When mistaking “car tyre” for “Kaitaia” is not okay - Towards quantifying language bias in speech technology

ECSE Project 18

Louis Chuo & Henry An

Supervisors: Jesin James & Josh Bensemann

Background

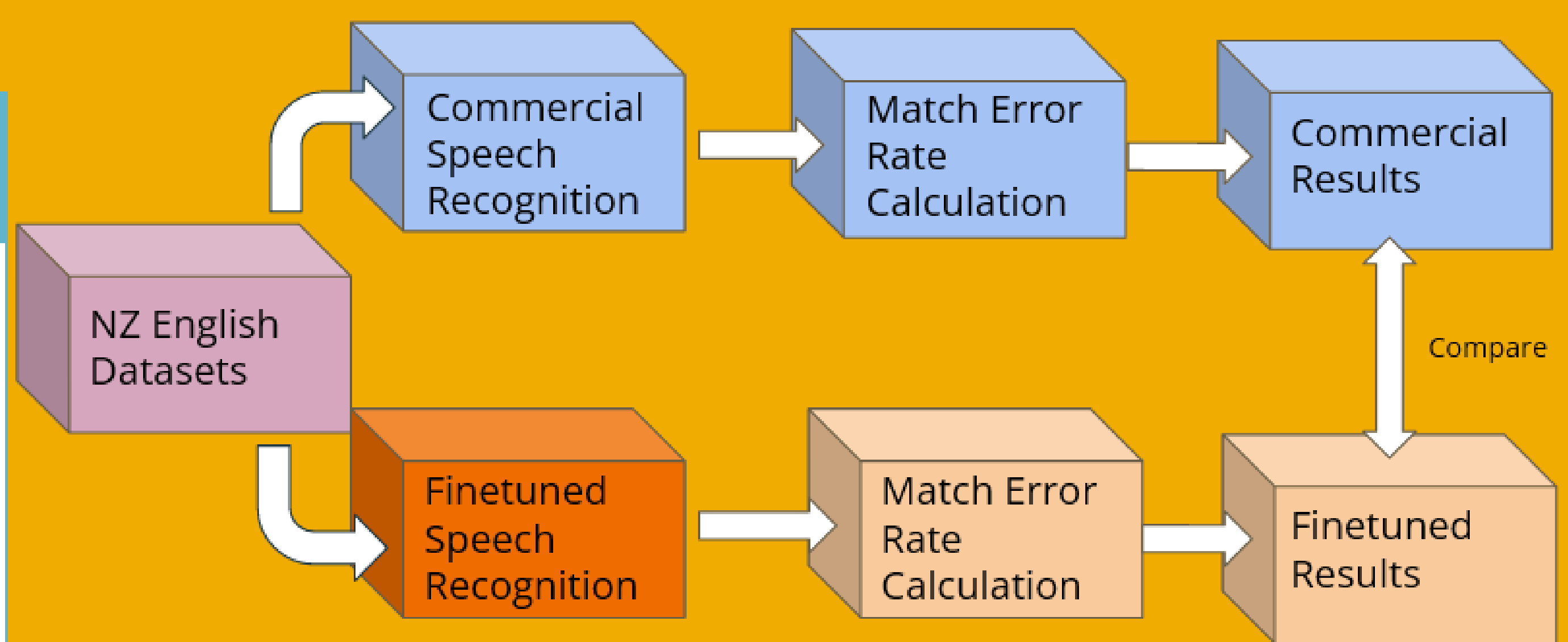
- Speech recognition systems need data to “learn” from - much like how toddlers need experience in order to learn around them.
- Different regions have different ways of pronouncing words, commonly referred to as accents, NZ English is distinct from US English[1].
- Accents can influence how a speech recognition system can learn, developing biases depending on the accents[2]
- As speech recognition technology becomes more widely adopted, it becomes increasingly important that the systems are able to effectively understand a wide variety of different English dialects.

Our goal: Minimize the effects of NZ English on ASR systems to make it so that speech recognition systems can properly “understand” NZ English.

Methodology

NZ English Datasets

Dataset Name	Is Open Source	Number of Lines	Metadata	File Type
JL Corpus	Yes	2,400	Emotion, Gender	.wav
Mozilla Common Voice	Yes	4,366	Gender, Age, Accent	.mp3
Mansfield	No	1,863	Gender	.wav


NZ English Datasets
Commercial Speech Recognition
Match Error Rate Calculation
Commercial Results
Compare
Finetuned Speech Recognition
Match Error Rate Calculation
Finetuned Results

Commercial Speech Recognition

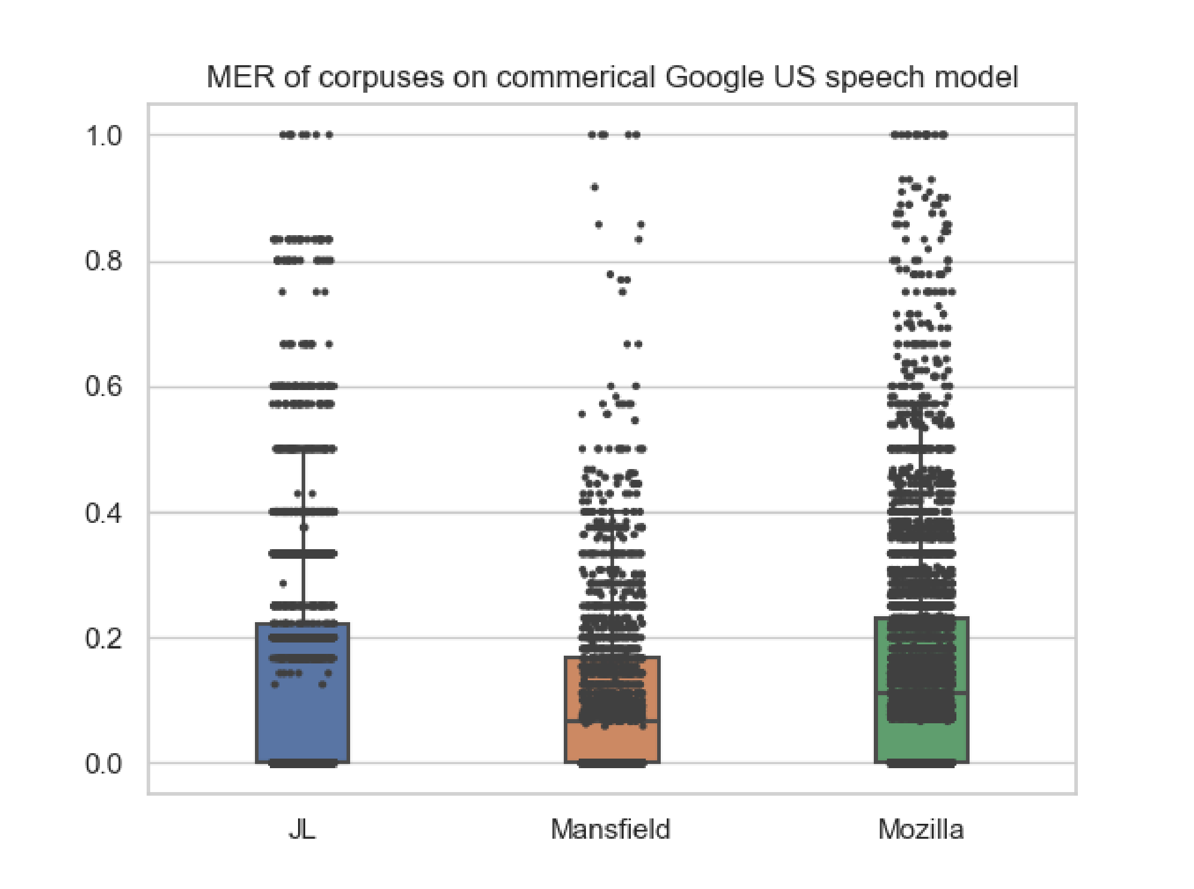
Commercial ASR	Free Credits	Pricing	Has NZ Model	Native mp3 support
Microsoft Azure	US $200	$1 per hour	Yes	No
Google	US $300	$0.006 per 15 seconds	Yes	Yes
Amazon Web Services	None	$0.024 per minute	Yes	yes

MER - Match Error Rate

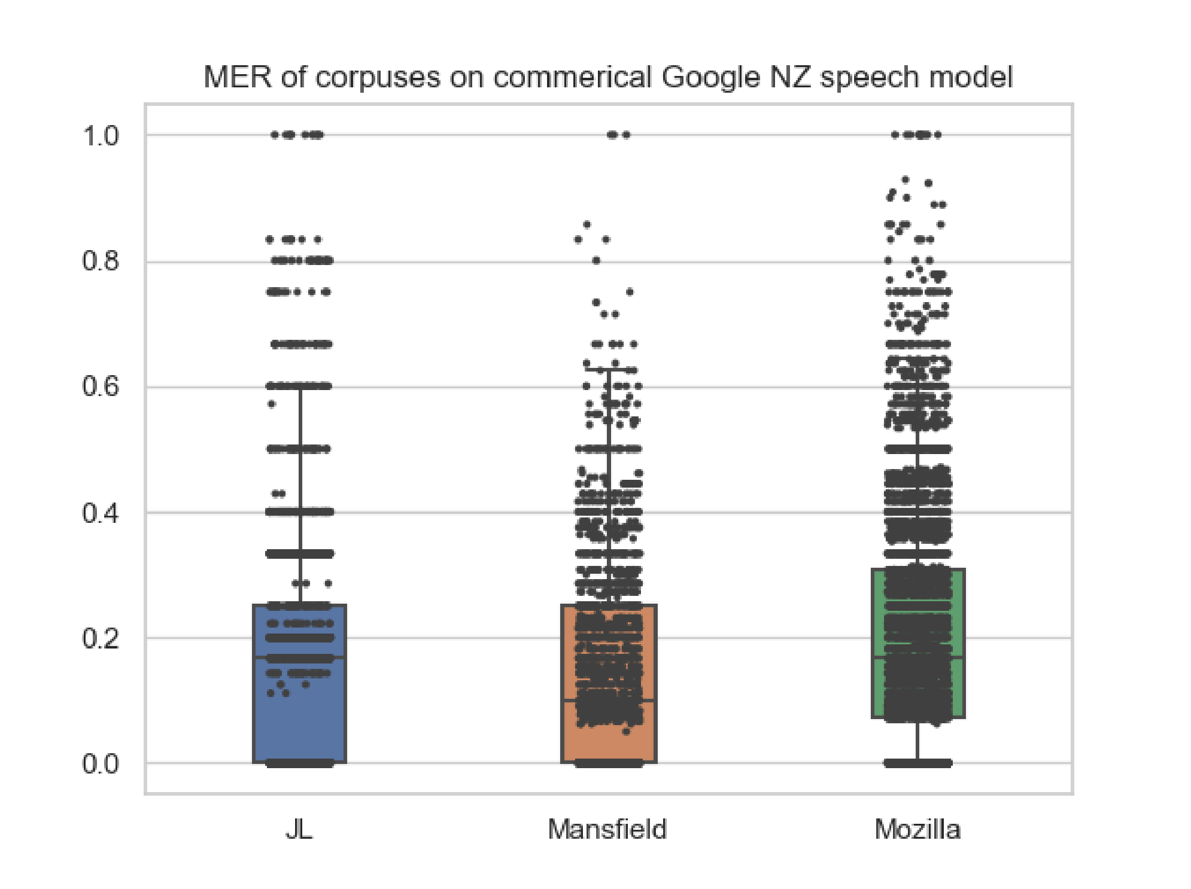
- Ratio of words mistranslated, added, or missing
- **Lower values = better**

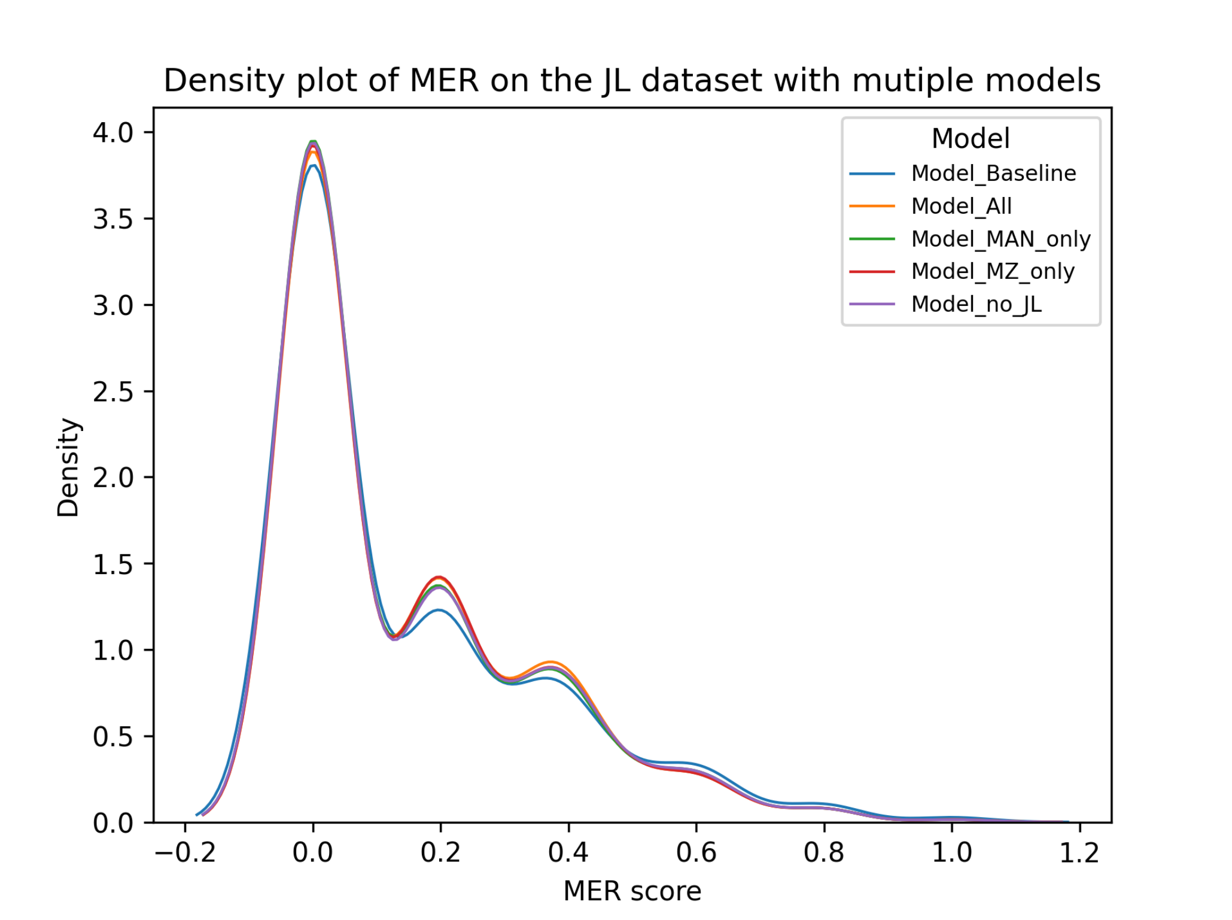
- **Speechbrain[3]** was used as basis for speech recognition system
- It has in-built support for finetuning, multiple pretrained systems available
- Automatically sorts and process audio files to be consistent for training and testing

Results



- Different commercial speech recognition systems were tested to observe their performance on NZ English
- Graphs show match error rate performance on Google speech-to-text US versus NZ models
- Google NZ model performs noticeably worse than the US model on NZ English data
- Same trend present on other commercial speech recognition systems tested like Microsoft





Finetuning Model Datasets

Model Name	Model Baseline	Model All	Model MAN only	Model MZ only	Model no JL
Datasets Used	None	JL, Mozilla, Mansfield	Mansfield	Mozilla	Mozilla, Mansfield

- Finetuning was done on an open-sourced model
- Different combinations of data were tested for training
- Best model performance uses only Mansfield dataset for training
- Best results appear when tested on the JL dataset, 1.6% decrease in average MER

Discussion

- Current datasets are insufficient for properly representing NZ English.
- Commercial systems trained on NZ English performed worse when tested.
- Finetuning results had a lower improvement than expected.
- Some datasets when used for training like JL worsened the performance of the system.

Conclusion: We need a better dataset in terms of quality and quantity for the NZ English dataset.

[1] C. Huang, T. Chen, και E. Chang, ‘Accent issues in large vocabulary continuous speech recognition’

[2] M. Najafian και M. Russell, ‘Automatic accent identification as an analytical tool for accent robust automatic speech recognition’

[3] Speechbrain: A Open-Source Conversational AI Toolkit : https://speechbrain.github.io/

Special thanks to Catherine Watson and Amelie Marchi for letting us access the Mansfield Corpus for research purposes

UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

ENGINEERING
DEPARTMENT OF ELECTRICAL, COMPUTER, AND SOFTWARE ENGINEERING