Results for fine-tuned model against base model

```
WER for generated_text vs text: 0.12
WER for finetuned_text vs text: 0.09
```

Observations and Proposed Steps for Accuracy

1. The WER did decrease after fine tuning was conducted on the wav2vec2-large-960h model, with the Common Voice dataset. A 3% decrease was observed.
2. This was to be expected as (a) the cv-valid-dev mp3 dataset was a subset of the Common Voice dataset, and (b) both base and fine-tuning models were trained mainly on US/UK(EN) data. Minor artefacts observed were of plain errors (Fig. 1)

Ground Truth	**Fine Tuned model transcript**
be careful with your prognostications said the stranger	be careful whit your prognostications said the stranger
the boy had met the alconist	the boy had met the alkenists

Figure 1. *Examples of WER errors*

3. With the relatively low WER% achieved by both base and fine tuned model, there could be steps to be taken to perhaps reduce or keep the WER% below 10%:
    a. Utilise other datasets to augment the vocabulary model, as well as using the spoken portion of the datasets to further fine-tune the ASR model. These may include: GigaSpeech, IMDA National Speech Corpus (Part 1, 2; non-codeswitched), LJSpeech
    b. Adjust the hyperparameters: those that could potentially impact WER include
        i. Mixed precision (affects precision for memory allocation)
        ii. Optimiser method (Adam was used, but AdamW?)
        iii. Warmup steps
    c. Use better memory: this indirectly impacts how large of a batch size can be used to feed into the fine tuning process, and reduces the need for accumulated gradient calculations.