Exercises on Speech Recognition

Ion Androutsopoulos, 2022–23

There are no exercises to be handed in as assignments in this part of the course.

1. Modify the formulae of exercise 5 of Part 4 (machine translation with BiLSTM encoder and LSTM encoder) for the case where the input to the encoder is a sequence of vectors representing speech frames (each vector represents a speech frame, e.g., vectors produced by wav2vec) instead of a sequence of word embeddings, and the decoder outputs at each time-step a single letter (as in slides 13–15).

2. Modify the formulae of exercise 4 of Part 5 (machine translation with CNN encoder and LSTM decoder) for the case where the input to the encoder is a sequence of real (or integer) numbers obtained by sampling an input speech signal (as in slide 7) instead of a sequence of word embeddings, and the decoder outputs at each time-step a single letter (as in slides 13–15). Hint: At the first convolutional layer, the filters are now applied to a window $[x_{i-k}, \ldots, x_i, \ldots, x_{i+k}]$ of $2k+1$ real numbers (or integers), where $2k+1$ is the size of the filters, instead of a window of $3 \cdot d^{(e)}$ real numbers obtained by concatenating three (for 3-gram filters) word embeddings of $d^{(e)}$ dimensions each.