1. An expected further improvement of model performance would come from utilizing the full traning dataset if we have sufficient RAM resources. Alternatively, a lower batch count during training and higher gradient_accumulation_steps can compenstate for the RAM usage with more traning time.

2. Data augmentation could be performed on the audio data, e.g. through low pass filtering, varying the tempo, or introducing artifects like echo to artificially increase the traning dataset.

3. The additional metadeta that came with the datasets: age, gender, accent could be incoporated into the model by adding additional layers on top of the base model for the extra inputs. These information can help the model make better predictions.