**Automated Speech-to-Text with Grammar and Pronunciation Correction Using Whisper and Grammarly**

**Authors = Rahul Sinha, Koushik Roy, Thomas Benardo.**

**Date – 29/09/1999**

### **Introduction**

This R&D project explores the integration of advanced models to build a seamless pipeline for **speech-to-text**, **grammar correction**, and **text-to-speech** applications. You are utilizing **Whisper v3** for speech recognition, **Grammarly/coedit-large** for grammar correction, and **Microsoft SpeechT5\_TTS** for converting text to speech.

Additionally, you're experimenting with other language models like **phi3** and **LLaMA 3.1 GGUF** to enhance text generation. The primary objective is to create a system that corrects spoken language for grammatical accuracy and then generates natural-sounding speech while maintaining voice personalization through speaker embeddings.

However, the project faces challenges due to hardware limitations, specifically with large models requiring high computational resources, and small models delivering suboptimal performance. The key aim is to develop an efficient, high-performing system within these constraints.

### **Overview of Setup**

1. **Speech-to-Text (Whisper v3)**: This model is used to convert spoken language into text. Whisper is known for its high accuracy in transcribing speech, handling multiple languages, accents, and noisy environments.
2. **Grammar Checking (Grammarly/coedit-large)**: After transcription, the text is processed using **Grammarly/coedit-large** to correct grammatical errors and improve the overall clarity of the text without altering its meaning.
3. **Text-to-Speech (Microsoft SpeechT5\_TTS)**: The grammatically corrected text is converted into natural, personalized speech using **Microsoft SpeechT5\_TTS**, which can also accommodate different speaker voices via embeddings.
4. **Experimenting with LLMs (phi3, LLaMA 3.1 GGUF models)**: Our plan to explore other large language models, like phi3 and LLaMA 3.1 GGUF, potentially for further improving text generation, understanding, or conversational capabilities in the future.

### **Expected Outcome**

You are developing a system that:

* Transcribes spoken language with high accuracy.
* Produces grammatically correct text that maintains the original meaning.
* Outputs natural and personalized speech, improving human-machine interactions.
* Leverages advanced LLMs for further enhancing language understanding or generation.

### **Challenges and Potential Failures (Updated)**

1. **Model Integration & Performance**:
   * Larger models such as **leslyarun/grammatical-error-correction**, **meta-llama/Meta-Llama-3-8B**, **EleutherAI/gpt-neox-20b**, and **google/flan-t5-large** require significant computational resources, particularly RAM. Due to your 29GB RAM limitation, these models may fail to load or run efficiently, leading to slower processing times or system crashes.
2. **Underperforming Small Models**:
   * Smaller models like **t5-small**, **vennify/t5-base-grammar-correction**, **textattack/roberta-base-CoLA**, and **gpt2-large** have been underperforming, resulting in insufficient accuracy in grammar correction or text generation, which affects the overall quality of your system.
3. **Resource Constraints**:
   * Given the hardware limitations, achieving the desired balance between performance and resource consumption poses a challenge. Finding an optimal solution with smaller or more efficient models while maintaining high-quality output is a key hurdle.
4. **Model Fine-Tuning**:
   * Fine-tuning smaller models for better performance can be time-consuming and resource-heavy, with no guarantee of significant improvements, further complicating the development process.

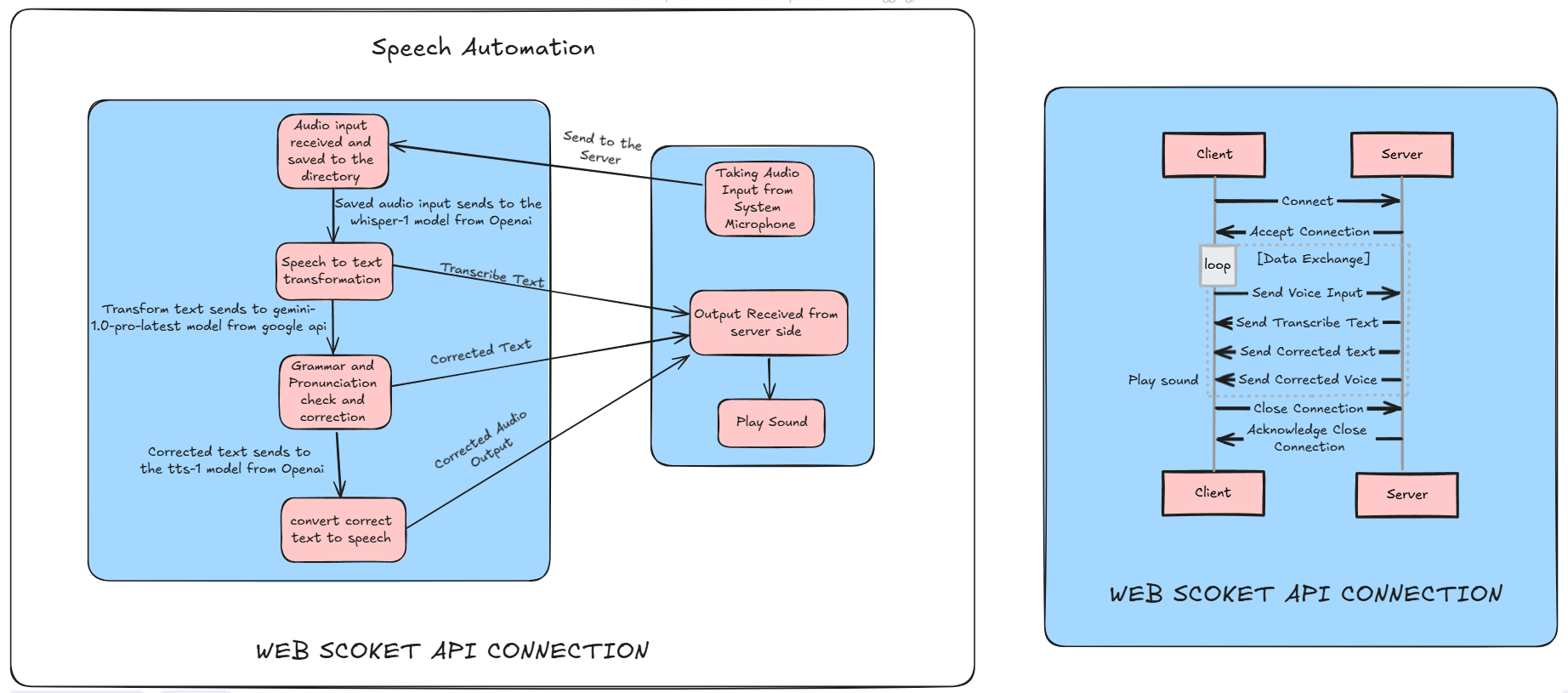
### **Comparision based on Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Model** | | **Purpose** | **Performance** | **RAM/Resource Usage** | **Challenges** |
| Whisper v3 | Speech-to-Text | High accuracy, handles noise well | **Moderate** | No major issues, integrates well into pipeline |
| Grammarly/coedit-large | Grammar Correction | Good correction, accurate | **High** | Requires significant RAM, may slow down processing |
| microsoft/speecht5\_tts | **Text-to-speech** | Natural, customizable with embeddings | **Moderate to High** | Needs speaker embeddings for high-quality output |
| meta-llama/Meta-Llama-3-8B | Text Generation/LLM | High-quality language generation | Very High (Fails to load) | Cannot run due to insufficient hardware resources |
| t5-small | Grammar Correction | Underperforms, lower accuracy | **Low** | Fails to deliver adequate correction results |
| gpt2-large | Text Generation/LLM | Moderate performance | **High** | Struggles with complex language tasks |
| vennify/t5-base-grammar-correction | Grammar Correction | Basic corrections, underperforms | **Low to Moderate** | Struggles with nuanced grammar issues |
| textattack/roberta-base-CoLA | Grammar Correction | Moderate accuracy, limited scope | **Low** | Does not handle complex grammar structures well |
| google/flan-t5-large | Text Generation/LLM | High-quality output, versatile | Very High (Fails to load) | Too large to run with current hardware |
| EleutherAI/gpt-neox-20b | Text Generation/LLM | State-of-the-art generation quality | Extremely High (Fails to load) | Completely unworkable with available resources |
| microsoft/Phi-3.5-mini-instruct | Text Generation/LLM | In testing, potentially powerful | Very High (Fails to load) | Too large to run with current hardware |
| LLaMA 3.1 GGUF | Text Generation/LLM | In testing, potentially powerful | **Low** | Time consuimg and |

**Output:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Processing Time** | **Input File/Text** | **Output File/Text** |
| **Whisper v3** | **8 Sec** | [Recording (3).mp3](https://ogmaitconceptions-my.sharepoint.com/:u:/g/personal/rahul_ogmaconceptions_com/ESyqdYyyy8RNhJCdkcCGa2ABXDMSShtgg6jw4lyH8UfT2A?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJPbmVEcml2ZUZvckJ1c2luZXNzIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXciLCJyZWZlcnJhbFZpZXciOiJNeUZpbGVzTGlua0NvcHkifX0&e=TTGNkZ) | Yesterday I got to the store to buy some vegetables and tomatoes. The cashier said I don't have enough money so I leave without buying anything. |
| **grammarly/coedit-large** | **6 sec** | Yesterday I got to the store to buy some vegetables and tomatoes. The cashier said I don't have enough money so I leave without buying anything. | Corrected text: Yesterday I went to the store to buy some vegetables and tomatoes. The clerk said I didn't have enough money so I left without buying anything. |
| lmstudio-community/Meta-Llama-3.1-8B-Instruct-GGUF | **5 Minute** | I am having two brothers and one sister. My elder brother is working in an IT company and my younger brother is studying in school. I am liking to read books and playing cricket in my free time. Yesterday, I did a party with my friends for celebrating my birthday. It was a good fun, and we enjoyed a lot. | Corrected and Improved Version: "I have two brothers and one sister. My elder brother works in an IT company and my younger brother is studying in school. I like reading books and playing cricket in my free time. Yesterday, I threw a party with my friends to celebrate my birthday. It was a good time, and we had a lot of fun." |
| microsoft/speecht5\_tts | **30 sec(on cpu)** | Corrected text: Yesterday I went to the store to buy some vegetables and tomatoes. The clerk said I didn't have enough money so I left without buying anything. | [speech (2).wav](https://ogmaitconceptions-my.sharepoint.com/:u:/g/personal/rahul_ogmaconceptions_com/EbUXtOYME1BHvRcPblF5jDgBU3xxK1-7lIiioJ0YkrriVw?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJPbmVEcml2ZUZvckJ1c2luZXNzIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXciLCJyZWZlcnJhbFZpZXciOiJNeUZpbGVzTGlua0NvcHkifX0&e=bvmsxM) |

**End to End Flow Chart:**



**Experiments and Experimental Success and Failure:**

1. Experiment in Grammar Correction = Initially we used **Grammarly** open source model which gives very good output but after that you are use **gemini-1.0-pro-latest** which give me more structured output so come to a decision we can use this model.
2. Experiment in Text to Speech = We are experimenting with Microsoft TTS, Google TTS all are open source models and give good output. We use OpenAi TTS paid API.
3. Experiment in Voice Cloning = While we are trying to clone voice in real time we are facing many challenges. We are experimenting on

* <https://github.com/CorentinJ/Real-Time-Voice-Cloning.git> After experimenting with this model, it gives very low-quality output and doesn't even complete entire text to voice.
* <https://github.com/coqui-ai/TTS.git> we have done a complete experiment in google colab which gives very good output but when we try to implement in local it required Ubuntu 18, and this does not support in local system.
* <https://github.com/jnordberg/tortoise-tts.git> experimenting on this model gives very low-quality output.
* <https://github.com/resemble-ai/Resemblyzer.git> This is a voice encoder.

We are experimenting other model like embedding model, synthesizer main issue occurs there are no such kind to good which can give good output if we need real time voice cloning, we need to train model either we need to use paid real time voice cloning models.

1. We use FastApi with Web Socket connection for real time communication.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Experiment** | **Tool/Model** | **Result/Outcome** | **Advantages** | **Challenges** | **Decision** |
| **Grammar Correction** | **Grammarly (Open Source Model)** | Very good output | Effective grammar correction, open-source | No significant challenges mentioned | Initially used |
|  | **Gemini-1.0-Pro-Latest** | More structured output | More structured and organized corrections | No significant challenges mentioned | Decided to use this model |
| **Text to Speech (TTS)** | **Microsoft TTS (Open Source)** | Good output | Open source, good performance | No significant challenges mentioned | Experimenting |
|  | **Google TTS (Open Source)** | Good output | Open source, reliable output | No significant challenges mentioned | Experimenting |
|  | **OpenAI TTS (Paid API)** | Good output | High-quality, reliable output | Paid service | Experimenting |
| **Voice Cloning** | **Real-Time-Voice-Cloning** (CorentinJ) | Low-quality output, incomplete text-to-voice conversion | Open-source, easy to experiment | Low output quality, fails to complete entire voice conversion | Not suitable for use |
|  | **Coqui-AI TTS** | Very good output when run on Google Colab | High-quality output, effective cloning in Colab | Requires Ubuntu 18, does not support local systems | Difficult to implement locally |
|  | **Tortoise-TTS** | Low-quality output | Open-source, easy experimentation | Poor voice quality, not suitable for high-quality voice cloning | Not suitable for use |
|  | **Resemblyzer** | Acts as a voice encoder | Useful for voice encoding | Does not perform voice synthesis directly | Useful as a supporting tool, but not standalone |
| **Other Voice Cloning Models** | **Various embedding models, synthesizers** | Mixed results, no models provided satisfactory real-time voice cloning | Some models show potential | None found satisfactory for real-time voice cloning without further training | Likely need to train custom models or use paid APIs |
| **Real-Time Communication** | **FastAPI with WebSocket** | Supports real-time communication | Efficient for real-time interactions | No significant challenges mentioned | Suitable for use in real-time applications |

**Experiment Notebooks:**

1. [**https://colab.research.google.com/drive/1ZdkNC0um1WNvHV92CnzF2Bm62xQssnYI?usp=sharing**](https://colab.research.google.com/drive/1ZdkNC0um1WNvHV92CnzF2Bm62xQssnYI?usp=sharing)
2. [**https://colab.research.google.com/drive/1ELk63dBXw441TowVNpxycTgTfxOgaENw?usp=sharing**](https://colab.research.google.com/drive/1ELk63dBXw441TowVNpxycTgTfxOgaENw?usp=sharing)
3. [**https://colab.research.google.com/drive/1GoTLwly6qE7EkUM7SQZ6s\_1GsUfRGu1U?usp=sharing**](https://colab.research.google.com/drive/1GoTLwly6qE7EkUM7SQZ6s_1GsUfRGu1U?usp=sharing)
4. [**https://colab.research.google.com/drive/1NxiY3zHN4Nd8J3YAqFsbYaOB71IiLE04?usp=sharing**](https://colab.research.google.com/drive/1NxiY3zHN4Nd8J3YAqFsbYaOB71IiLE04?usp=sharing)
5. [**https://colab.research.google.com/drive/1N8lXviiwfmrS9vYzuFdQnLHlddXMMMvU?usp=sharing**](https://colab.research.google.com/drive/1N8lXviiwfmrS9vYzuFdQnLHlddXMMMvU?usp=sharing)
6. [**https://colab.research.google.com/github/tugstugi/dl-colab-notebooks/blob/master/notebooks/RealTimeVoiceCloning.ipynb**](https://colab.research.google.com/github/tugstugi/dl-colab-notebooks/blob/master/notebooks/RealTimeVoiceCloning.ipynb)
7. [**https://colab.research.google.com/github/olaviinha/NeuralTextToAudio/blob/main/TTS\_voice\_cloning\_pub.ipynb**](https://colab.research.google.com/github/olaviinha/NeuralTextToAudio/blob/main/TTS_voice_cloning_pub.ipynb)
8. [**https://colab.research.google.com/drive/1IZGsMRRN-eoidVlN6crp2q5HCN85kYN\_?usp=sharing**](https://colab.research.google.com/drive/1IZGsMRRN-eoidVlN6crp2q5HCN85kYN_?usp=sharing)