Design and Implementation of a Text-to-Speech Dataset Creation Pipeline for the Irish Language

Introduction

The development of a Text-to-Speech (TTS) system for the Irish language presents unique challenges and opportunities. The primary goal of this project was to design and implement a data engineering pipeline capable of creating a comprehensive TTS dataset from YouTube content in Irish. This document details the approach, design decisions, challenges encountered, and a representative sample of the dataset.

Approach

The pipeline's construction was guided by the following steps:

- Data Collection: Identify and download Irish language content from YouTube.
- Audio Processing: Convert video files to audio in a format suitable for TTS applications.
- Transcription: Generate transcriptions for the audio files.
- Dataset Formatting: Organize the audio and text data into a format akin to the LibriTTS dataset.

Design Decisions

Data Collection

- YouTube API: Employed to search for and identify videos in the Irish language, prioritizing clear and articulate speech content.
- pytube: Chosen for its efficiency in downloading videos, ensuring that the audio streams are of the highest possible quality.

Audio Processing

- ffmpeg: Utilized for its versatility in converting various media formats, enabling the conversion of video files to WAV format with specific requirements for audio codecs, sample rates, and mono channels.
- AudioSegment (pydub): A critical tool for segmenting audio files into shorter, more manageable pieces, aiding in the transcription process.

Transcription Using a Fine-tuned Model

- Link to the model: https://huggingface.co/kingabzpro/wav2vec2-large-xls-r-1b-Irish
- Facebook's Wav2Vec2 XLS-R Model: We leveraged the "kingabzpro/wav2vec2-large-xls-r-1b-Irish" model, a fine-tuned version of Facebook's Wav2Vec2 XLS-R for the Irish language. This model was chosen for its state-of-the-art performance in speech recognition, particularly for its ability to understand and transcribe the nuanced phonetics of the Irish language.
- Transformers and Librosa Libraries: These were instrumental in processing the audio files to match the input requirements of the model, ensuring that the transcriptions are as accurate as possible.
- Below are the parameters of the model

Training hyperparameters

The following hyperparameters were used during training:

- learning_rate: 7.5e-05
- train_batch_size: 32
- eval_batch_size: 8
- seed: 42
- gradient_accumulation_steps: 4
- total_train_batch_size: 128
- optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
- lr_scheduler_type: linear
- lr_scheduler_warmup_steps: 200
- num_epochs: 100
- mixed_precision_training: Native AMP

- The result statistics of the model is shown below:

Training results

Training Loss	Epoch	Step	Validation Loss	Wer	Cer
6.3955	12.48	100	2.9897	1.0	1.0
2.3811	24.97	200	1.2304	0.7140	0.3106
1.0476	37.48	300	1.0661	0.5597	0.2407
0.7014	49.97	400	1.1788	0.4799	0.1947
0.4409	62.48	500	1.2649	0.4658	0.1997
0.4839	74.97	600	1.3259	0.4450	0.1868
0.3643	87.48	700	1.3506	0.4312	0.1760
0.3468	99.97	800	1.3599	0.4236	0.1768

Dataset Formatting

- Custom Scripts: Developed to align the transcriptions with their respective audio segments, formatting the dataset in a manner consistent with the LibriTTS standards. This includes adherence to naming conventions, directory structure, and file formats.

Challenges Encountered

- Transcription Accuracy: Despite the advanced capabilities of the fine-tuned Wav2Vec2 XLS-R model, achieving consistently high transcription accuracy was challenging, highlighting the need for ongoing model training and fine-tuning.
- Variability in Audio Quality: The diverse nature of YouTube content meant that audio quality varied significantly across videos, affecting both the ease of transcription and the consistency of the dataset.
- YouTube API Limitations: Encountered challenges with the API's rate limits, necessitating efficient query management and planning to ensure continuous data collection.

- Quality of Transcriptions: Achieving high accuracy in automated transcriptions for the Irish language was challenging due to the limited availability of pre-trained models specialized in Irish.

Execution and Dataset Sample

The pipeline was executed to collect approximately 1 GB of data, demonstrating its functionality. The dataset includes a variety of content types to ensure diversity in speech patterns and vocabulary. Each entry in the dataset consists of a WAV audio file and a corresponding text file containing the transcription, formatted according to LibriTTS standards.

Output

The directory structure that will be outputted by the model will be in a format similar to LibriTTS. The parent directory “Result” will contain the sub-directories with the author/creator’s name. Each author folder contains the directory for each of the videos that have been downloaded and each of these directories will accommodate the video files (.mp4), audio files (.wav), and transcription files (.txt). A sample tree directory is shown below.

```
Result
├──60 Minutes-20240305T024246Z-001
│   └──60 Minutes
│       └──Hurling  Ireland&#39_s national obsession
├──AllThingsGAA-20240305T024429Z-001
│   └──AllThingsGAA
│       └──The Best of Gaelic Football 2022 - GAA
├──An Ghaeilge
│   └──Barack Obama speaking the Irish language   Barack Obama ag caint as Gaeilge
├──An Ghaeilge-20240305T024518Z-001
│   └──An Ghaeilge
│       └──Barack Obama speaking the Irish language   Barack Obama ag caint as Gaeilge
├──Aontú Ireland
│   ├──Acht na Gaeilge Anois
│   └──Cuir Acht na Gaeilge I Bhfeidhm Anois
├──Ballsdotie-20240305T024541Z-001
│   └──Ballsdotie
│       └──UFC&#39_s
```

Conclusion and Future Work

This project successfully designed and implemented a data engineering pipeline for creating a TTS dataset for the Irish language from YouTube content. While challenges such as transcription accuracy and audio quality variability were encountered, the

pipeline provides a solid foundation for further refinement and expansion. Future work will focus on improving transcription accuracy through the development or adaptation of more sophisticated models for the Irish language and enhancing audio processing techniques to better handle variability in source material quality.