The complete finetuning of the model could not be carried out due to time constraints. This report outlines the proposed steps to improve the accuracy.

Proposed methods to improve the accuracy:

Improving the accuracy of an audio transcription model, especially in terms of reducing word error rate (WER):

1. Data collection and pre-processing
   - A more diverse audio transcription dataset can be acquired, covering various accents, noise levels, and speaking styles to ensure robustness.
   - Audio pre-processing was not carried out in this task. Noise removal, volume normalisation, and possibly segmentation would have been carried out for easier processing.
2. Evaluation metrics
   - WER was used as the only evaluation metric in this finetuning procedure. Other seq2seq metrics could be additionally tracked.
3. Error analysis
   - Error analysis can be conducted to identify common types of transcription errors made by the model. An example would be spelling errors (as the tokenizer is on character level), such as extra repetition of letters, or misclassifying ‘th’ as ‘d’. This can guide subsequent improvement efforts by highlighting where the model struggles.
4. Data augmentation
   - Data augmentation was not carried out in this task. Augmenting the training data with techniques such as speed perturbation, adding background noise, and pitch shifting can help to simulate greater variations in the audio data.
5. Additional features
   - Although wav2vec2 was only trained on raw audio samples, the common voice dataset provides additional data for each sample, such as age, gender, accent, etc. These additional data can be extracted and used as features, combined with the output vector of the model (via concatenation for example).
   - The additional data can also be used in a multitask setting, where an additional classification loss is calculated by a separate output from the final vector to predict these metadata.
   - Personally, this is a less recommended procedure, because of how sparsely these data exist in the common voice dataset (a lot of blank values).
6. Language model integration
   - If there are sufficient compute resources, language models can be integrated into the transcription pipeline to improve the accuracy of the

generated transcripts. Techniques such as beam search decoding can also help in this regard.