Ways to improve the model:

1. A larger model can be used. We originally used the tiny version to speed up training, but larger versions with more layers, larger inputs, and more attention heads could significantly improve accuracy, but at higher computational cost.
2. Using more training data. Because of memory limits, we originally only used 15% of the data. The more data that would be available for fine-tuning, the better the results could become, especially with larger models that have more weights to update.
3. A better-quality dataset could be used. When going through the dataset and trying out the different recordings, I often had a hard time understanding my own language. The sentences often didn’t make sense, the transcriptions contained a lot of mistakes, and accents often weren’t really Dutch either.
4. The maximum number of training steps could be increased. To speed up training, this was set to 2000 or 3.12 epochs. Setting this to 5000 to 10000 could help improve performance further, although it would take even longer to train.
5. There may exist a better metric to measure model performance. WER, ‘word error rate’ was used, which equals (substitutions + deletions + inserts)/reference\_length. This does not look at how large the effect of a mistake might be, and the equal weighting of substitutions, deletions and inserts might not be optimal. An alternative could be NER, which weights errors based on their impact.
6. Different models could be tried, some options are: wav2vec2, Hubert, and librispeech-asr.