Speech Enhancement for Robust Automatic Speech Recognition

Problem description. End-to-end automatic speech recognition (ASR) systems are built on the idea that leveraging the powerful feature extraction capability of neural models and huge amounts of data can alone address the challenging ASR task.

However, this approach has several issues:

- It does not take into account what has been learned about the speech signal in the past decades.
- It requires an enormous amount of speech data, and it is not easily portable to dialects and low-resources languages
- It may miserably fail in unseen adverse environmental conditions, and fine-tuning on the target test conditions may be needed to improve performance.

In contrast, it has constantly been reported that "*good signal processing can lead to top ASR performance in challenging acoustic environments*". Indeed, in the last few years, several works had focused on developing neural network approach for speech enhancement, speech dereverberation, source separation, voice activity detection and speech bandwidth expansion, outperforming the performance of all state-of-the-art digital signal processing (DSP) competing algorithms that we have compared with.

With respect to speech enhancement, several strategies can be used to perform signal processing using deep neural networks, for instance:

1. Mapping based schemes.
2. Masking based solutions.
3. GAN-based solutions.
4. Distributional based schemes

The project will focus on implementing at least **one** of the speech preprocessing 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, 3-4 to the possible extension as master's thesis.

1. Literature review on speech enhancement.
2. Implementation of at least one of the speech enhancement strategies testing the improvement on the related ASR task.
3. Test the new pipeline on Norwegian data and compare with baseline models.

Data. We provide three main datasets for speech enhancement, source code to generate training data. Moreover, data fully processed and ready for ASR tasks would be ready, namely: 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. An important aspect to note about these datasets is the degree of noise and impurities in the speech signal, in descending order TNCS > NPSC > NST.

Objective. The ultimate objective is to improve the performance of ASR for low-resource languages such as Norwegian in noisy conditions. This is an important contribution to *Scribe* and the ASR research community.

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