University of Sheffield

TITLE HERE

Adam Spencer

Supervisor: Professor JP Barker

A report submitted in fulfilment of the requirements
for the degree of BSc in Artificial Intelligence and Computer Science

in the

Department of Computer Science

April 10, 2023

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Adam Spencer

Signature:

Date: April 10, 2023

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims and Objectives

1.2 Overview of the Report

Chapter 2

Literature Survey

2.1 Understanding Transcription

2.1.1 Manual Transcription

2.1.2 Semi-Automatic Transcription

2.1.3 Fully-Automatic Transcription

2.2 What is ASR?

Automatic Speech Recognition (ASR) is a subfield of computer science which develops technology aimed at allowing computers to understand human speech. At the most basic level, ASR operation consists of running a recording of human speech through a computer model and producing a transcript of said speech. Nowadays most ASR is conducted using machine learning models trained on large datasets of human speech and using a variety of pre- and post-processing steps.

The metric upon which ASR systems are compared is their accuracy, often reported as a percentage word error rate (WER) where lower WER indicates higher per-word accuracy. WER is calculated as

$$\text{WER} = \frac{S + D + I}{N}$$

Where S,D, and I are substitutions, deletions, and insertions respectively, and N is the total number of words in the reference transcript [1]. Such systems may also be compared across other metrics (for example, time complexity) but won't be covered because this work is only concerned with system accuracy.

2.3 Speech Corpora

2.4 Modern ASR

2.4.1 Whisper

In late September 2022, the OpenAI research laboratory (known for their 'GPT-3' language model) released a new open-source ASR system known as 'Whisper' [2]. Whisper is unique in being very large (trained on 680,000 hours of speech data), open-source, and fully supervised; all the training data used to create the model has been accurately labeled and quality-checked by humans, unlike the much larger unsupervised 'BigSSL' model (1,000,000+ hours of data) [3].

Whisper uses a natural language model to perform next-token prediction (in layperson's terms, there is a secondary system trying to ensure the intelligibility of sentences produced from transcription). Practically, this means that conversational speech (i.e. speech which flows as sentences rather than disjoint terms) should be transcribed with a higher degree of accuracy.

Their accuracy results are promising across a range of different speech corpora, outperforming previous state-of-the-art 'wav2vec 2.0' [4]. Interestingly, Whisper doesn't achieve higher performance than some other models on specific corpora, for example on LibriSpeech test-other [5] it is outperformed by two models built atop wav2vec [6, 7], though across a more diverse set of speech corpora Whisper achieves much lower WER [2].

2.5 Collection of Example Data

2.6

2.7

2.8

2.9

2.10 Summary

Chapter 3

Analysis

3.1 Project Requirements

3.2 Ethical, Professional and Legal Issues

Chapter 4

Design

4.1 Risk Analysis

4.2 Project Plan

Chapter 5

Implementation and Testing

5.1 Preparing the Data

While the LifeLUCID corpus[8] consists of conversational audio recordings, each of these recordings are presented as individual stereo WAVE files approximately 10 minutes in length, with each speaker recorded seperately in either the left or right channel. Time-aligned transcriptions accompany these data in *Praat TextGrid* format.

5.1.1 The TextGrid Format

Praat is a piece of software for speech recording and analysis[9] and a *TextGrid* is used to align individual *speech tokens* with the time in which the are uttered in the recording. When viewed in a text editor, *TextGrid* files appear as a descending series of intervals, indexed in the order they occur; with start- and end-times, and individual speech tokens. To illustrate the format, here is a snippet taken from *LifeLUCID*, the utterance is simply "a bush with a yello duck on top";

```
intervals [12]:
  xmin = 20.899
  xmax = 20.971783458461772
  text = "SIL"
intervals [13]:
  xmin = 20.971783458461772
  xmax = 21.05
  text = "a"
intervals [14]:
  xmin = 21.05
  xmax = 21.47
  text = "BUSH"
intervals [15]:
  xmin = 21.47
```

```
  xmax = 21.66
  text = "with"
intervals [16]:
  xmin = 21.66
  xmax = 21.720024609817834
  text = "A"
intervals [17]:
  xmin = 21.720024609817834
  xmax = 22.1
  text = "SIL"
intervals [18]:
  xmin = 22.1
  xmax = 22.49
  text = "yellow"
intervals [19]:
  xmin = 22.49
  xmax = 22.84
  text = "duck"
intervals [20]:
  xmin = 22.84
  xmax = 23.06
  text = "ON"
intervals [21]:
  xmin = 23.06
  xmax = 23.769
  text = "top"
```

Considering that this file contains over 1000 of these intervals, this example should hopefully demonstrate that the *TextGrid* format is not particularly readable. In order to simplify quality checking as well as to allow more accompanying metadata (e.g. ASR results), the utterances shall be moved into *JSON* format.

Due to Whisper being written entirely in Python, to maintain language-homogeneity a Python library named `textgrid.py`[10] was used to read and manipulate TextGrid files rather than dealing with the transcription data using *Praat*.

According to their documentation, the *TextGrid* files for LifeLUCID[8] contain some special, non-speech tokens to denote certain parts of the speech recordings as follows:

- `<SILP>` denotes time where one participant is silent and the other is talking.
- `<SIL>` denotes silent time between words, where the speaker is silent but the other participant is also silent, such as when the speaker is taking a breath.
- `<GA>` denotes either the time before the task begun but the recording had started or external noises picked up by the microphone.

- `<BELL>` replaces moments when a participant has pressed their bell, these moments are also silent in the recording.

Given that these special tokens are marked by the times at which they begin and end, it was possible to segment the large audio files into hundereds of short utterances.

5.1.2 Generating Utterances

The contents, beginning, and end of every utterance where computed using the data available in the *TextGrid* files using a Python script named `get_utterances`. This script operates over a directory containing *TextGrid* files, writing out the utterances as files in *JSON* format.

The script also takes as args; a minimum time between tokens required to end the utterance and a maximum pause time allowed within one utterance. These thresholds allow utterances to be fine-tuned by a user, leading to fewer drawn-out or unreasonably short utterances.

JSON was selected due to its ability to be easily read and understood by a human, unlike *TextGrids*. This allowed for simple verification of the data without the need for more specific software to view the files.

5.1.3 Audio Segmentation

Another Python script named `segment_audio` was created to generate audio files for each utterance. Given two directories as input; one containing `.json` files (as output by the `get_utterances` script) and the other containing `.wav` files representing each audio recording, the audio is split along the beginning and end times of each utterance and output to a new directory.

This script uses the *python-soundfile* module[11] to load audio files into *NumPy*[12] arrays. By multiplying the sampling rate of the audio by the start- and end-times of each utterance, the array indices at the start and end of each utterance are computed. Array slices between these indices represent each utterance, which can then be saved to new audio files using the *python-soundfile* module.

5.2 ASR With Whisper

Whisper is available as a Python module named `whisper`[13]. The module features a `transcribe()` function to transcribe audio files given as a parameter to the function and return an object containing the output of Whisper.

Chapter 6

Results and Discussion

Chapter 7

Conclusions

Bibliography

[1] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, "A review on speech recognition technique," *International Journal of Computer Applications*, vol. 10, no. 3, pp. 16–24, 2010.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision." arXiv:2212.04356, 2022.

[3] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. C. Sim, B. Ramabhadran, T. N. Sainath, F. Beaufays, Z. Chen, Q. V. Le, C.-C. Chiu, R. Pang, and Y. Wu, "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1519–1532, oct 2022.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 12449–12460, Curran Associates, Inc., 2020.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[6] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020.

[7] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," 2021.

[8] O. Tuomainen, L. Taschenberger, and V. Hazan, "LifeLUCID Corpus: Recordings of Speakers Aged 8 to 85 Years Engaged in Interactive Task in the Presence of Energetic and Informational Masking, 2017-2020," *UK Data Service*, May 2021.

[9] "Praat: doing Phonetics by Computer," Mar. 2023. [Online; accessed 8. Apr. 2023].

[10] K. Gorman, "textgrid.py," Apr. 2023. [Online; accessed 8. Apr. 2023].

[11] B. Bechtold, "python-soundfile," 2013. [Online; accessed 8. Apr. 2023].

[12] C. R. Harris, K. J. Millman, S. f. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.

[13] "openai-whisper," Apr. 2023. [Online; accessed 10. Apr. 2023].

Appendices

Appendix A

An Appendix of Some Kind

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.

Appendix B

Another Appendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.