Université de Lorraine - IDMC
Masters SC & TAL – M1
2020-2021

Fiche de projet tutoré / Project form

Signal Processing & Linguistics in Self-Supervised Learning for Automatic Speech Recognition

Encadrement / Supervisors

1. équipe, laboratoire / team, lab
   Multispeech
2. encadrant·e principal·e (nom, email) / main supervisor (name, email)
   Imran Sheikh (imran.sheikh@inria.fr)
3. autres encadrant·es / other supervisors

Description / Description

1. projet global/global project

Automatic Speech Recognition (ASR) systems [1] are typically trained on large amounts of speech data comprising spoken utterances and their text transcriptions. Obtaining such human transcribed or crowdsourced annotated speech corpora is known to be highly expensive or non-realistic for practical applications. Moreover, such a process is also not feasible for vast majority of languages which are underrepresented in ASR research and development.

More recently, self-supervised learning approaches have been successfully applied in training of ASR systems. These approaches can start learning ASR relevant representations of speech without annotated data. A typical trend with these approaches is (a) to rely on thousands of hours of unlabeled speech data, and (b) use of heavy machine learning models which are trained on auto-regressive or encoding-decoding tasks [2,3]. However, methods from signal processing and linguistics can add complementary information to these approaches, and hopefully scale down or simplify the structure and training of the involved models [4].

The aim of this project is to:

(a) survey the direct or in-direct use of signal processing and linguistic methods in existing self-supervised learning approaches for ASR, and
(b) identify and analyse areas where such methods can be complmentary [5].

2. biblio. UE 705 (semestre 7)

The students will have the opportunity to gradually read and learn:

(a) theory on building ASR systems
(b) recent approaches related to self-supervised learning in ASR

Both (a) and (b) will be accompanied with joint discussions, as often as bi-weekly. Students are expected to make regular notes on the reading materials, which will ultimately contribute to the bibliography report at the end of the semester. In case of (b), knowing or learning details about the heavy machine learning models will be out of scope of this project. The focus will be on understanding the overall scheme and identifying steps that relating to signal processing and linguistics.

3. réalisation. UE 805 (semestre 8)
Realisation of the project will involve hands on with existing models including an analysis on:

(a) phoneme recognition and seperability
(b) speaker invariance
(c) generalization to other langauges

Based on the bibligraphy report and the interest of the students, there will be an opportunity to build a self-supervised phoneme discovery model based on simple Recurrent Neural Networks (RNN) and spectral transition measure from speech processing.

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

Corpus :

- Librispeech (https://www.openslr.org/12/) ,
- Common Voice (https://commonvoice.mozilla.org/en)

Tools:

- S3PRL toolkit (https://github.com/andi611/Self-Supervised-Speech-Pretraining-and-Representation-Learning),
- PASE toolkit (https://github.com/santi-pdp/pase)

Livrables et échéancier / Deliverable and schedule

October-December: Reading ASR theory, self-supervision methods. Writing bibliography.
January: Introduction to speech corpus and tools.
February-March: Analysis on existing models and tools. Experiments on new ideas.
April-May: Compilation and analysis of the results and final report writing.

Bibliographie /References (max. 4-5)

[1] Maël Fabien, Introduction to Automatic Speech Recognition, https://maelfabien.github.io/machinelearning/speech_reco/#
[2] Alexei Baevski et. al., wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, https://arxiv.org/abs/2006.11477
[3] Alexander H. Liu et. al., Towards Unsupervised Speech Recognition and Synthesis with Quantized Speech Representation Learning, https://arxiv.org/abs/1910.12729
[4] Mirco Ravanelli et.al., Multi-task Self-supervised Learning For Robust Speech Recognition, https://arxiv.org/abs/2001.09239
[5] Maria Andrea Cruz Blandon and Okko Räsänen, Analysis of Predictive Coding Models for Phonemic Representation Learning in Small Datasets, https://arxiv.org/abs/2007.04205