Laboratoire d'**I**nformatique et **S**ystèmes
LIS – UMR CNRS 7020

Stage master recherche
Deep learning for speech perception
(Modélisation de la perception de la parole par apprentissage profond)

Length of internship: 5-6 months

Start date: between January and March

Contact: Ricard Marxer <ricard.marxer@lis-lab.fr>

Context

Recent deep learning (DL) developments have been key to breakthroughs in many artificial intelligence (AI) tasks such as automatic speech recognition (ASR) [1] and speech enhancement [2]. In the past decade the performance of such systems on reference corpora has consistently increased driven by improvements in data-modeling and representation learning techniques. However our understanding of human speech perception has not benefited from such advancements. This internship sets the ground for a project that proposes to gain knowledge about our perception of speech by means of large-scale data-driven modeling and statistical methods. By leveraging modern deep learning techniques and exploiting large corpora of data we aim to build models capable of predicting human comprehension of speech at a higher level of detail than any other existing approaches [3]. This internship is funded by the ANR JCJC project MIM (Microscopic Intelligibility Modeling). It aims at predicting and describing speech perception at the stimuli, listener and sub-word level. The project will also fund a PhD position, the call for applications will be published in the coming months. A potential followup in PhD could be foreseen for the successful candidate of this internship.

Subject

The goal of this internship is to produce the first DL-based models that predict human intelligibility at the sublexical level. This translates into predicting the positions in the audio stimuli where confusions will occur, and the type of confusions, in other words, which phonemes are confused with which others on an individual basis.

We will use data from a public corpus of consistent confusions that we have produced (http://spandh.dcs.shef.ac.uk/ECCC/): speech-in-noise stimuli that evoke the same misrecognition among multiple listeners. In order to simplify this first approach to microscopic intelligibility prediction, we restrict ourselves to single-word data. This reduces the lexical factors to aspects such as usage frequency and neighborhood density, significantly limiting the complexity of the required language model. Consistent confusions are valuable experimental data about the human speech perception process. They provide targets for how intelligibility models should dif-ferentiate from automatic speech recognition (ASR) systems. While ASR models are optimised to recognise what has been uttered, the proposed models should output what has been perceived by a set

of listeners. A sub-task encompasses creating baseline models that predict listeners' responses to the noisy speech stimuli. We will target predictions at different levels of granularity such as predicting the type of confusion, which phones are misperceived or how a particular phone is confused.

Several models regularly used in speech recognition tasks will be trained and evaluated in predicting the misperceptions of the consistent confusion corpora. We will first focus on well established models such as GMM-HMM and/or simple deep learning architectures. Advanced neural topologies such as TDNNs, CTC-based or attention-based models (CPC, wav2vec2) will also be explored, even though the relatively small amount of training data in the corpora is likely to be a limiting factor. As a starting point we envisage solving the 3 tasks described in [3] consisting of 1) predicting the probability of occurrence of misrecognitions at each position of the word, 2) given the position, predicting a distribution of particular phone misperceptions, and 3) predicting the words and the number of times they have been perceived among a set of listeners. Predictions will be evaluated using the metrics also defined in [3] and random and oracle predictions will be used as references. These baseline models will be trained using only in-domain data and optimized on word recognition tasks.

Profile

The candidate shall have the following profile:

- Master 2 level or equivalent in one of the following fields: machine learning, computer science, applied mathematics, statistics, signal processing
- Good English written and spoken language skills
- Programming skills, preferably in Python

Furthermore the ideal candidate would have:

- Experience in one of the main DL frameworks (e.g. PyTorch, Tensorflow)
- Notions in speech or audio processing

Application procedure

If interested please contact ricard.marxer@lis-lab.fr

References

[1] Barker, J., **Marxer, R.**, Vincent, E., & Watanabe, S. (2017). The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes. Computer Speech & Language, 46, 605–626.

[2] **Marxer, R.**, & Barker, J. (2017). Binary Mask Estimation Strategies for Constrained Imputation-Based Speech Enhancement. In Proc. Interspeech 2017 (pp. 1988–1992).

[3] **Marxer, R.**, Cooke, M., & Barker, J. (2015). A framework for the evaluation of microscopic intelligibility models. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (Vol. 2015-January, pp. 2558–2562).