Task 4

Comparing the mean Word Error Rate (WER) (computed using the Python package jiwer) of the transcriptions by the model in Task 2a (0.118) to the finetuned model in Task 3a on the cv-valid-dev data (0.115), the fine-tuning of the wav2vec2-large-960h-cv model on the common voice dataset has improved the performance of the model on the Common Voice valid-dev data slightly by 2.5%.

To analyze the model performance with and without finetuning in more detail, Tables 1-3 below show the breakdown of the WER for each category for accents, age, and gender respectively. When we look at the variability in WER across the different accents for both models in Table 1, both models have the most difficulty predicting Indian accents, followed by Australian and Scottish accents as compared to the other more common US, England and Canadian accents. Looking at the improvements in WER after finetuning in Table 1, US, Canadian and Scottish accents show the least improvement in WER. The percentage of speakers with a Scottish accent being less 1% could be the reason for the little improvement in WER, but for US and Canada, there is a substantial proportion of speakers with those accents (Table 1), suggesting that finetuning on these speakers does not improve the WER by much. From this, I would suggest that decreasing the proportion of speakers with US and Canadian accents and increasing the proportion of speech data with Australian and Indian accents in the finetuning dataset would help in improving the performance of the finetuned model.

Looking at Table 2 and 3, the variability of WER across age and gender is much less as compared to accent, indicating that age and gender are not as significant in affecting model performance as accent type. However, one observation is that the WER for the fifties age category is higher than the other age categories. Zooming in on the accent types of people in their fifties, 41 speakers in their fifties have an Indian accent, which is nearly half out of the total 89 speakers with an Indian accent across the entire cv-valid-dev dataset. This suggests that the high WER for the fifties age category is likely due to the large number of speakers with an Indian accent, rather than their age itself.

Table 1: WER of the model before and after finetuning for each accent type. Categories with less than 30 samples are excluded from the table.

Accent	No. of Occurrences in cv-valid-dev	Mean WER of model in 2a	Mean WER of finetuned model	Percentage (%) of data in finetuning dataset
US	637	0.084485	0.082219	15.6
England	327	0.110236	0.108008	7.8
Australia	102	0.199628	0.190361	2.2
Indian	89	0.264624	0.258191	2.2
Canada	81	0.048042	0.047925	2.1
Scotland	32	0.157862	0.154638	0.6

Table 2: WER of the model before and after finetuning for each age group. Categories with less than 30 samples are excluded from the table.

Age	No. of Occurrences in cv-valid-dev	Mean WER of model in 2a	Mean WER of finetuned model
Teens	113	0.116378	0.107766
Twenties	487	0.137620	0.134356
Thirties	345	0.101825	0.100901
Fourties	244	0.085373	0.080583
Fifties	203	0.157783	0.152247

Sixties	95	0.082487	0.083096
Seventies	37	0.112647	0.114224

Table 3: WER of the model before and after finetuning for each gender. Categories with less than 30 samples are excluded from the table.

Gender	No. of Occurrences in cv-valid-dev	Mean WER of model in 2a	Mean WER of finetuned model
Male	1135	0.115173	0.112443
Female	394	0.127258	0.123838

To improve the accuracy of finetuning model, I would propose the following steps. The first step is to determine the optimal proportion of data with speakers of various accents in the finetuning dataset. To do so, we can conduct experiments with different weights assigned to different accent types, and create a finetuning dataset by sampling the training data based on the accent type (without replacement) with the assigned weights. After finetuning the model on the dataset, evaluate the WER of transcriptions on cv-valid-dev. Based on the current proportions, we can conduct experiments by decreasing or increasing the proportions of various accent types accordingly. Table 4 shows an example of a set of experiments to determine the proportion of accent types.

Table 4: Example percentages for sampling various accent types for the finetuning dataset. The WER numbers are random and not accurate.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
US	17.6%	10%	5%	8%
Canada	2.1%	1%	1%	1%
India	2.2%	5%	7.5%	6%
Australia	2.2%	5%	7.5%	6%
…	…	…	…	…
WER on cv-valid-dev	0.115	0.101	0.107	0.097

After determining the dataset with the optimal mix of data, the next step is to tune the hyperparameters. Here, I will focus on determining the optimal batch size and learning rate, but other hyperparameters can be tuned as well. For the finetuned model in Task 3a, the batch size is 8 with gradient accumulation steps of 2 which is similar to a batch size of 16, and the learning rate is 1e-6. Based on the current hyperparameters, we can conduct experiments by decreasing or increasing the batch size or learning rate one at a time to determine the best set of hyperparameters greedily. However, the greedy approach may not always be the best as there may be interaction effects between batch size and learning rate where sometimes a large batch size requires a smaller learning rate to avoid overshooting, or a small batch size requires a larger learning rate to escape local minima. Table 5 shows an example of a set of experiments to tune the hyperparameters.

Table 5: Example hyperparameters for hyperparameter tuning experiments. The WER numbers are random and not accurate.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Batch size	16	32	64	32
Learning rate	1e-6	1e-6	1e-6	1e-7
WER on cv-valid-dev	0.115	0.109	0.113	0.105

In order to reduce the training time taken to conduct all the experiments, I would also suggest to conduct these experiments on a smaller dataset (e.g. 20,000 samples). Once the optimal proportion of data and hyperparameters are determined, we can then finetune the model on a larger dataset, and evaluate the WER of the transcriptions on cv-valid-dev.

If more data of the less common accent types are required, cv-other-train can be considered as additional training data. However, as the text labels are not verified to match the audio by at least 2 listeners, some additional steps should be taken to ensure the text and the audio match. I would suggest using the model in Task 2a to generate transcriptions and to filter out the samples with a WER higher than a certain threshold.