COMP47700 - Project Proposal

ChatGPT Voice Assistant

Declan Atkins
14388146
March 25, 2023

Outline

In this project I will build a ChatGPT voice assistant. The voice assistant will transcribe audio from the user, use this as the input to a query to the ChatGPT API, before synthesising the resulting text to speech. The initial goal of this project will be to build the application using pre-trained ASR and TTS models, following this the advanced goals will be to retrain a TTS model to synthesise a different voice, and fine tune an ASR model to work better with Irish accent inputs.

SMART Goals

Specific - Project Outline

This project will involve interaction between two different speech/audio models as well as an external API. In essence it is a simplistic version of voice assistants such as Siri or Alexa[1] with the tasks/skills of the assistant being replaced by an external API. There are numerous examples of projects online for building your own voice assistant such as this one [1], however with the release of the ChatGPT API, I think there is a possibility to extend on these to something a bit more interesting, by linking with the API and performing speech synthesis.

Measurable - How to measure success

The success of this project will be measured as follows:

1. Is the fundamental application implemented, i.e. Can you speak and receive spoken output.

[1] https://towardsdatascience.com/build-your-own-ai-voice-assistant-to-control-your-pc-f4112a664db2

2. Has the produced voice been altered? To measure this I will attempt to get the assistant to speak with my voice, as discussed in [2].

3. Has analysis been performed on the accuracy of audio transcription of my voice, and fine tuning been performed to attempt to improve this.

Once these conditions are met, the project will be complete.

Achievable

As I mentioned there are a number of samples of similar projects which will serve as inspiration for this project.

For Speech Synthesis YourTTS is implemented in this repo[2] so that will be a good starting point. As the aim is synthesis with low data, I should be able to provide the dataset myself, however LibriSpeech[3] could also be used.

For ASR Librispeech, combined with self annotated data for fine tuning will be used.

Relevant

This project relates to ASR and speech synthesis. It is quite similar in concept to the sample project for generating commands from speech.

Time Bound

The timeline for the project will be as follows:

Week 1: Build the initial application with pretrained models.

Week 2-3.5: Modify the synthesis to use my voice.

Week 3.5-5: Analyse and fine tune the performance of ASR with my voice.

References

1. Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

2. Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

3. Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[2] https://github.com/coqui-ai/tts