Speech processing

Part 2

1. The key difference here which we should note is that we have is a lot of data (relatively)

This difference would lead to differences in training time, for KNN based on DTW, we don't need any training, just save the points, thus it takes just about no time, in comparison to the DNN which takes time to train and learn from those examples, the more examples, the more training time we will have.

Inference time for the DNN is only affected by the input length and the network size (which doesn't change), unlike the KNN which would need to calculate the distance for the predicted example with **every** other example in our train set and is also affected by the size of the sample, when n is the size of the train set. Thus for a big dataset like GCommands the neural network would be faster given a GPU (which is a fair assumption).

Memory consumption of the network is affected only by the input size and the network size (which is constant), compared to the KNN which would again, need to store the full dataset. thus, memory consumption for the DNN would probably be smaller (assuming that the network size is smaller than the size of the train set)

2. For dataset A we would choose DTW and for dataset B we would choose CNN. In DTW, we would need to store the entire dataset (for the KNN) and storing 1,000,000 samples is a lot and is not needed for CNN. In addition, neural networks usually need many examples relative to classical models such as DTW.

3. To predict 5 digits instead of 1, We would replace the classification layer (usually a FC or MLP) that gives the scores to each digit with 5 different FC/MLP – one for each digit. Then, the first one will predict the first digit, the second one will predict the second and so forth. Meaning, the models will have the same CNN backbone to extract the latent space from the audio and the classification layers will be responsible for generating the digits.