Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Final Report

Project Number: 18

Addressing Accent Bias in Speech Recognition Systems

Louis Chuo

Henry An

Supervisors: Jesin James, Josh Bensemann

11/10/2022

Declaration of Originality

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

Louis Chuo

Name: Louis Chuo

Abstract

Accents introduce skews in the statistical learning of speech recognition systems, leading to systems performing worse overall, often referred to as accent bias. This report analyses the current performance of three commercial automatic speech recognition systems, Google Speech-to-Text, Microsoft Azure, and Amazon Transcribe, against three New Zealand English datasets, JL Corpus, Mansfield Corpus, and Mozilla Common Voice Corpus. Furthermore, a finetuned automatic speech recognition system is also developed using the three New Zealand English datasets to address accent bias. The produced pipelines, methodology, and experimental setup are outlined for these experiments. Results show a 1.6% match error rate increase for the best-finetuned model, showing the possible improvement possible with the current datasets available. Commercial New Zealand English models performed similarly or worse than their United States English counterparts. As such, a conclusion is made that further improvement is possible for improving the performance of automatic speech recognition systems on the New Zealand accent. This can be done by developing and improving upon the currently available datasets by gathering audio data that is a good representation of the New Zealand English accent.

Acknowledgements

We would like to thank Jesin James and Josh Bensemann for providing excellent support as supervisors during the project and helping us immensely to progress the experiments.

We would also like to thank Catherine Watson for allowing us to access the Mansfield dataset to aid our research.

Table of Contents

List of Figures and Tables, Formulas and Acronyms

Tables

Figures

Acronyms

ASR – Automatic Speech Recognition

AID – Automatic Accent Identification

WER - Word Error Rate

MER - Match Error Rate

WIL – Word Information Lost

API – Application Programming Interface

CRDNN – Convolutional, Recurrent, and Deep Neural Network

CTC - Connectionist Temporal Classification

RNNLM – Recurrent Neural Network Language Modelling Toolkit

1. Introduction

Current societies have increasingly relied on speech recognition technology for tasks such as speech-to-text applications and virtual assistants such as Apple's Siri. Artificial intelligence is used to sample voice audio and process them into machine-understandable information to make use of [1]. Due to human nature, there will be variations in the pronunciations and speech patterns of different people, commonly referred to as accents [2]. An example of accents would be the pronunciation of certain words in New Zealand (NZ) versus United States (US) residents. Another example are accents on a regional scale, with southern US accents such as Texan being distinct from New York or Californian accents. These accents can introduce statistical skews and inaccuracies that can lead to worse speech recognition system performance, often referred to as accent biases [3]. Research has been and is being done to accommodate this and allow for more accurate speech recognition systems. As NZ accented English is underrepresented, it was elected to focus on New Zealand accented English as the main accent for developing the project. Moreover, the research itself should be NZ based, furthering the motivation.

The research questions addressed in this research are:

1. What is the accuracy of current speech recognition systems on the NZ accented speech?
2. What is an appropriate quantitative measure for identifying errors in current speech recognition systems?
3. How can ASR systems' accuracy be improved regarding recognition of the NZ English accented speech?

The remainder of the report is structured as follows. The following subsection outlines the general definition of a speech recognition system and the most common error metric. Section 2 outlines the literature review on the current solutions for addressing accent bias. Section 3 describes the metrics datasets and automatic speech recognition (ASR) systems chosen for the project and an overview of the methodology. Section 4 describes the experimental setup and results attained for commercial and finetuned ASR models. Section 5 discusses and justifies the results. The conclusions follow this in Section 6.

1.1 Automatic Speech Recognition Definition

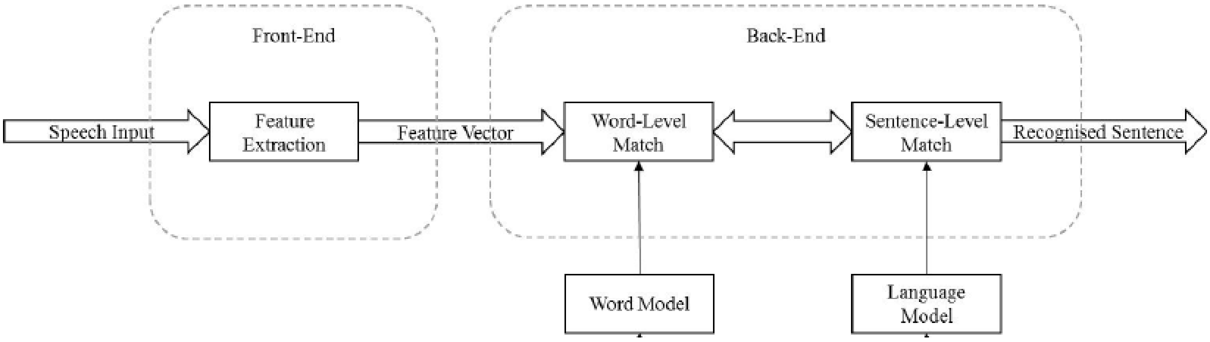

Fig. 1. General model of an ASR system [4]

ASR systems are neural network models that work by converting audio files to statistical graphs, called embeddings, that can be quantitively evaluated, called feature extraction. These values are then matched, both per word and per sentence, by the model. Training a model uses datasets that contain both the audio file and its accurate transcription, matching the extracted embeddings to the transcription for the model to "learn ." Testing models receive audio files that the model will try and infer words based on its learning. Finetuning is when an already pretrained model is then trained again with additional sets of data. This helps the model to adapt to the new dataset while also not having to relearn embeddings directly from scratch, allowing for a faster learning rate.

The most used metric for evaluating the performance of a model, word error rate (WER) (Appendix A), compares the correctly and incorrectly inferenced words and calculates a ratio of that. As such, the lower the WER score is, the better.

2. Literature Review

2.1 Current Systems and their performance

Research has been done to determine if accent affects the performance of ASR systems. Huang et al. [5] found that accents and gender are the most important to consider of all the variabilities in speech. Further research was done to determine the exact effects and the initial steps to solve these biases. Huang et al. [6] showed that current speech recognition systems showed a 40-50% error increase when trying to classify an accent with a model trained with a non-matching accent. Moreover, they have also determined two main methods for addressing these biases, pronunciation adaptation and automatic accent identification. Pronunciation adaptation is where the ASR system adapts precisely to the pronunciation of the speech, tracking a user's speaking speed and style and modifying the system to account for these based on the standard US English accent. Automatic accent identification classifies different accents based on their features and matches pre-trained accent models so that accent bias is minimized. This can be a problem on more generalized ASR systems that would want to consider all accents, as the number of accents that can be considered is determined by the availability and quality of the databases available for each accent.

Wassink et al. [7] researched more phonetic variations in languages, specifically on a multi-ethnic sample of Pacific-northwest American English accents. They utilize socio-phonetic methods to identify and classify their dataset's different vocalizations, allowing a more accurate representation of how each accent diverges from the standard American English accent. They then used Microsoft's speech recognition software to test the error rates to determine the exact errors produced while translating accurately. They determined that around 20% of the errors are caused by these specific socio-phonetic features, such as mispronunciations and devoicings. Moreover, most of the errors detected were from speech inputs that use conversational language, compared to lexical tasks and reading passages. This implies that pronunciation is not the only root cause of word errors but also that speaking styles can have a significant impact. These systems can be leveraged, specifically for fine phonetic details, to improve speech recognition accuracy further.

Current commercial speech recognition systems do not yet account for accent bias. Koeneke et al. [8] examined five state-of-the-art ASR systems: Apple [9], Amazon [10], Google [11], IBM [12], and Microsoft [13], using the Corpus of Regional African American Language (CORAAL) [14] to see if there is specifically racial bias on these systems. CORAAL is an open-source corpus with regional varieties of African American speech and around 27 GB of audio files. Racial bias can be equated to accent bias as both similarly have socio-phonetic features that are divergent from the standard data that commercial speech recognition systems have. They have found that the average word error rate for black American speech is considerably worse than for white American speech, consistently through all the five ASRs they used, showing racial bias toward white American speech. Although ASR systems are trained to translate with a vast vocabulary, there is a chance that they mistranslated slang or uniquely structured grammar. Thus WER may not be the most accurate representation of accented speech recognition.

Tatman [15] also did similar research, using YouTube's automatic captions and evaluating their accuracy regarding the video's tagged country as a reference for their accent. YouTube's automatic captions use Google's ASR system, and YouTube has a large user base making it ideal for testing. Accents from California, Georgia, New England, New Zealand, and Scotland were sampled, and the videos that were sampled were on the topic of the "accent challenge," where people were to recite a list of words as a "challenge" to see the effects of various accents. This meant that ASRs could not use language models that try to contextualize sentences, making sure that the system would only recognize the words. The results in accent WER showed that Californian English had lower overall WER than the other accents, further solidifying that current commercial ASRs are biased towards the Californian American accent. More importantly, they also found that the New Zealand English accent has an unusually high confidence interval compared to the others, suggesting that New Zealand English has a high degree of variability.

2.2 Current Trends and Solutions

Current research is being done into potentially combining accent recognition with improving the recognition itself by identifying and accounting for the deviations from the standardized US English accent. Multiple research papers have been published [16, 17] that follow this logic, using two separate neural network models to allow for flexible speech bias learning while also being able to do accent recognition. Oyo & Kalema [16] used African English as the base dataset to develop a dual speech recognition engine. This works by first using a speech classifier engine to segregate the different accents; then, a second recognizer is used to classify the differences between the current and standard English accents. This difference is then used in a measure that can be used to provide feedback to the person on the utterances that have the most deviation from the standard accent. Quantifying the difference between current accents and the standard US English accent can be extremely helpful in allowing a multi-accent ASR system to recognize accents dynamically without having to be explicitly pretrained.

Jain et al. [17] expand further on this research. A similar dual-model framework is used, with the first model being an accent classifier and phone recognition to translate speech to text. This is trained in conjunction with another model that classifies the differences between the detected accent to the standard US English accent. This then produces accent embeddings that can be used as auxiliary inputs to the phone recognition model, allowing for accented speech to be transcribed more accurately. Moreover, they also used a different, more robust dataset, the

Mozilla Common Voice Corpus [18], which compiles over 14,000 validated hours of different audio files. It also has various accents so that accent recognition can be more robust. The research uses the most well-represented accents in the datasets and the model to try and detect accents that are not used to train the model, labeled as "unseen" accents. Using the accent embeddings on the model achieved a 15% overall WER reduction on the trained accents and a 10% WER reduction on the untrained accents, showing promise in the field to develop further and improve the system.

2.3 Summary and Potential Research Avenues

It was shown that the New Zealand accent has a high degree of variability compared to other English accents, and there is very little research done to classify and solve this issue; this may be a potential avenue for research. There is also potential research on combining the current multi-framework model being researched with the socio-phonetic analysis to improve upon the accuracy of accent translation further, allowing for a clearer definition of the differences in untrained languages to detect these better. Lastly, word error rate, which is the most common measure for ASR accuracy, has been shown to potentially not be the most accurate specifically on quantifying errors from accent purely. As such, further research can be conducted to determine if there is a better measure.

3. Research Methods

3.1 Error Metrics

The main issue with the WER calculation is that the ratio produced counts the errors over the initial baseline words. This results in WER scores potentially being higher than a value of 1, in which it was found that this results in a less accurate representation of the errors of words.

TABLE 1: WER, MER, and WIL Score Comparison of Three Example Sentences

Baseline Sentence	ASR Hypothesis	WER	MER	WIL

“Markazi lies in western Iran”	“Mark as he lies in. Whiston around it.”	1.2	0.75	0.9
“Avengebrol Avencebrol Avicebrol and finally Avicebron”	“Avenger bro bro”	1.0	1.0	1.0
“Lights flickered as the last remaining fuel drained from the generator”	“let's flicking is the last remaining fuel drain from the generator”	0.3636	0.3636	0.595

As shown in Table 1, The first sentence set has a worse WER score than the second set, even if it has more correct individual words. The first two examples are on the extreme case of the present errors but need to be addressed to have a more appropriate representation of the calculated errors.

Morris et al. [19] explored alternative metrics for addressing the problem, proposing two metrics as alternatives, Match Error Rate (MER) (Appendix B) and Word Information Lost (WIL) (Appendix C). As seen in Table 1, both MER and WIL give a more appropriate numbered representation for the extreme cases while still being consistent if the hypothesis is completely wrong. Moreover, on lower error cases, such as in the third sentence set, MER and WER performed almost identically while WIL gave different values. As most papers use WER, it was deemed that MER is the most appropriate metric as it strikes a balance between more appropriately representing the extreme cases while also performing similarly enough on normal cases that it can still be compared to the results of other papers.

3.2 New Zealand English datasets

Datasets were needed to represent the NZ accented English for the general purpose of the project. As such, three different datasets containing NZ accented English were acquired: JL Corpus, Mansfield Corpus, and Mozilla Common Voice Corpus. A comparison of the datasets is shown in Table 2.

TABLE 2: Comparison of NZ Accented English datasets

Database Name	Is Open Source?	Sentence Number	Average Sentence Length	Metadata	File Type	Sentence Type	Baseline Format
JL Corpus [20]	Yes	2,400	5.39	Emotion, Gender	.wav	Set of sentences, repeated per emotion	Separate text files
Mozilla Common Voice [18]	Yes	4,366	10.411	Gender, Age, Accent	.mp3	Unique sentences sampled from internet	CSV
Mansfield Corpus [21]	No	1,863	9.78	2 Females, 2 Males	.wav	Unique sentences from NZ literature	CSV

The JL Corpus [20] was developed primarily for emotional speech recognition. However, it is spoken by native NZ English speakers. Mansfield Corpus [21] was made specifically to represent NZ accented English, making it appropriate for the project. It, however, is not open-sourced and only has four speakers that it represents. Finally, the Mozilla Common Voice Corpus [19] includes multiple accent representations. Still, all the files are labeled based on accent; thus, the appropriate files for NZ accented English will need to be extracted. The corpus itself has the highest average sentence length by word and the greatest number of lines available, making it the most varied dataset out of all the ones explored. As all the explored datasets have unique properties, all three should be used for the project in conjunction to allow for the best representation of NZ accented English.

3.3 Commercial Systems

The different commercial ASRs explored by [8] was used as a basis for the project's chosen commercial ASRs. Three of the five listed were available for public testing and use. A comparison between the ASRs is shown in Table 3.

TABLE 3: Comparison of Different Commercial ASRs

Commercial ASR	**Free Credits**	**Pricing**	**Has NZ Model**	**Local File Support**	**Native mp3 file support**
Microsoft Azure Speech-to-text [13]	US $200	$1 per hour	Yes	Yes	No
Google Speech-to-text [11]	US $300	$0.006 per 15 seconds	Yes	Yes	Yes
Amazon Transcribe [10]	None	$0.024 per minute	Yes	No	Yes

All three Commercial ASRs have NZ models available, so a comparison between the NZ and US English models can be made. Unlike the other two, Amazon does not have local file support or free credits available. As these services require payment, results on the Amazon Transcribe ASR will be limited. All three systems will be tested to have a better understanding of the performance of commercial ASR systems.

3.4 Open-sourced Systems

TABLE 4: Comparison of Different Open-Source ASR APIs

Toolkit	**Base Framework**	**Main Advantages**	**Main Disadvantages**
Vosk [22]	Kaldi	Portability, ease of use	Hard to customise, not based in PyTorch
SpeechBrain [23]	PyTorch	Flexibility, ease of use, customisable	Documentation not comprehensive, actively being developed
FairSeq [24]	PyTorch	Comprehensive, well-documented	Complicated, hard to set up, hard to customise

Different open-sourced APIs were also researched for developing and building an open-sourced model. Three were chosen as the best-fitting API for the project, Vosk [22], SpeechBrain [23], and FairSeq [24]. Table 4 compares the different APIs. SpeechBrain was chosen as the most appropriate API for the project, as it allows for the flexibility of developing and finetuning different ASR models versus the other options. Moreover, it was found to be the easiest to develop, including tools that make ASR development tasks faster, such as an integrated dataset loading system and automatic audio normalization.

3.5 SpeechBrain

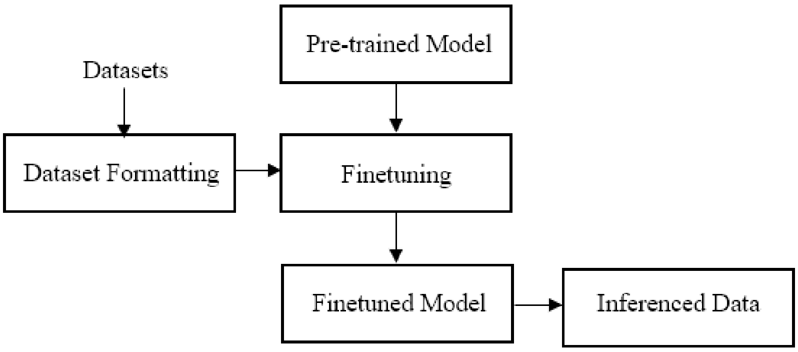

Fig. 2. General Structure for SpeechBrain finetuning

Speechbrain was mainly used for finetuning a model for the project. Figure 2 outlines the general structure of finetuning a pretrained model in SpeechBrain. The dataset needs to be formatted using the DynamicItemDataset class [25] so that it can match up with its accurate transcriptions. Speechbrain itself has a finetuning Brain class [26] that inputs the formatted data and a pre-trained model available from their repositories as inputs and outputs a finetuned model. This model can then be used directly for testing and inferencing using its EncoderDecoder class [27].

3.6 Methodology

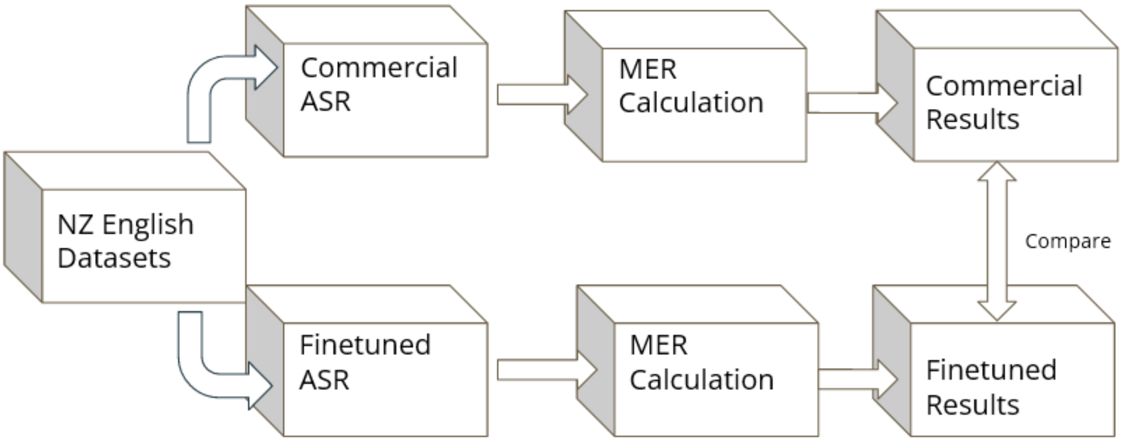

Fig. 3. Flow chart plan of the project

Figure 1 displays the general plan for the experiments for the project. It outlines the general steps that will be taken to produce the results that will be needed for answering the research questions. The plan is to use the chosen commercial ASRs and calculate their MER scores. The same steps will be done for the finetuned ASRs, then a comparison of the results will be made to discuss and draw a conclusion. Finetuning a pre-trained model instead of training a new model was chosen due to time constraints. Finetuning would take less time to process, as is the nature of it, versus training from scratch while still garnering a similar effect on the model as training it from scratch.

4. Experiments and Results

4.1 Experiment Hardware and Specifications

All experiments were done on a desktop computer with the following specifications:

Processor: AMD Ryzen 7 3700X 8-Core Processor 3.6 GHz

Installed RAM: 16 GB

GPU: NVIDIA GeForce RTX 3060 12 GB VRAM

OS: Ubuntu 20.04

All written code was developed and run in the Python language.

4.2 Dataset Preprocessing

The Mozilla Common Voice Corpus contains a variety of accents and is in the .mp3 audio file format. As such, preprocessing steps are needed to allow the dataset to contain only the NZ accent files in .wav format. The CSV file present for the corpus containing all the labels is imported to Microsoft Excel to take advantage of Excel's label filter.

This filtered data is then fed into a file sorter processing code to be sorted in a new file path containing only the files labeled to have the NZ accent. The python package, Pydub [28], is then used to convert the .mp3 files to .wav.

4.3 Commercial ASR testing

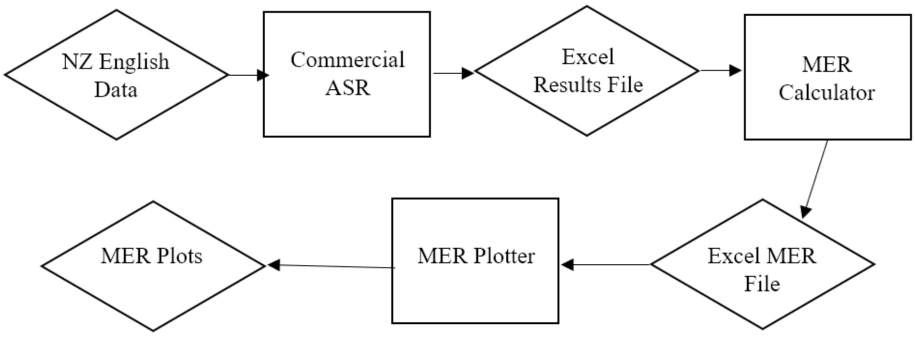

Fig. 4. Flow chart for commercial ASR processing pipeline

A pipeline was set up to streamline multiple datasets being tested on multiple different commercial ASR models, as seen in Figure 3. All the boxes represent the processing code developed, and all the diamonds represent the various data saved in Excel sheets that are produced for the pipeline.

Per the Mozilla and Microsoft ASR, code was developed that leveraged the provided API to allow for sending the local audio files to their respective cloud services. These were then used as functions within a loop to go through each file in their respective dataset, as all the audio files are in one folder per dataset. These results were then saved in a CSV file, including their respective audio file names. They were then imported to Excel and combined using different Excel sheet pages for further processing. Each API call per audio file took around 3 seconds to process, as such, to process all the data on one model:

$$2{,}400 + 4{,}366 + 1{,}863 = 8629 \ \times 3 = 25{,}887 \ seconds \ or \ 7.2 \ hours$$

Four models were tested (Microsoft NZ and US, Google NZ and US), and the total processing time took around 29 hours.

For the Amazon ASR, as there were limitations per the cost and the Mansfield corpus being closed source, only the JL Corpus was uploaded to their cloud services, and thus only the JL Corpus was tested on Amazon. Unlike the other ASR systems, Amazon had no local file support and could only be used if the audio files were uploaded to their cloud servers. The code was for API calls to transcribe all the audio files in the cloud and save the results locally.

Code was then developed to compare the generated results to the baseline by matching up their filenames to the provided baseline files, matching up the filenames, then using the python package JiWER [29] to calculate the MER score. These are then saved in another CSV file and compiled in Excel for plotting and evaluation.

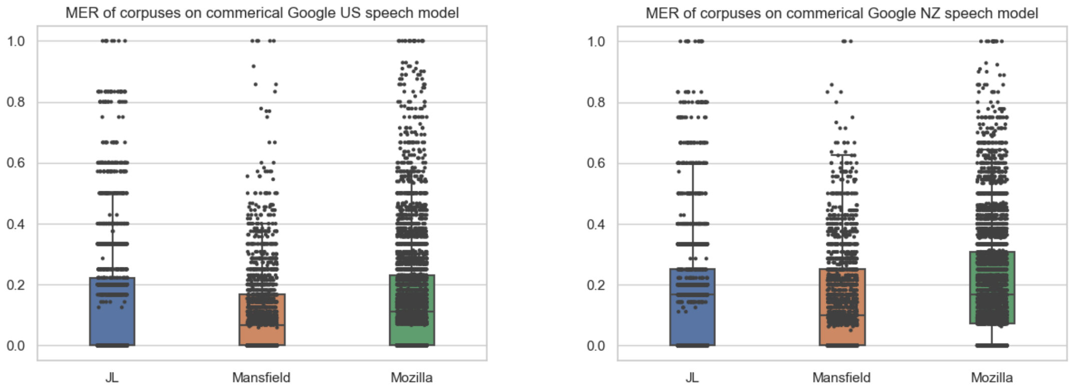

Fig. 5. MER for NZ Corpuses on Google US vs NZ Model

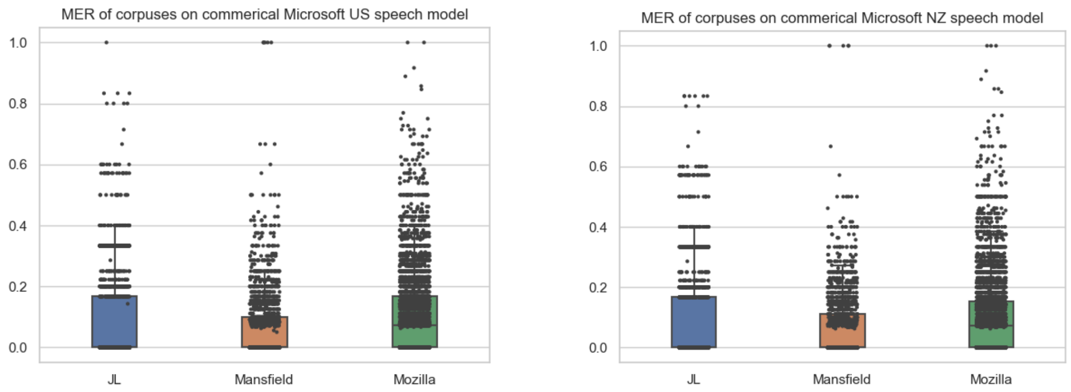

Fig. 6. MER for NZ Corpuses on Microsoft US vs NZ Model

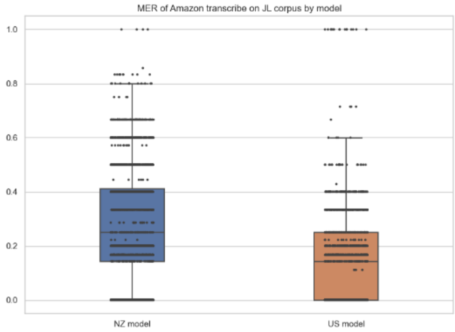

Fig. 7. MER for JL Corpus on Amazon US vs NZ Model

4.4 Dataset Embeddings

To further explore the effects of the NZ English accent on ASR models, the accent embeddings were plotted versus US English.

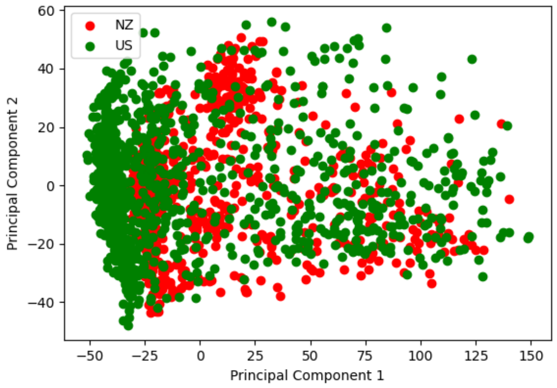

Fig. 8. Accent Embeddings of NZ vs US English using the Mozilla Corpus

As the Mozilla Common Voice Corpus contains multiple accents, it was used as the dataset for comparing accent embeddings. This was achieved by extracting the US accented audio files from the corpus, similar to how the NZ accent files were extracted. Then, a random set of 200 audio files was sampled from each accented dataset and fed into SpeechBrain’s VoxLingua107 spoken language identification model [30] to produce the required accent embeddings and plotted.

4.5 Splitting Datasets

The datasets are mixed and split into 75% training, 5% validation, and 20% testing, saved in sorted CSV files containing the audio file names and their respective baseline texts. This is needed as training data cannot be the same as validation and testing data for finetuning a model, as it will promote model overfitting.

4.6 Finetuning

SpeechBrain’s DynamicItemDataset feature [25] was used to simplify the loading of data for processing. As such, code was developed for reformatting the CSV files to allow for proper data loading in the framework. The audio normalization, as well as the pipeline for tokenizing the baseline words, were done. Allowing the dataset to be run directly into finetuning an ASR model.

The chosen pre-trained ASR model for finetuning is a CRDNN with CTC/Attention and RNNLM trained on the Librispeech Corpus [31]. The model was chosen as it is trained on a different dataset than the project's, allowing the finetuning to gain a more distinct result. Moreover, using both CRDNN and RNNLM allows for more future work to be done as they both are heavily customizable for improving the model.

Different combinations of the datasets were explored, detailed in Table 5, for finetuning the model to see the effectiveness of each dataset in representing the NZ accented English.

TABLE 5: Different Finetuned Models and datasets used

Model Name	Model_All	Model_MAN_only	Model_MZ_only	Model_no_JL
Datasets Used	Mozilla, Mansfield, JL	Mansfield	Mozilla	Mansfield, Mozilla

Separate sets of data were produced by filtering out the training list and then were sued for finetuning the CRDNN model.

Validation was also done per each training epoch, testing the average MER of the validation set to see if the dataset has started overfitting and, thus, being less accurate. Only one epoch was needed to fully train the model, as each epoch resulted in worse validation results. Each epoch took 8 minutes, with the validation taking 2 minutes; thus, it took 40 minutes to train all the models.

4.7 Finetuned Data Testing

Much like the Commercial ASR testing, a similar pipeline was established by testing the finetuned data, only switching the commercial ASR processing code with the finetuned processing code and combining it with the WER calculation. Moreover, as the testing results were extremely similar, an alternative graphing solution, density plots, was used to represent the data's differences better.

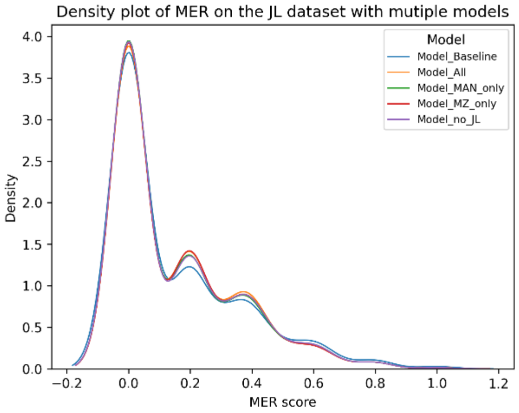

Fig. 9. Density plot of MER on JL dataset

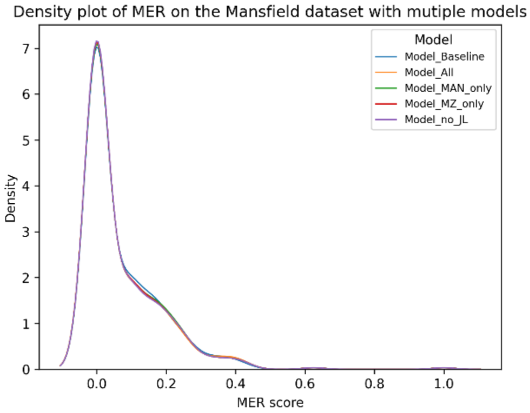

Fig. 10. Density plot of MER on Mansfield dataset

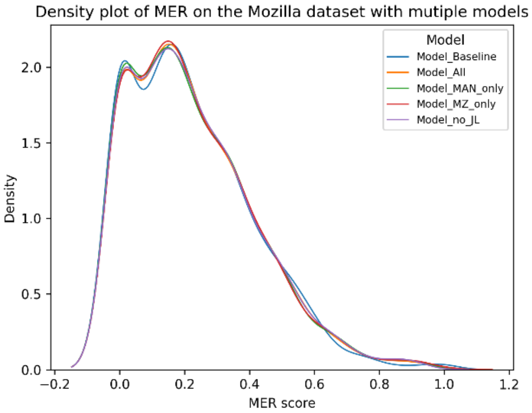

Fig. 11. Density plot of MER on Mozilla dataset

5. Discussion

5.1 Commercial Models

Trends from Figures 5, 6, and 7 suggest that Mansfield gets the best MER score, followed by JL and lastly Mozilla. This is supported by the average MER values in Table 6.

TABLE 6: Average MER of different commercial models

	Google US	Google NZ	Microsoft US	Microsoft NZ	Amazon US	Amazon NZ
JL	0.146	0.165	0.078	0.08	0.18	0.23
Mansfield	0.102	0.144	0.064	0.0622	-	-
Mozilla	0.154	0.205	0.1024	0.098	-	-
Average	0.134	0.171	0.0815	0.801	-	-

Interestingly, for both Google and Amazon, their NZ English models performed worse than their US models, which is unexpected, given that the test dataset is in NZ English. Microsoft's models performed very similarly to each other, and overall is the best performing commercial model. As shown in the embeddings in Figure 6, there should be a noticeable difference in NZ versus US English accents. So, to have the NZ models of the commercial ASR models perform similarly or worse than their US models is an unexpected result. A potential reason for the results is that the data used for training the commercial models on NZ English is insufficient or of low quality, such as being noisy or not well enunciated. This helps explain why it would perform worse, as bad training data usually leads to worse ASR results.

5.2 Finetuned Models

TABLE 7: Average MER of different finetuned models

	Baseline	All	MAN only	MZ only	No JL
JL	0.1477	0.1464	0.1437	0.1449	0.1445
Mansfield	0.0695	0.0684	0.0689	0.0683	0.0679
Mozilla	0.2217	0.2245	0.223	0.2245	0.2245
Average	0.1463	0.1464	0.1452	0.1459	0.1456

Table 7 shows the average MER for each tested MER model. Surprisingly, the model trained to include all the datasets performed worse overall than the baseline data, where it was not finetuned. Overall, the model trained with only the Mansfield dataset performed the best, suggesting that the Mozilla and JL dataset may not be the best representation of NZ accented English.

All finetuned models performed worse overall on the Mozilla dataset. However, the density plot from Figure 11 shows a higher peak of MER scores on the lower end while also having higher MER score peaks on the higher end. This suggests that finetuning the model allowed for the better MER scoring sentences to be transcribed but also worsened the performance of sentences with higher MER scores.

The presence of the JL and Mozilla datasets makes the finetuned model perform worse. This is most likely as the JL corpus is focused on emotional language; thus, it will affect the accuracy of the NZ English representation. Likewise, the Mozilla dataset is very noisy and full of jargon compared to the Mansfield, meaning there will be a lot more variance in the data.

5.3 Commercial Versus Finetuned

The finetuned models performed worse overall than the commercial models, potentially due to the commercial models being better as more research and work has been done for them compared to the open-source models. As the finetuned models performed at best 1.6% better compared to the baseline on the JL corpus, it can be concluded that the datasets can be improved to better represent NZ accented English, especially since both the JL and Mozilla datasets on the training data worsened the results. This also supports why the commercial NZ models performed worse, as this implies that the datasets that were used for training the commercial models may not be like the Mozilla and JL datasets and thus has the potential for improvement to better represent NZ accented English.

5.4 Limitations

As the commercial models are proprietary, exploring them and the datasets used for their training is not currently possible. As such, only assumptions can be made as to the commercial ASR results achieved. Moreover, the Mansfield dataset is closed-sourced, limiting its usage for testing on commercial datasets.

5.5 Research Questions Met

Research into identifying MER as the most appropriate measure satisfies research question 2, the results obtained from testing the ASR systems satisfy research question 1, and the conclusions drawn will satisfy research question 3 as to what improvements need to be made.

6. Conclusions

6.1 Technical Summary

The report outlines the processes developed for finetuning a pre-trained ASR model on the SpeechBrain framework using three different NZ English-based datasets; the JL, Mansfield, and Mozilla Common Voice corpus. Processes were also developed for testing three commercial ASR systems; Google Text-to-speech, Microsoft Azure, and Amazon Transcribe. Moreover, MER is also determined to be the most appropriate quantitative measure for evaluating the performance of ASR models.

6.2 Definitive Conclusion

Finetuning a model using the Mansfield dataset showed a 1.6% increase versus baseline MER performance, which is lower than expected. Moreover, the presence of the JL and Mozilla datasets in the training data resulted in worse results, most likely due to the undesirable aspects of both datasets, including emotional and noisy data. Thus, it can be concluded that improvements could be made with the current NZ accented English databases to allow for better ASR performance, such as gathering more comprehensive data and increasing its quality by filtering out unwanted noise.

6.3 Future Work

As per the conclusion, a potential future work is the compilation of a more comprehensive dataset to represent the NZ English accent. The Mansfield dataset, which had the best performance, could be expanded to include a wider variety of speakers following the same format. Furthermore, further testing can be done on finetuned models, using different pre-trained models, or even potentially building a new model that can be trained purely on an NZ English dataset to see any performance increases versus finetuning.

References

[1] S. Alharbi κ.ά., ‘Automatic speech recognition: Systematic literature review’, IEEE Access, τ. 9, σσ. 131858–131876, 2021.

[2] S. Deshpande, S. Chikkerur, και V. Govindaraju, ‘Accent classification in speech’, στο Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID’05), 2005, σσ. 139–143.

[3] M. Benzeghiba κ.ά., ‘Automatic speech recognition and speech variability: A review’, Speech communication, τ. 49, τχ. 10–11, σσ. 763–786, 2007.

[4] R. Salaja, R. Flynn, και M. Russell, ‘ALife-Based Classifier for Automatic Speech Recognition, 06 2014, τ. 679.

[5] C. Huang, T. Chen, S. Z. Li, E. Chang, και J.-L. Zhou, ‘Analysis of speaker variability’, στο INTERSPEECH, 2001, σσ. 1377–1380.

[6] C. Huang, T. Chen, και E. Chang, 'Accent issues in large vocabulary continuous speech recognition, International Journal of Speech Technology, τ. 7, τχ. 2, σσ. 141–153, 2004.

[7] A. B. Wassink, C. Gansen, και I. Bartholomew, ‘Uneven success: automatic speech recognition and ethnicity-related dialects’, Speech Communication, τ. 140, σσ. 50–70, 2022.

[8] A. Koenecke κ.ά., 'Racial disparities in automated speech recognition, Proceedings of the National Academy of Sciences, τ. 117, τχ. 14, σσ. 7684–7689, 2020.

[9] Apple. Speech Recognition API [Online]. Available: https://developer.apple.com/documentation/speech

[10] Amazon. Amazon Transcribe [Online]. Available: https://aws.amazon.com/transcribe/

[11] Google. Speech-to-Text [Online]. Available: https://cloud.google.com/speech-to-text

[12] IBM. Watson Speech to Text [Online]. Available: https://www.ibm.com/cloud/watson-speech-to-text

[13] Microsoft. Azure ASR. [Online]. Available: https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/#overview

[14] T. Kendall and C. Farrington. 2021. ‘The Corpus of Regional African American Language’. Version 2021.07. Eugene, OR: The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal.

[15] R. Tatman, ‘Gender and dialect bias in YouTube’s automatic captions’, στο Proceedings of the first ACL workshop on ethics in natural language processing, 2017, σσ. 53–59.

[16] B. Oyo και B. M. Kalema, ‘A preliminary speech learning tool for improvement of African English Accents’, στο 2014 International Conference on Education Technologies and Computers (ICETC), 2014, σσ. 44–48.

[17] A. Jain, M. Upreti, και P. Jyothi, ‘Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning’, στο Interspeech, 2018, σσ. 2454–2458.

[18] Mozilla. Common voice. [Online]. Available: https://commonvoice.mozilla.org/

[19] A. C. Morris, V. Maier, και P. Green, 'From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition, στο *Eighth International Conference on Spoken Language Processing*, 2004.

[20] J. James, L. Tian, και C. I. Watson, 'An open source emotional speech corpus for human-robot interaction applications', στο *Interspeech*, 2018, σσ. 2768–2772.

[21] C. I. Watson και A. Marchi, ‘Resources created for building New Zealand 29nglish voices’, στο *Proc. 15th Australas. Int. Conf. Speech Science and Technology*, 2014, σσ. 92–95.

[22] VOSK Offline Speech Recognition API. Alpha Cephei. [Online]. Available: https://alphacephei.com/vosk/

[23] M. Ravanelli κ.ά., ‘SpeechBrain: A General-Purpose Speech Toolkit’, arXiv [eess.AS]. 2021.

[24] M. Ott κ.ά., ‘fairseq: A Fast, Extensible Toolkit for Sequence Modeling’, στο Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

[25] SpeechBrain. speechbrain.dataio.dataset module [Online]. Available: https://speechbrain.readthedocs.io/en/latest/API/speechbrain.dataio.dataset.html

[26] SpeechBrain. speechbrain.core module [Online]. Available: https://speechbrain.readthedocs.io/en/latest/API/speechbrain.core.html

[27] SpeechBrain. speechbrain.pretrained.interfaces module [Online]. Available: https://speechbrain.readthedocs.io/en/latest/API/speechbrain.pretrained.interfaces.html

[28] J. Robert. Pydub [Online]. Available: https://github.com/jiaaro/pydub

[29] Jitsi, JiWER: Similarity measures for automatic speech recognition evaluation [Online]. Available: https://github.com/jitsi/jiwer

[30] J. Valk και T. Alumäe, 'VoxLingua107: a Dataset for Spoken Language Recognition, στο Proc. IEEE SLT Workshop, 2021.

[31] SpeechBrain, CRDNN with CTC/Attention and RNNLM trained on LibriSpeech [Online]. Available:

https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech

Appendixes

Appendix A: Equation for Calculating Word Error Rate

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

Where *I, S, and D* denote the insertions, substitutions, and deletions of the sentence hypothesized by the ASR, *N* is the total number of words present in the baseline input words. Lower values mean better performance.

Appendix B: Equation for Calculating Match Error Rate

$$MER = \frac{I + S + D}{H + S + D + I} = 1 - \frac{H}{N}$$

Where *I, S, and D* denote the insertions, substitutions, and deletions of the sentence hypothesized by the ASR, *N* is the total number of words present in the baseline input words, and *H* is the number of correct words hypothesized. Lower values mean better performance.

Appendix C: Simplified Equation for Calculating Word Information Loss

Subject to $H >> S+D+I$:

$$WIP = \frac{H}{N_1}\frac{H}{N_2}, WIL = 1 - WIP$$

Where *I, S, and D* denote the insertions, substitutions, and deletions of the sentence hypothesized by the ASR, N_1 is the total number of words present in the baseline input words, N_2 is the total number of words present in the hypothesized words, H is the number of correct words hypothesized. Lower values mean better performance.