Observation

generated_text	generated_text_finetuned	Word_ Error_ Rate
BE CAREFUL THAT YOU PROGNOSTICATIONS SAID THE STRANGER	BE CAREFUL WHAT YOUR PROGNOSTICATIONSSAID THE STRANGER	0.5
THEN WHY SHOULD THEY BE SURPRISED WHEN THEY SEE ONE	THEN WHY SHOULD THEY BE SURPRISED WHEN THE SEE ONE	0.1
A YOUNG ARAB ALSO LOADED DOWN WITH BAGGAGE ENTERED AND GREETED THE ENGLISHMAN	A YOUNG ARABALSO LOADED DOWN WITH BAGGAGEENTEREDAND GREETED THE ENGLISHMAN	0.3846 15385

Table 1: comparison results from task 2a

In the first sentence, the fine-tuned model modifies "THAT YOU" to "WHAT YOUR," attempts a correction, although it introduces new errors like "PROGNOSTICATIONSSAID" where it combined words incorrectly. A WER of 0.5 indicates a moderate number of errors here.
In the second sentence, the correction from "THEY SEE" to "THE SEE" is incorrect, though the WER remains low at 0.1, suggesting fewer word errors overall in the sequence.

The third sentence reveals issues where the fine-tuned model combines words incorrectly ("ARABALSO" and "BAGGAGEENTEREDAND"), resulting in a WER of 0.3846, indicating significant errors.
The low WER in the second sentence shows that, despite some errors, the model performs well overall in short and simple sentences, but struggles with more complex structures, as seen in the third example.

Improving accuracy

One can increase the number of Epochs for training as it allows the model to have more steps to learn from the dataset, however this can cause overfitting if the model memorizes the training data. It is good practice to gradually increase the number of epoch and monitor the validation loss to ensure that overfitting is not occurring, this can happen when the validation loss begins to flat out or started to increase.

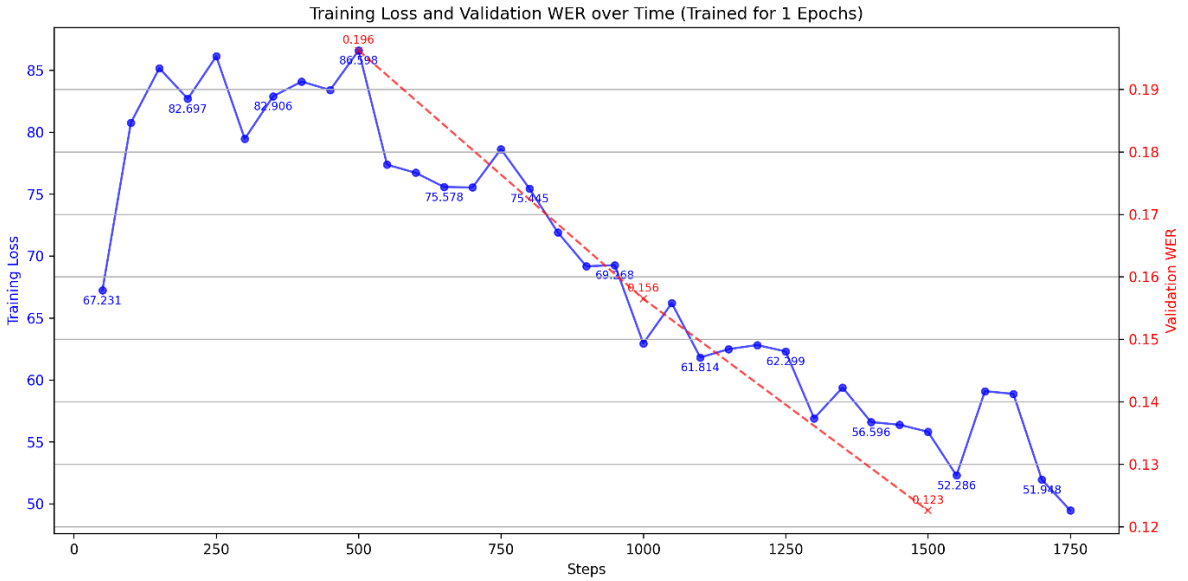

Figure 1: training loss and validation WER graph of 1 epoch model

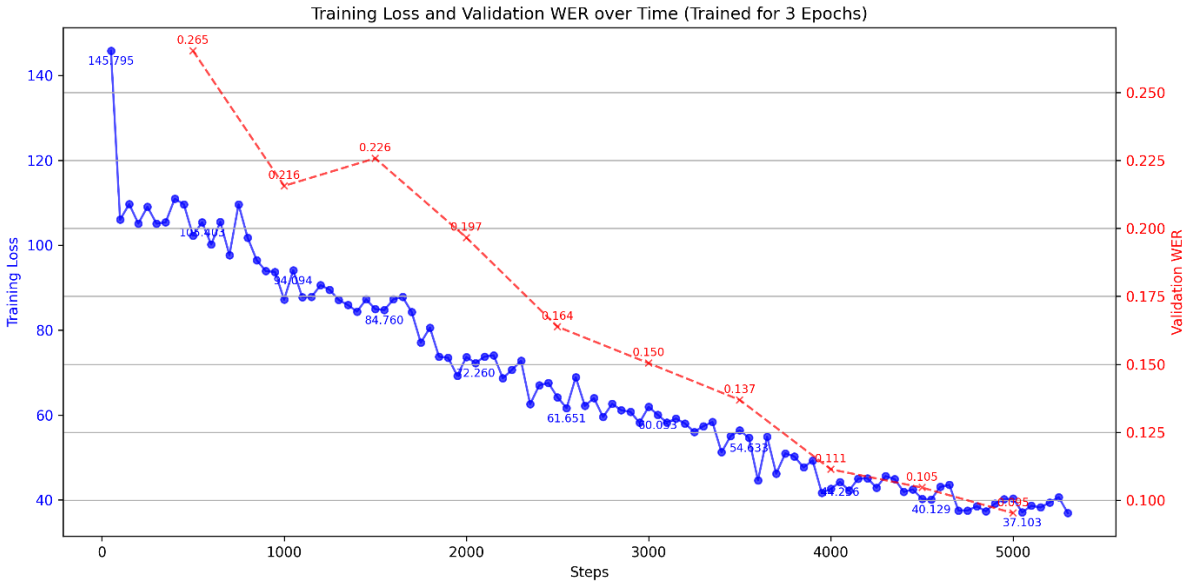

Figure 2: training loss and validation WER graph of 3 epoch model

Comparing 1 epoch versus 3 epoch model finetuned with facebook/wav2vec2-large-960h using common voice dataset with sampled rate of 16k and length of mp3 not more than 3 seconds:

Figure 1 shows training loss decreases but there are some fluctuations in the middle steps, similarly, the Word Error Rate (WER) shows a decline as the training progresses, it indicates that the accuracy is improving over time.

Figure 2 shows a longer training with a more significant decrease in training loss and a substantial reduction in Word Error Rate (WER), suggesting that the training for 3 epochs

results in better performance and lower WER rate compared to 1 epoch.

In summary, while both models show improvements in training loss and validation WER, figure 2 still demonstrated the benefits of prolonged training with lower loss and WER values at the end.

Aside by increasing the number epoch to lengthen the training process, decreasing the learning rate can allow the model to make smaller adjustments to the weights, is especially good when the model needs to make smaller changes or improvements. One can try to decrease the learning rate incrementally and track the performance improvements, it is possible to use a learning rate scheduler that reduces the learning rate when the validation performance stops improving.

Using a larger, more diverse dataset will help the model learn better patterns and generalize over different inputs, reducing overfitting on smaller datasets. If it’s possible, add more examples of similar types of speeches or texts that the model is struggling.

Using larger batch size can improve model stability, it can stabilize training by averaging more data points but can also smooth out fine details. Users can experiment with different batch sizes, smaller sizes may allow model to learn more specific details while larger batches can improve training efficiency and stability.