End-to-end attention-based Automatic Speech Recognition

Problem description. End-to-end sequence-to-sequence models have been popularized in the last years for tasks such as machine translation and automatic speech recognition (ASR). In particular, attention mechanisms have proved to be very successful in modelling contextual, semantic, and long-term dependencies inherent to natural language. Thus, they have been applied to many tasks concerning language such as language generation, question answering, machine translation, speech recognition, etc.

Open source releases typically include models pre-trained for the most common languages, e.g English, but the community still needs to mature in terms of open-sourcing the code for attention-based ASR models in a unified framework, similarly to DeepSpeech for end-to-end RNN-based ASR. The availability of open source code allows the community to train their own models in any language and test different modifications of the baseline model.

Some organizations such as HuggingFace and Speechbrain are already taking a step in this direction, and the goal is to use the mentioned frameworks to build our own end-to-end pipeline that allows us to train attention-based ASR models using Norwegian data.

This thesis will be part of the research project *Scribe*, which aims to provide a Norwegian ASR engine applicable to real-life conversations, and contribute to language technology development in general. As such, the student will be integrated within a large team of researchers.

Main tasks include:
Tasks 1-2 pertain to the specialization project, 3-4 to the possible extension as master's thesis.

1. Literature review on end-to-end attention-based ASR.
2. Assessment of open source frameworks and their usability for the project.
3. Use the available resources to build an end-to-end attention-based ASR pipeline.
4. Use the pipeline to train models on Norwegian data, predict speech transcriptions, and compare with baseline models based on recurrent architectures.

Data. We provide three main datasets fully processed and ready for ASR tasks: Nordisk Språkteknologi (NST), Norwegian Parliament Speech Corpus (NPSC) and Telenor Norway's Customer Service (TNCS), consisting of approximately 400h, 60h and 15h, respectively, of audio recordings and their corresponding transcriptions. In addition, RNN-based ASR models (in PyTorch) already trained on these datasets are provided for comparison of results.

Objective. The ultimate objective is to put together the available open-sourced components of attention-based ASR to form an end-to-end pipeline, and compare with baseline models. This is an important contribution to *Scribe* and the ASR research community.

Academic supervisor: Giampiero Salvi, NTNU.
Co-supervisor: Pablo Ortiz, Telenor Research (pablo.ortiz@telenor.com), Marco Siniscalchi, NTNU.