Improved Automatic Speech Recognition through better decoding

Problem description. In automatic speech recognition (ASR) systems, a transcription is generated when a sound utterance is given to the system. In the ASR system under study, this transcription is the result of a prediction model that first generates a number of candidates *n* based on a trained acoustic model (AM), a process called *beam search decoding*. Then it re-scores those *n*-best candidates based on a *separately trained* language model (LM). Given the weights of both AM and LM, the system outputs the best candidate.

However, this decoding is far from optimal due to several factors:

- The weights of AM and LM are adjusted based on grid-search, not learnt by training.
- The weights of AM and LM are constants and don't adapt to new input data.
- The *n*-best candidates are all very similar and models struggle to distinguish which one is really the best candidate, causing word accuracy to fall dramatically.

Several strategies can potentially alleviate these problems, for instance:

1. Fully differentiable beam search decoder, such that AM and LM can be jointly trained.
2. Weights that are functions of the input data.
3. Lattice decoding and rescoring.
4. Diverse beam search decoding, to produce more diverse *n*-best candidates.

Already provided is the data and baseline ASR system to work with, and this project will focus on implementing at least **one** of the strategies mentioned above, depending on what the student and supervisors jointly decide, and of course on the progress.

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, 2-3 to the possible extension as master's thesis.

1. Literature review on optimization and decoding in ASR systems.
2. Implementation of at least one of the strategies in the provided ASR pipeline.
3. Test the new pipeline on Norwegian data and compare with baseline models.

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 already trained on these datasets are provided for comparison of results.

The ASR system under study is based on DeepSpeech 2. It is mostly written in Python and PyTorch, while the decoding component is in C++, so we encourage applicants with previous knowledge of those programming languages.

Objective. The ultimate objective is to improve the performance of ASR for low-resource languages such as Norwegian, by means of more intelligent methods, in this case during the decoding phase. 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.