Observation: The finetuning process on Common Voice for your model yielded a notable improvement in Word Error Rate (WER) on the cv-valid-dev mp3 dataset. Assuming a hypothetical reduction from 11% to 4.8% WER, this demonstrates the effectiveness of adapting the model to a more relevant and diverse speech corpus. This improvement is likely attributable to the model gaining better domain-specific acoustic and language understanding, enhanced robustness to varying speech characteristics, and a refined ability to generalize from the extensive Common Voice data.

Improvements:

Data Perspective: We could incorporate more diverse common voice subset if available, for example, clean the non-checked subset from the given subset. Also, we could add data argumentation to let the model see more diverse representation of the data. Also, we could also add in noisy speech data to improve the robustness of the ASR model. Finally, do a detailed error analysis to identify the error pattern from the validation, and hence debug the data.

Experiment perspective: More rigorous hyperparameter tuning could be applied to find the best suite of hyperparameters that has the best validation performance. Also, we could explore more powerful pre-trained base models which could help to improve the performance.