you’re using Python 3.10 or 3.9.
* **FFmpeg:** Ensure FFmpeg is installed and added to your system’s PATH.
* **Windows Issues:** Refer to SETUP.md for platform-specific help.
* **Still Stuck?** Try the fallback installation methods in the documentation or seek help by raising an issue on the project repository.

**9. Contributing**

* Contributions are welcome!  
  If you have suggestions, bug fixes, or new features, submit a Pull Request.
* The project is open-source under the MIT License (see LICENSE file).

**10. Frequently Asked Questions (FAQ)**

**Q: What audio formats can I use?**  
A: The tool supports .wav, .mp3, and .m4a files.

**Q: Do I need a GPU?**  
A: No, the application is optimized for CPU-only operation.

**Q: Is my audio data stored?**  
A: By default, audio is processed in memory and not stored, but always check the privacy policy if deploying publicly.

**Q: Can I use this tool for other languages?**  
A: Yes, Whisper supports multiple languages, but accuracy may vary.

**11. Final Notes**

This Speech-to-Text Recognition Tool is a powerful, user-friendly solution for anyone needing accurate speech transcription and analysis. Its open-source nature means you can adapt and extend it for your own needs.

If you have questions, suggestions, or need help, please refer to the project’s documentation or reach out to the community.

**Happy transcribing!**