1 Question 8

The models used for the end to end ASR were:

- RNN and CTC
- RNN and Attention/CTC

Using the Census Database as the corpus the results obtained for both models were very similar.

Two different decoding configuration files were used:

- decoding_ctcweight0.5.yaml - with ctc_weight = 0.5
- decoding_ctcweight1.0.yaml - with ctc_weight = 1.0

The correct token percentage for the RNN and CTC model results were slightly better using the decoding configuration file decode_ctcweigth1.0.yaml and the RNN and Attention/CTC results were slightly better using the configuration file decode_ctcweigth0.5.yaml.

This differences might be due to the alignment model of the RNN and Attention/CTC model which improve the results with noisy data.

2 Question 9

2.1 9.9

The results obtained for our recordings are very poor not even achieving a correct token rate of 45%.
Looking at the results the model only seems to understand the numbers said, but any other word said is most likely to have every token deleted or being understood completely wrong (example : MILK understood as SEVEN). Isolated letters are understood a little better but there are still some mistakes.

2.2 9.10

The recording from "goncalo" folder are somewhat noisy so we tested for both models and both decoding configurations. With the RNN and Attention/CTC model is able to handle the noise very well compared to the model with no Attention.
The results with the first model were almost equal for both students but with the second one the lack of the alignment model makes the correct token rates for "goncalo_recx" to go down and the results for "ricardo_recx" to go up. The utterance which performs better is the license plate utterance since it doesn't have any words apart from isolated letters and numbers.