Apply AI - Project Proposal

Leena, Ethar, Ruei-yuan, Charles

Summary of project proposal

We aim to develop an advanced audio silence extraction system that leverages deep learning techniques, specifically targeting and removing awkward pauses. This will enhance audio clarity, offer smoother playback experiences, and streamline voice-based technologies

Topic of Interest

Audio in the Digital Age: Pervasive use of audio content: podcasts, lectures, etc. Important of clear and concise audio for user experience

The Challenges of Pauses: Silences and pauses can disrupt flow and comprehension

Relevance in different Applications: Transcription Services, Speech Recognition

Why: Enhancing listener experience, smoother playback, better comprehension, time saving in manual editing, and potential for improved voice based technologies and applications

Share lessons learned about AI impact and ethics

AI can be used to service those who are not proficient in public speaking. We hope to start with pauses and if successful consider filler words like, “um” or “like”.

We have considered ethical issues such as the sourcing of datasets. For this, we stuck to open-sourced data. Another issue is the biases, which can prevent some voices from being detected. We considered this through diverse datasets.

Type of machine learning algorithm(s) the team intends to use

We plan to use an LSTM (Long Short-Term Memory), a type of recurrent neural network (RNN) Recurrent Neural Network (RNN) are networks with loops in them which allows information to persist.

LSTMS are a special kind of RNN designed to remember information for long periods of time LSTMs are widely used for tasks that involve sequences such as natural language processing

Utilizing an LSTM: LSTMs can be used for audio processing tasks including detention and removal of pauses or silences in speech

Steps

- Feature Extraction: Convert audio to features like spectrograms.
- Sequence Labeling: Use LSTM to label sequence as "speech" or "silence/pause".
- Training: Use labeled audio data for supervised training.
- Post-Processing: Smooth out LSTM predictions to reduce false positives.
- Filtering: Keep segments labeled as "speech" to remove pauses.
- Enhancements: Consider combining LSTMs with architectures like CNNs for improved accuracy.

Research question

- Different accents within audio
- Variation between female v. male voice

Dataset(s) to be used

Words to Extract out of Audio:

- First focus → Long Pauses

Stretch Goals:

- “Ums”
- “Like”
- “Er”
- “Uh”
- “Well”
- “So”
- “You know”

Examples of Pauses in Conversations We Are Analyzing:

- Phone Call Conversation Pause
- Long Pause Interview
- Speech Pause
- Interview Pause

At least 3 to 4 citations

- https://github.com/jim-schwoebel/voice_datasets
- https://urbansounddataset.weebly.com/urbansound.html
- https://web.stanford.edu/class/cs224s/datasets/
- https://www.youtube.com/