Fine-tuned WER Score: 1.0266326158336374
Original WER Score: 1.0130557148069006

The fine-tuned model and the original model exhibit similar Word Error Rate (WER) scores, which is to be expected since the fine-tuned model was trained on only a subset of the full training dataset. As a result, the model may not have reached its full potential yet.

However, if I were to train on the complete dataset, I would anticipate the fine-tuned model to outperform the original pre-trained model used in Task 2a, as fine-tuning on domain-specific data should improve performance.

Upon reviewing the predictions of the fine-tuned model, it showed that the model performed worse (having bad text predictions) for audio samples with stronger accents, such as the Indian accents, or when articulation is less distinct. This indicates a need for more examples from the underrepresented accents in the training dataset to help the model learn and generalise better across diverse speech patterns.

To improve the model's accuracy and robustness, several strategies can be employed:

1. Include audio samples representing a wider variety of accents, speaking speeds, and background environments. Increase the representation of underperforming accents, such as Indian accents, in the dataset.

2. Add background noise and overlapping music. Alter speaking speeds, apply pitch shifting, and adjust volume levels. These techniques can help the model generalise better to real-world scenarios.

3. Conduct systematic hyperparameter tuning, experimenting with parameters like learning rate, batch size, gradient accumulation steps, and dropout rates. Techniques such as grid search, random search, or Bayesian optimisation can be utilised to identify the optimal configuration.

4. Start training with simpler audio data before progressively introducing more complex examples. This approach could improve training efficiency and model accuracy.

5. Integrate language models (e.g., KenLM or Transformer-based models) during the decoding phase to improve contextual understanding and grammatical accuracy. Use post-processing techniques like spell-checking and grammar correction to refine transcriptions further.

While the fine-tuned model’s performance is currently similar to the original pre-trained model, there is substantial potential for improvement. By incorporating a full dataset, employing advanced data augmentation, optimising hyperparameters, and integrating language models, the fine-tuned model can achieve significantly better accuracy and robustness. These enhancements will enable it to handle diverse accents, challenging audio conditions, and broader use cases more effectively, making it a more reliable and adaptable speech recognition system.