AI Whisperer - Statement of Work

This Statement of Work (SOW) outlines the scope, objectives, deliverables, and timeline for the implementation of our project - AI Whisperer.

Project Lead: Aamir A. Ansari

Team Members:
Shashank Srivastava
Ashwin Selvendran
Siddartha Rachakonda
Anoushka Tandon

Contact Information:

shashank.srivastava5523@gmail.com

ashwinselva97@gmail.com

siddartha.rachakonda@iitdalumni.com

anoushkatandon21@gmail.com

Background and Motivation:

Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. The Whisper architecture is a simple end-to-end approach, implemented as an encoder-decoder Transformer. Input audio is split into 30-second chunks, converted into a log-Mel spectrogram, and then passed into an encoder. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.

Accessibility: One of the primary motivations is to make information more accessible to individuals with disabilities, particularly those who may have difficulty reading or typing.

Convenience: Speech-to-text technology enhances user convenience by allowing individuals to input information using their voice rather than typing.

Multilingual Support: Speech recognition systems can support multiple languages, allowing users to interact with devices and applications in their preferred language.

Problem Statement:

Automatic Speech Recognition (ASR) is a technology that converts spoken language into written text, enabling machines to understand and process human speech. In online education, ASR plays a crucial role in video transcription by automatically converting spoken words in educational videos into accurate and searchable text. This facilitates the creation of subtitles, enhances accessibility for diverse learners, and enables efficient content indexing, making educational materials more inclusive and easily navigable.

The objective of this project is to develop an Automatic Speech Recognition using the architecture of the state-of-the-art ASR, OpenAI's Whisper model. This model will be used specifically to identify Professor Pavlos's voice and hence we will fine-tune the model on his voice and generate accurate transcriptions.

Key Objectives:

1. Collect and preprocess speech recognition datasets.
2. Develop a OpenAI Whisper based transformer model to generate transcripts from Professor Pavlos's videos.
3. Implement a scalable backend to handle multiple queries.
4. Design an intuitive and user-friendly frontend.
5. Integrate a large language model to provide summary and key points of the video using the generated transcripts.

Potential Datasets:

http://www.openslr.org/12

http://www.openslr.org/94

https://github.com/cricketclub/gridspace-stanford-harper-valley

https://www.openslr.org/51

https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

https://fluent.ai/fluent-speech-commands-a-dataset-for-spoken-language-understanding-research

https://research.google.com/audioset

https://urbansounddataset.weebly.com/

Learning Emphasis:

The project will focus on building a complex transformer-based ASR model and building a scalable application to make it accessible to the end users. Also, we will learn about large language models and prompt engineering.

Research and Development:

We will review the literature about Automatic Speech Recognition systems and their working including the OpenAI Whisper. Additionally, we'll look into signal processing and techniques like Fourier Transform and its variations, that would be relevant for us in preprocessing. Finally, we'll read about large language models and zero shot inference.

Mock Design:

The application will comprise of the following:

1. An interface to upload the videos or take the URL of the videos.
2. Separate web page to show the processed videos.
3. Separate web pages to display the results of each video that comprise of the transcript, the video, the summary and the key points.

Fun Factor:

Exploring the intersection of signal processing and deep learning is a fascinating aspect, which makes this project more attractive.

Limitations and Risks:

Computational limitations when training and deploying complex models.

Milestones:

1. Data collection and preprocessing
2. AI Wisperer model development
3. Backend implementation
4. Frontend development
5. Final testing and deployment