Master's internships 2021/2022

A limited number of positions for a Master's internship are available this year. These internships should be for A MINIMUM OF 5-6 MONTHS. Please send your application by email (CV + LM required) to syntheticlearner@gmail.com.

1. Self-supervised learning at the word level

Supervision in machine learning is a paradigme that requires labelled datasets which is often the result of substantial human time and efforts. For that reason, unsupervised or self-supervised methods are becoming increasingly used in several areas of machine learning: vision, text, and very recently speech. For instance, contrastive predictive coding or Wav2vec2.0 have been used to discover speech representations without supervision [1,2]. These models can embed fixed duration (usually 10ms) of speech into a vector but cannot represent variable-length speech sequences. These latter representations can be very useful in a variety of tasks ranging from information retrieval to speech segmentation into words (i.e can you find word boundaries in an audio recording without label nor prior knowledge of the language?).

The aim of this internship is to search for new and more robust methods to build variable length speech embeddings. You will implement new loss functions and regularisation schemes in a pre-existing deep learning model to improve its performance. The resulting model will be used as input to a speech segmentation model and hopefully improve the current state-of-the-art in that domain. This internship will be done in collaboration with researchers at Facebook AI Research.

[1] Aaron van den Oord, Yazhe Li, Oriol Vinyals (2019). Representation Learning with Contrastive Predictive Coding https://arxiv.org/pdf/1807.03748.pdf
[2] https://arxiv.org/abs/2006.11477
[3] https://arxiv.org/abs/2007.13542

2. Universal speech synthesis

Text-to-speech synthesis (like the voice of Siri, or of Google Translate) has seen stunning advances in the last five years, sounding extremely natural. However, like many other speech tools, it requires massive amounts of labelled training data to construct a good speech synthesis model, and this means concentrating on single language, a single, usually highly standardized, accent, and, for best results, a single speaker. As a consequence, there are not many languages with high-quality speech synthesis. A new wave of speech synthesis attempts to move towards *multilingual* speech synthesis, so that, on the basis of training data in a handful of languages, we would have a single model

that could generate speech in other languages. The goal of this internship is to advance *universal* speech synthesis, a single model which can synthesize speech in any language in the world.
Several papers have attempted tasks going in this direction, including recent work in our lab. A promising approach is that taken by [https://arxiv.org/abs/2008.04107], which uses meaningful articulatory-inspired features as input, to allow for a more general model, as well as the task of cross-lingual voice transfer [https://arxiv.org/abs/1907.04448], which takes speech in an unknown language as input and re-speaks it, without knowledge of this language. Tasks for the intern may thus include attempting to take the articulatory features approach, and seeing whether it continues to be useful for the voice transfer task, or developing a more rigorous evaluation in order to push the limits of these models, depending on the interests and aptitudes of the intern.

3. Domain discovery for blind distributional robustness in deep learning models

Current neural network-based machine learning models are able to attain high accuracy across a wide range of modalities and tasks. However, they often fail when used on data from different domains: for example, a machine translation model trained on news articles will tend to perform poorly when translating social media comments, a very different kind of text [1]. This has unfortunate implications for their use in real-world scenarios where distribution shifts abound.
There has been a recent surge of interest in developing models that are robust to domain shift. However, most such approaches rely on having examples from a variety of different domains at training time to learn invariant representations [2], or make strong assumptions on the nature of the shift (eg. that the target domain overlaps with the training data [3]).
The subject of this internship will be to develop a training algorithm that is able to yield models that are more robust to domain shift, without the aforementioned limitations. A potential idea to explore would be to uncover domains present in the training data, and use this information to learn invariant features. However, other approaches could be explored depending on the interests of the intern. For inquiries, please contact pmichel31415@gmail.com.

[1] Michel, P., & Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. https://arxiv.org/pdf/1809.00388.pdf
[2] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. https://arxiv.org/pdf/1907.02893.pdf
[3] Oren, Y., Sagawa, S., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust language modeling. https://arxiv.org/pdf/1909.02060.pdf

4. Using deep learning to study children's multimodal behavior in face-to-face conversation

If you want to apply for this internship subject, please send your inquiries and application directly to Abdellah Fourtassi (abdellah.fourtassi@gmail.com)

The study of how children develop their conversational skills and how these skills help them learn from others is an important scientific frontier at the crossroad of social, cognitive, and linguistic development with important applications in health (e.g., mitigating communicative difficulties), education (e.g. improving teaching practices), and child-oriented AI (e.g., virtual learning companions). Recent advances in Natural Language Processing and Computer Vision allow going beyond the limitations of traditional research methods in the lab and advance formal theories of conversational development in real-life contexts. In this internship, we will leverage some of these recent techniques (e.g., multiscale recurrent neural network, see [1]) to build a model that mimics how children behave in face-to-face conversations with their caregivers and how this behavior develops across middle childhood. The intern will have access to the child-caregiver conversation data collected by our team [2]. The data has already been hand-annotated for non-verbal behavior (e.g., nods, smiles, and frowns) and is currently being transcribed for verbal data and processed for extraction of vocal/acoustic features. The interns will contribute to the development of a model (building on an existing pipeline in PyTorch) that aims at studying how multimodal cues from the vocal, visual, and verbal dimensions contribute to predicting the child's coordination behavior in conversation (e.g., turn-taking management, negotiating shared understanding with the interlocutor, and the ability for a coherent/contingent exchange). The intern will collaborate closely with several members of our team, involving computer scientists, psychologists, and linguists (see our website www.cocodev.fr) as well as members from the CoML team.

[1] Roddy, M., Skantez, & Harte (2018). Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. *In Proceedings of the 20th ACM International Conference on Multimodal Interaction*

[2] Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., & Fourtassi, A. (2021). ChiCo: A Multimodal Corpus for the Study of Child Conversation. *In Proceedings of the International Workshop on Corpora and Tools for Social Skills Annotation. 23rd ACM* International Conference on Multimodal Interaction

5. Machine learning and speech modelling for Neurology

Current methods for quantifying cognitive and neurological impairments are exhausting, time consuming, stressful and expensive. Speech production invokes various cognitive, linguistic, emotional and motor skills, and its realisation is greatly influenced by the mental and neurological state of the individual.
The goal of this internship is to find and analyze spontaneous speech from individuals with neurodegenerative disorders such as Huntington's Disease and Parkinson's Disease. First, the intern will build custom deep learning architectures specific to the speech of neurological patients. Then, the intern will build specific representations and models to quantify the severity of the disease.
This internship will be done in collaboration with Neurologists, Neuropsychologists and Linguists.

1) Riad, R., Titeux, H., Lemoine, L., Montillot, J., Bagnou, J. H., Cao, X. N., ... & Bachoud-

Lévi, A. C. (2020). Vocal markers from sustained phonation in Huntington's Disease. Interspeech 2020
2) Perez, Matthew, et al. "Classification of Huntington Disease Using Acoustic and Lexical Features." Interspeech. 2018.
3) Al Hanai, Tuka, Rhoda Au, and James Glass. "Role-specific Language Models for Processing Recorded Neuropsychological Exams." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018.

6. Learning speech sounds like an infant

Current methods for speech processing (speech recognition, speech synthesis) require supervised learning: they require mountains of speech data which are *labelled* with the sequence of speech sounds or letters which was uttered. In contrast, infants learn to perceive the sounds of their native language well before their first birthday, and can articulate most or all sounds in an adult-like manner well before they can read. There is therefore something unnecessary about the supervised approach to speech processing, as demonstrated by the fact that infants and young children do not need subtitles or transcriptions to learn the sounds of their first language. As of yet, however, we do not have reliable methods for “discovering” the sounds (consonants and vowels) of a language in an unsupervised fashion. Recent methods have shown excellent promise (1,2) but are still limited to discovering relatively short-duration acoustic events, and attempts to deal with the problem of *segmenting* continuous speech into meaningful units in an unsupervised fashion have seen limited success for *discovery* (3,4).

This internship will deal with the problem of unsupervised discovery of speech units which has been developed by the team in the last five years into the Zero Resource Speech Challenge (5,6). The intern will deal with the questions of (i) how to appropriately evaluate discovered units and (ii) how to improve them. We will explore the idea that placing constraints grounded in speech articulation can lead to better units.

(1) Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
(2) Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.
(3) Bhati, S., Villalba, J., Żelasko, P., Moro-Velazquez, L., & Dehak, N. (2021). Unsupervised Speech Segmentation and Variable Rate Representation Learning using Segmental Contrastive Predictive Coding. *arXiv preprint arXiv:2110.02345*.
(4) Ondel, L., Vydana, H. K., Burget, L., & Černocký, J. (2019). Bayesian subspace hidden Markov model for acoustic unit discovery. *arXiv preprint arXiv:1904.03876*.
(5) Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., ... & Dupoux, E. (2017, December). The zero resource speech challenge 2017. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 323-330). IEEE.
(6) Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T. A., de Seyssel, M., Rozé, P., ... & Dupoux, E. (2021). The Interspeech Zero Resource Speech Challenge 2021: Spoken language modelling. *arXiv preprint arXiv:2104.14700*.

7. Speech and language tools for under-documented languages

The aim of our LAAC (Language Acquisition Across Cultures) team is to shed light on the mechanisms and processes involved in early language acquisition in a variety of cultures

and language communities. To this end, we use an interdisciplinary approach (ranging from computational modelling to laboratory experiments and advanced data analysis) in the context of open, collaborative and publicly engaged science.

Most research on early language acquisition has documented input and learning cues in only a handful of cultures, the assumption being that the mechanisms postulated to explain acquisition in these cultures are universal. However, there is too little research on some languages.

The team now has data from many uncommonly studied languages (such as Tsimane', Yélî Dnye, and many others). We now want to analyse these corpora:

- systematise and clean the data
- for corpora of texts, generate orthographic and phonological dictionaries
- for corpora of speech and texts, generate structured alignments
- where possible, generate tests at different linguistic levels as in the ZR Speech Challenge

Internship's objectives:

- Develop speech and text tools for low-resource languages
- Join an interdisciplinary team
- Learn about open and cumulative science (a response to the replication crisis)
- Experience life in the Lab
- Exposure to research methods in experimental psychology and language science