WHEN MISTAKING "CAR TYRE" FOR "KAITAIA" IS NOT OKAY – TOWARDS QUANTIFYING LANGUAGE BIAS IN SPEECH TECHNOLOGY

by

Henry An

Department of Electrical, Computer, and Software Engineering

Part IV Research Project Final Report

The University of Auckland

October 14, 2022

Supervisors: Jesin James, Josh Bensemann

Teammate: Louis Martin Chuo

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Henry An

Name: Henry An

THE UNIVERSITY OF AUCKLAND

ABSTRACT

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

by Henry An

Supervisors: Jesin James, Josh Bensemann

This report presents a study aiming to improve ASR model performance in regarding to the recognition of NZ accented speech. Our approach to improving was to fine tune an existing model to suit our needs. The adapted model was fine-tuned and tested on 3 NZ English datasets. Three existing commercial systems were tested for comparison purpose. Within the tested existing systems, the Microsoft US model performed the best. An improvement was found in the average MER on NZ English of the speech brain wav2veq model by fine tuning it on samples of NZ English. However, no evidence was found for the proposed model to achieve better performance than existing commercial models. There could be many reasons for this, such as the size of the dataset, the research time allocated, etc. Further study is required on this.

Acknowledgements

The author would like to express sincere appreciation to the supervisors Jesin James and Josh Bensemann for their support and guidelines in the process of this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Accent A distinctive way of pronouncing a language.

Dialect A form of a language which is particular to a specific region or social group

ASR. A research field called Automated Speech Recognition.

Bias When a system produces systematically prejudiced due to properties of the data used in training

CER. Character error rate.

MER. Match error rate.

NZE. An "accent" of English.

PER. Phrase error rate.

WER. Word error rate.

1. Introduction

This report presents the research studies conducted by project group 18 in the first semester and second semester of 2022. The members of the project group are Henry An and Louis Martin Chuo. The research field is in Automated Speech Recognition (ARS)

1.1 The general background

An ASR system allows a computer to take the audio file or direct speech from the microphone as an input and convert it into the text, preferably in the script of the spoken language, and then subsequently use the recognized words as an input to another machine so that some “action” can be performed on [1].

Recent developments in deep learning have provided significant improvements in ASR performance [2], [3]. Accent and dialect recognition is an important part of ASR [4]. ASR has provided many systems for speech recognition, aiming to allow natural communication between humans and computers via speech [5]. These speech recognition systems can help to create voice control for robots [6].

A major challenge in ASR systems is to handle speech from a diverse set of accents [7]. It can be seen that the American variety of English, when measuring by word error rate (WER), achieves the best accuracy results across a variety of different popular modern ASR systems [8]. Further investigation into different types of American English finds that the accuracy is not equal among groups within the US either [9].

1.2 The research intent

The existing literature shows that although there has been tremendous development in the field of ASR, many issues in ASR are yet to be addressed effectively.

In particularly, regarding the accent and dialect recognition, improvements are necessary on interpretation accuracy, recognition error measurements, accent bias etc. Methods such as end to end modelling, multitask modelling or ensemble modelling provide potentials for the improvement. It has been have found that New Zealand dialect generally performs worse than a general American dialect of English by comparing the performance of ASR systems between different dialects. Further investigation into the circumstances of the performance of NZ dialect English is needed. We would like to study the following research questions to address the issues. In this study NZ accent and dialect are not distinguished as the samples we have used can be considered examples of both.

1.3 The research questions

Three research questions were defined as following:

- What is an appropriate quantitative measure for identifying errors in current speech recognition systems?
- How accurate are current accent-aware speech recognition systems in detecting accents?

- How can the accuracy of ASR systems be improved with regards to recognition of NZ English accented speech?

This report presents the research work in an effort to address the above research questions. It starts with literature review, followed by research methods, then the experiments and results, discussion, finally conclusions and future work.

2. Literature Review

2.1 The challenges with existing ASR models

The challenges with the existing ASR models regarding accent and dialect recognition include interpretation accuracy, recognition error measurements, accent bias etc. These systems require a large body of data to achieve good results, however this is often times not available for all dialects [10]. Much of the data available to sample is of the American variety, for example in the Vox Forge corpus over half the body is classified as American. As the ASR software has likely been trained on a greater quantity of American English, it has superior performance when encountering American English again. Given the vast number of varieties of English, each with different number of available samples, it may be infeasible to simply sample English varieties such that a similar performance is obtained for all of them, in addition to considerations on performance overall.

A major challenge in Automatic Speech Recognition (ASR) system is to handle speech from a diverse set of accents [7]. Focusing specifically on English language ASR, it can be seen that the

American variety of English, when measuring by word error rate, achieves the best accuracy results across a variety of different popular modern ASR systems [8]. Further investigation into different types of American English finds that the accuracy is not equal among groups within the US either with Caucasian American varieties achieving the highest accuracy rates [9].

Despite the efforts made by recent research work on improving the accuracy of ASR systems, such as [7], [11], [12], [13], [14], further improvements are necessary when handling accented/ dialectal speech of less resourced dialects in order to bring the performance of ASR systems on them into parity with more prominent dialects such as American. This importance will increase in coming years as technology relying on ASR systems becomes more common.

2.2 Quantitative measures

The quantitative measurement for identifying errors is the main evaluation approach in current Speech Recognition Systems.

Current methods of measuring accuracy in speech recognition systems include measurements on a system level such as word error rate (WER), phrase error rate (PER), character error rate (CER) [8], and match error rate (MER) [9], [15]. These measurements are calculated from the recorded results of testing speech on different ASR systems. Measurements can also be taken and calculated on a sample level such as perplexity. The suitability of the measures depends on the context and purpose for taking the measurements. For example, the CER measurement is a more appropriate for measure for accuracy on Chinese Dialects than English since characters correspond to a whole syllable in Chinese while in

English they are not directly linked to sounds. Comparing WER to PER, WER would likely require less preparatory work as the total number of words can be counted automatically while total phrases will need to be manually classified. However, PER may be considered more suitable for measuring the performance of an ASR system as in many situations, such as for search, the corruption of any words would require manual intervention. In addition WER would bias the results to speech with greater word density.

2.3 Recognition of NZ English

Studies comparing the performance of ASR systems between different dialects have found that although the New Zealand dialect of English is not the worst performing dialect overall, it generally performs worse than a general American dialect of English. For example, when tested with the IBM Watson system NZ English had a word error rate for ~ 0.45 which was greatly lower than the 0.9 of Indian English but still significantly higher than the WER of 0.25 recorded for American English [8].

A potential shortfall of this result is the fact there is not one NZ accent. The general New Zealand dialect fall on a scale between Cultivated, which has traits more similar to British RP, and broad which has more distinctively local characteristics. In addition to this there are also distinct variations of English spoken in New Zealand such as the Southland dialect or the Māori English.

The weaker performance of ASR on New Zealand English when compared to American English is likely due to multiple reasons. Firstly, there is relatively small body of NZ English to work with. For example, in the VoxForge corpus only 2 percent of the total sample are classified as New Zealand

English [10]. In addition to this New Zealand is a relatively small country. At the time of this writing New Zealand has a population of approx. 5 million which is 5 times smaller than Australia, 13 times smaller than the UK and 66 times smaller than the US. Thus, it is less likely for NZ English to be prioritized for the purposes of training ASR models.

Despite these factors the performance of NZ English is similar to or even better than other non-American dialects. For example, in a study by [16] on the performance of the YouTube Captioning system on various dialects, the New Zealand dialect achieved a similar word error rate to New England English with an average WER of ~0.4 and a superior performance when compared to Scottish English which averaged a WER of ~0.55. Considering the significantly larger population and economy of Australia, it is unlikely that this discrepancy was due to a greater sampling of New Zealand speech in the training data. Taking these results, there are likely factors specific to the dialects themselves that influence the accuracy of ASR systems. Further investigation to clarify the performance on New Zealand English is needed.

One limiting factor in the current research is that there is not necessarily a classification of the task being given when recording the sample. [17] shows that the type of task being given to the speaker has an impact on the performance of ASR systems. In the study it was found that while Yakama speech had the highest phrase error rate when sampled during a conversation, it had the lowest PER while sample when performing a reading task. Classifying the tasks being performed could give some insights into the performance of NZ Dialect of English.

2.4 Relevant approaches

There are many existing approaches on improving ASR performance for accented speech. These approaches generally utilise machine learning techniques, such as semi supervised learning, unsupervised learning, multi-task learning, etc. The results of implementing these approaches usually improve WER or CER. For example, [18] used multi-task learning technique, which jointly learns a multi-accent acoustic model and an accent classifier. The results gave 10% relative WER reduction for an unseen accent; 15% relative WER reduction for seen accents. [19] used similar technique, where they jointly learn an accent classifier and a multi-task acoustic model. As a result, 5.94% relative WER improvement on British English, and 9.47% relative WER improvement on American English were obtained. [20] used supervised approaches when transcription of the accented data is available and unsupervised approaches when they are not. The results showed up to 30% of WER reductions. [21] used multi-task learning for native English and Non-native English, achieved 11.95% better CER performance for Hispanic accents and 17.55% CER performance for Indian accents.

3. The research methods

Based on the three research questions outline in Section 1, a quantitative measure for an ASR system needed to be found to measure the accuracy of the ASR system to indicate whether a particular approach has improved the system, so a quantitative approach has been taken. The proposed approach also needed to be implemented to test whether it can improve the performance of ASR on the NZ accent. In order to better understand the performance of existing ASR systems on NZ English and also to provide a baseline to compare the new model to, experiments were performed on several commercial ASR systems.

3.1 The methods

The approach taken in this project was to fine tune an existing pretrained model. This approach was chosen as we did not have the time and resources to construct our own ASR system from scratch. The model used is called wav2vec [22].

For implementation, python was used to prepare the datasets before feeding data into the models, as well as to operate the models and analyse the data. The seaborn python library was used to prepare the graphs. The PyTorch based Speechbrain framework was used as the base for our attempt to create a model with improved performance on NZ English. The new wav2vec2 model was specifically used as it showed promising results. The pretrained model was trained on the LibriSpeech dataset.

The performance of the models on samples of speech were evaluated and compared using the MER. Initially the WER was used like much other research [18], [19, [20], [21], but it was noticed that for cases where the transcribed text had more words than the sample, the WER could be greater than 1. For small samples this could significantly distort the result so we elected to use MER instead as it would allow us to still have a result comparable to WER which is commonly used in research.

3.2. The model

The wav2vec2 [22] model selected is a state-of-the-art model for self-supervised learning. It was chosen due to that it is able to produce good results with a relatively small amount of labelled data. This suited us as we had a limited amount of data to work with as well a limited to tune the model. The model works by first using unlabelled data to construct speech representations before assigning those representations to language.

3.3 The experiment process

As illustrated in Figure 1, the research process began by testing a variety of commercial open-source ASR models using samples of New Zealand. Then MER for each ASR was calculated on each sample, then the distribution of the MER for each model on each corpus was found. After this the same tests were performed on a pretrained model from an open-source framework, Speech Brain. Then each of the datasets were separated into a training, validation and testing set before fine tuning the pretrained model, testing the match error rate again and comparing against the original speech brain model. Different models were created with different combinations of data used for fine tuning with each model being evaluated independently and compared to one another.

As stated earlier, it was intended that WER be used to gage the correctness of the ASR on a sample. However early in the process it was found that in some cases the WER could be extremely high due to the way it is calculated. As such Alternative measurements that did not suffer the same issues were found; MER and WIL[23]. MER was chosen as the measurement for this study as it is less computationally expensive than WIL.

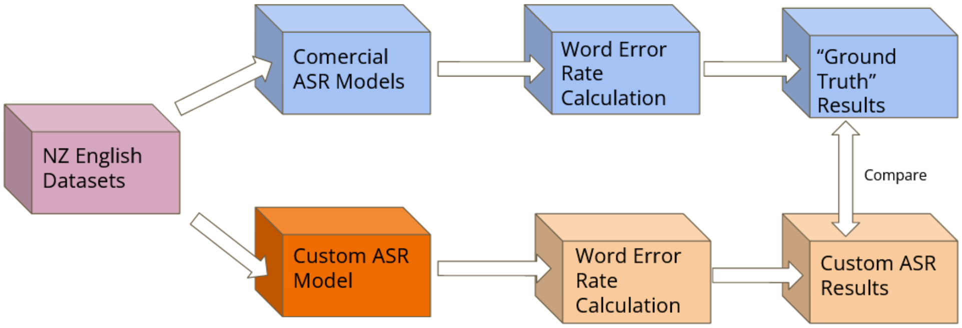


Figure 1. Experiment Process

This process was designed to answer all 3 questions posed with the 2nd question answered through additional research and experience.

The embeddings were then plotted and compared to each other in order to attempt to understand how different the datasets were from each other as well as how NZ and US English compared to each other and the internal variation within the datasets. Due to limited time, only a subset of the data was compared, consisting of the samples with the highest proportion of the most common words between NZ and US English.

4. The experiments and the results

The experiment process has been illustrated in Figure 1. The Figure depicts our planned process, which we only deviated from by a selection of MER over WER.

4.1 The datasets

3 datasets were used in the project, the JL, Mansfield and Mozilla Common Voice corpus. These were selected because they represented a broad array of samples of both American and New Zealand English. Details of the datasets can be found in Table 1. The Mozilla common voice corpus contains a large array of sample from different languages and dialects. Only samples of American and New Zealand English from the Mozilla corpus were used for this research project. The Mansfield and JL corpus are both exclusively NZ English. The JL corpus was made for usage in emotional speech recognition and consists of actors reading a small set of lines expressing different emotions. This may have impacts when testing and training with this dataset as the same expected output may correspond to quit different sounding inputs. The Mansfield corpus consists of 3 individuals reading a large variety of lines while the Mozilla corpus contains both a variety of speakers and lines but also varied in terms of audio quality.

Table 1. Description of different datasets used

Database Name	**Is Open Source?**	**Number of Lines Used**	**Average Sentence Word Length**	**Metadata**	**File Type**	**Sentence Type**	**Baseline Format**
JL Corpus	Yes	2,400	5.39	Emotion, Gender	.wav	Set of sentences, repeated per emotion	Separate text files
Mozilla Common Voice	Yes	4,366	10.411	Gender, Age, Accent	.mp3	Unique sentences sampled from internet	CSV
Mansfield Corpus	No	1,863	9.78	2 Females, 2 Males	.wav	Unique sentences from NZ literature	CSV

4.2 The results

The results from testing the Microsoft and Google ASR systems are graphed as boxplots and can be found in the appendices. The Amazon set was tested differently due to requiring the datasets to be uploaded, which prevented the usage of the Mansfield Corpus. This in addition to uploading time and cost meant it was only tested on JL corpus.

The newly trained models and baseline model performance on the datasets are graphed on a density plot for MER in Figure 2. The density of low MERs can be seen to be higher for the new models, especially the All model and Mansfield model, compared to the baseline model. This indicated that the new models are more likely to have a lower MER on a sample from the datasets we tested and thus performed better.

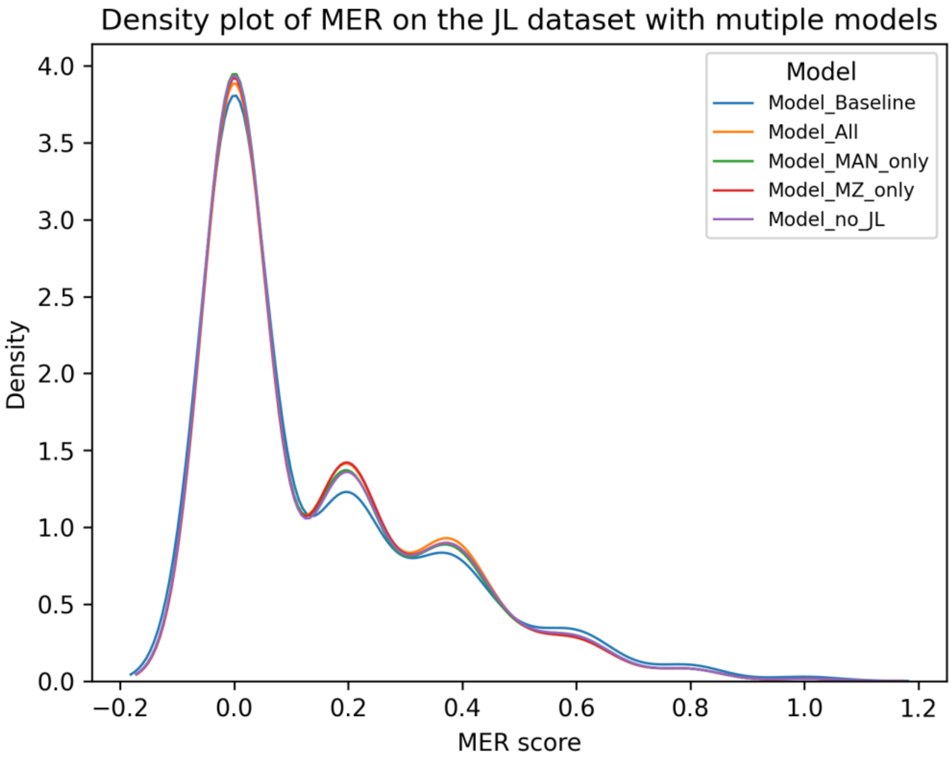


Figure 2. Density plot of MER

Figure 3 shows the plotting of the embeddings, it was found that although NZ and US English were quite similar, there were still noticeable differences. The embeddings plot takes a multi-dimensional comparison of the 2 sample sets and compresses them into 2 principal components. The samples are them plotted according to their principal components in order to visualise the differences between samples.

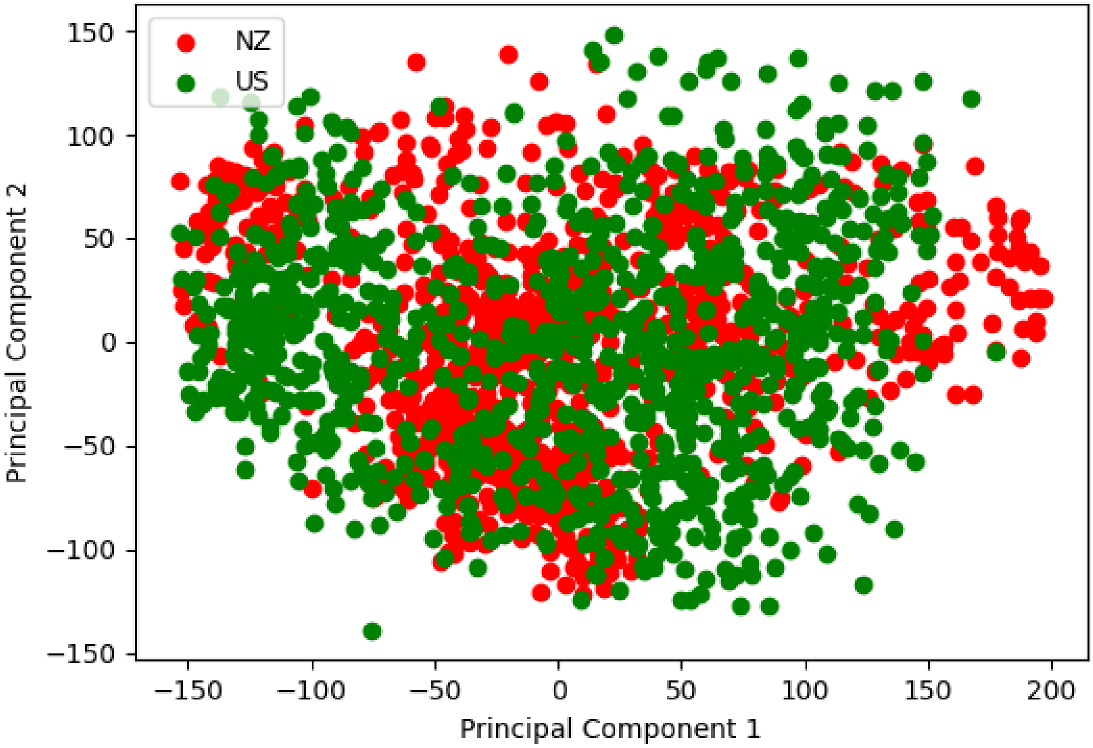


Figure 3. The plotting of the embeddings

5. Discussion

5.1 Data analysis

When testing the commercial as systems, it was found that the Microsoft US model performed the best out of all the models tested, achieving the lowest median MER. Investigating this, it was found that the Microsoft model used a private dataset to train its model. Differences in the quality of the training data could have been the source of the differences in performance.

Of note is the relative performance of the US and NZ English models. The Amazon and Google services both had NZ English and US English models available but when tested on NZ speech datasets, the US models consistently performed better than their NZ counterparts, across all datasets used. This is potentially due to the datasets that were used as well as the resources used to train the models.

Table 2 shows the Median MER for the new models on different datasets. It can be seen that the newly fine-tuned models obtained a lower median MER on the JL and Mansfield corpus, especially the models trained on the Mansfield data only and the one trained on the data excluding the JL corpus. The increase in performance can be seen in the density plot where the density of low MER values is greater for the newly trained models.

Table 2. Median MER for the new models on different datasets

	baseline	**all**	**MAN only**	**MZ only**	**no JL**
Mozilla	0.2217	0.2245	0.223	0.2245	0.2245
Mansfield	0.0695	0.0684	0.0689	0.0683	0.0679
JL	0.1477	0.1464	0.1437	0.1449	0.1445

When testing the different models, the ASR systems consistently performed the best on the Mansfield corpus. This can be seen in the Appendix A. Investigating this, several reasons are proposed. Firstly, the audio quality of the Mansfield corpus was found to be better on average than the Mozilla Corpus. In addition, many of the samples of speech used in the Mozilla Corpus contained proper names which could have affected the results negatively. With regard the JL corpus, it contained many samples of the same speaker repeating the same lines with different emotions so this could have also negatively affected the results. This could explain why the median MER differed so significantly across the three datasets.

5.2 Responses to the research questions

Ultimately, this report has determined the accuracy of several ASR systems on a variety of datasets of NZ English. However, it is advisable to refrain from using the results as a general representation of the performance of ASR on NZ English as the datasets are not a representative sample but rather a collection of data that was accessible. The performance of ASR systems varied greatly across different datasets and models so any results from this study should be considered in the scope that the recognition was performed in.

Match error rate was identified as a suitable measure of error in ASR as it is similar to WER in non-extreme cases but does not lead to the issues of excessively high scores that can happen in edge cases of WER. As the MER is restricted to between 0 and 1 it is also easier to make graphical comparisons with compared to WER which is uncapped. The MER is also less computationally expensive than WIL which is useful in an environment when a large number of calculations must be made.

It has been previously identified that commercial ASR systems lag behind in their performance on NZ English compared to US English. In this project it has been found that major commercial ASR models for NZ English perform more poorly on NZ English when compared to the US model. Experimentally the performance of an ASR model on NZ English was improved by fine tuning it with NZ English datasets. The fine tuning of existing non-NZ English models with NZ English data may be a promising way forward for the improvement of ASR performance on NZ English.

In terms of improving performance, an improvement on a pretrained non-NZ English model was made but even after fine tuning the new model performed worse than commercially available models. Of particular note was which models generated the greatest performance. The models trained on no JL and only Mansfield netted the greatest increase as compared to the baseline model. This suggests that the JL corpus was not suited for usage in fine tuning ASR models, possibly due to the large number of similar lines.

5.3 Research limitations

The project was constrained by limited access to datasets, where only open source and the Mansfield corpus were available to use as well as the limited time and computational resources. The approach taken in this project was influenced by this, the decision was made to adapt an already existing system rather than create a new one as it was believed that there were not enough resources to do so. As this study adapts an already existing model, the ability to change it was limited and as a result the improvement in performance was also limited. Another limitation of the project is the lack of original datasets. The quality of the datasets used could not be controlled easily as there was not enough time or expertise to collect original data for this study.

6. Conclusions

The current state of accuracy of ASR systems on NZ English historically has been worse than that of American English. Investigating in more detail, we found that the performance of some ASR systems is worse when using the NZ model as compared to the US model. We found an improvement in the average MER on NZ English of the speech brain wav2vec model by fine-tuning it on Samples of New Zealand English. With a limited access to datasets, time and computational resources, we were unable to achieve a better performance than the currently available commercial systems. Considering the superior performance of US models of ASR on NZ English as well as the improvement in performance from fine tuning, fine tuning existing models can be considered a viable method for improving the quality of ASR systems on NZ English. MER was identified and used as a good measure for the level of error for an ASR on a sample and was used to gage the performance of the different ASR models.

7. Future Work

In the future more in-depth research could be conducted on this topic, with a greater investment in collecting datasets as a potential avenue to explore. The highest quality dataset in terms of median MER also produced the model with the best performance when used to fine tune. The collection of a high-quality dataset, representative of NZ English, could be used to adapt existing non-NZ models or potentially form the basis of a wholly NZ English ASR. Research into what constitutes a high-quality dataset for ASR training is another potential avenue for future work.

References

[1] M. Malik, M.K., Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey. *Multimedia Tools and Applications,* 80(6), pp.9411-9457, 2021,”

[2] E. Trentin, and M. Gori, “A survey of hybrid ANN/HMM models for automatic speech recognition,” *Neurocomputing,* 37(1-4), pp.91-126, 2001.

[3] J. Padmanabhan, and M.J. Premkumar, Machine learning in automatic speech recognition: A survey. IETE Technical Review, 32(4), pp.240-251, 2015.

[4] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, L., “The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods,” In proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6918-6922, June, 2021.

[5] V. Këpuska, and G. Bohouta, “Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx),” *Int. J. Eng. Res. Appl,* 7(03), pp.20-24, 2017.

[6] R. Matarneh, S. Maksymova, V. Lyashenko, and N. Belova, “Speech recognition systems: A comparative review”, 2017.

[7] A. Jain, V.P. Singh, and S.P. Rath, “A Multi-Accent Acoustic Model Using Mixture of Experts for Speech Recognition,” in Proc. Interspeech, pp.779-783, 2019.

[8] C. Ike, S. Polsley, and T. Hammond, “Inequity in Popular Speech Recognition Systems for Accented English Speech,” in proceedings of 27th International Conference on Intelligent User Interfaces, pp. 66-68, Mar 22, 2022.

[9] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, and S. Goel, “Racial disparities in automated speech recognition,” in Proceedings of the National Academy of Sciences, 117(14), pp.7684-7689, 2020.

[10] M. Najafian, and M. Russell, “Automatic accent identification as an analytical tool for accent robust automatic speech recognition,” *Speech Communication*, 122, pp.44-55, 2020.

[11] M.Y. Tachbelie, S.T. Abate, and T. Schultz, “Multilingual speech recognition for GlobalPhone languages,” *Speech Communication,* vol. 140, pp. 71–86, 2022.

[12] T. Viglino, P. Motlicek, and M. Cernak, “End-to end accented speech recognition,” in Proceedings of Interspeech, pp. 2140–2144, 2019.

[13] A. Jain, M. Upreti, and P. Jyothi, “Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning,” in Proceedings of Interspeech, pp. 2454-2458, 2018.

[14] B. Oyo, BM. Kalema, “A preliminary speech learning tool for improvement of African English accents,” in proceedings of International Conference on Education Technologies and Computers (ICETC), pp. 44-48, Sep 22, 2014.

[15] W. Ying, L. Zhang, and H. Deng, “Sichuan dialect speech recognition with deep LSTM network,” *Frontiers of Computer Science,* 14(2), pp.378-387, 2020.

[16] R. Tatman, “Gender and dialect bias in YouTube’s automatic captions,” in Proceedings of the first ACL workshop on ethics in natural language processing, pp. 53-59, April 2017.

[17] AB. Wassink, C. Gansen, and I. Bartholomew, “Uneven success: automatic speech recognition and ethnicity-related dialects,” *Speech Communication*. Mar 12, 2022.

[18] A. Jain, M. Upreti, and P. Jyothi, “Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning,” in Proceedings of Interspeech, pp. 2454-2458, 2018.

[19] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, “Joint modeling of accents and acoustics for multi-accent speech recognition,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5.

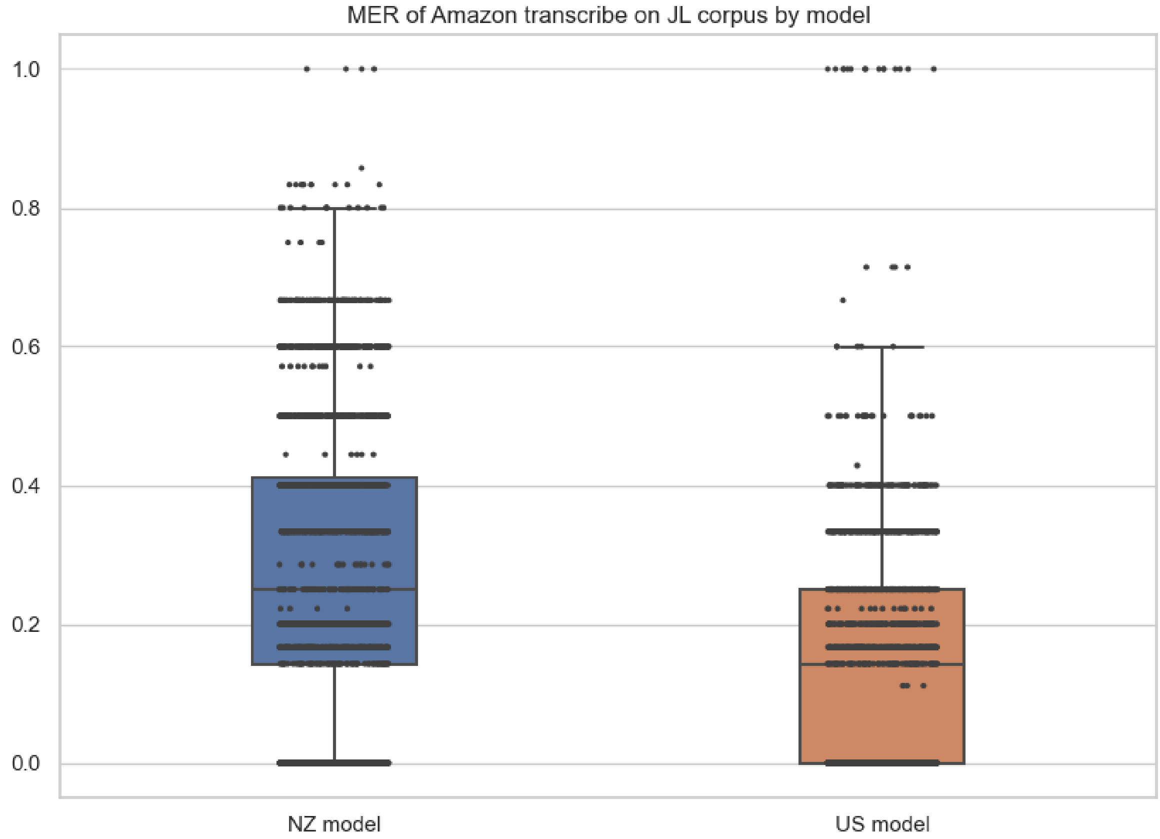
[20] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, “Data Augmentation Improves Recognition of Foreign Accented Speech,” Proceedings of Interspeech, No. September, pp. 2409-2413, 2018.

[21] S. Ghorbani, J.H.L. Hansen, “Leveraging Native Language Information for Improved Accented Speech Recognition,” in Proceedings of Interspeech 2018, pp. 2449- 2453.

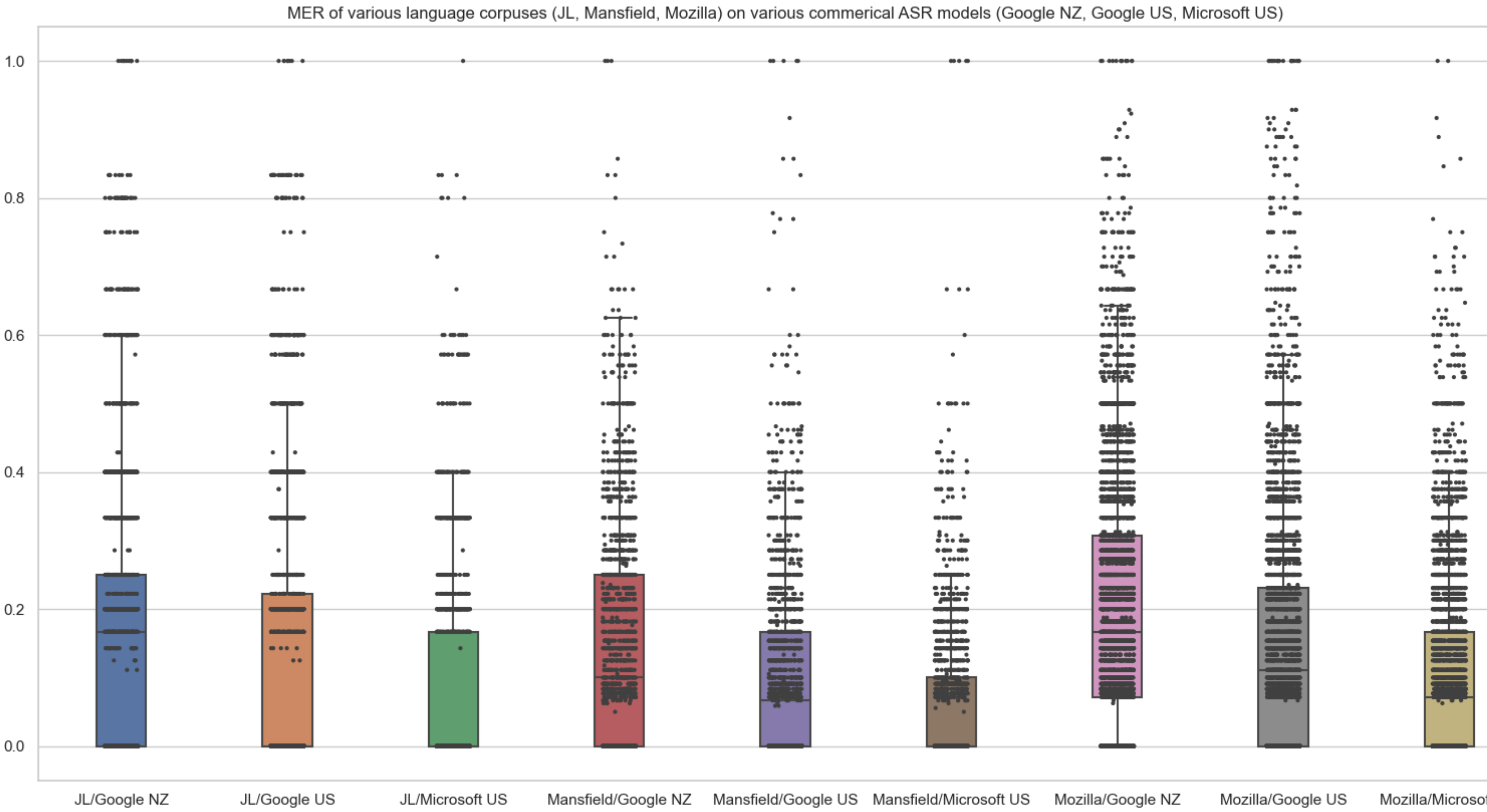
[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, 33, pp.12449-12460, 2020.

[23] A.C. Morris, V. Maier and P. Green, “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition,” in Eighth International Conference on Spoken Language Processing, 2004.

Appendix A. MER of Amazon Transcribe on JL corpus by model



Appendix B. MER of various language corpuses


MER of various language corpuses (JL, Mansfield, Mozilla) on various commerical ASR models (Google NZ, Google US, Microsoft US)
1.0
0.8
0.6
0.4
0.2
0.0
JL/Google NZ
JL/Google US
JL/Microsoft US
Mansfield/Google NZ
Mansfield/Google US
Mansfield/Microsoft US
Mozilla/Google NZ
Mozilla/Google US
Mozilla/Microsof