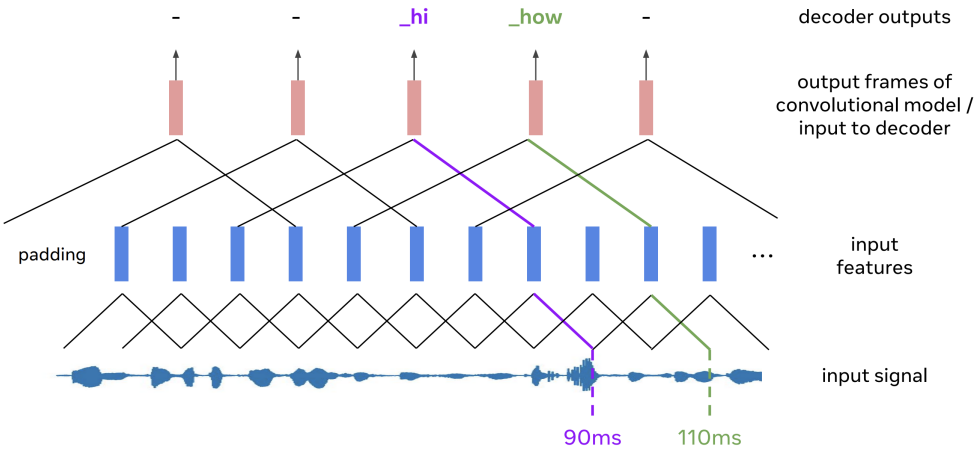
Example of per-word timestamps

In this example, a convolutional model is used. Inference is run continuously (not in a chunk-based manner). In particular:

- Feature extraction is performed with a window of 20 ms with shift of 10 ms
- Convolutional model processes the features. The model sub-samples the input sequence twice and for each output frame has receptive field covering 4 past and 2 future frames.
- Decoder decodes the sub-sampled sequence of output frames.



The timestamps for the output words are:

- *hi*: 90 ms (the word was emitted after seeing 3 output frames → 8 input frames → 90 ms of input signal)
- *how*: 110 ms (the word was emitted after seeing 4 output frames → 10 input frames → 110 ms of input signal)