********** Post-doctoral position **********

Post-Doctoral Research: "Disentangled and Interpretable Representation Learning for Audio-visual Speech"
Team: MULTISPEECH, Inria Nancy - Grand Est, France.
Contacts: Mostafa Sadeghi (mostafa.sadeghi@inria.fr), Romain Serizel (romain.serizel@loria.fr), Emmanuel Vincent (emmanuel.vincent@inria.fr)

Context: Representation learning has long been an important step in machine learning and signal processing [1]. Over the past few years, deep latent variable generative models, including variational autoencoder (VAE) and normalizing flow (NF), have been extensively utilized to model data generative processes and to learn a representation of data. A representation vector provides a concise description of data and is desired to capture and disentangle interpretable factors of variation in data. This enables learning explainable, robust, and universal representations that generalize well across different domains and tasks, and provides human-controllable data generation [2].
In particular, speech as a complex signal carries various sources of information, ranging from (speaker-independent) linguistic content to identity, age, emotion, and health status of the speaker, which are entangled and combined in a sophisticated way to produce speech. In privacy-preserving speech processing, for instance, disentangling the speech representation in order to keep only the required attributes and to remove the identity-related ones is of paramount importance. In some other applications, e.g. expressive text-to-speech and voice conversion, it is desired to manipulate the generative factors of speech.
Unsupervised disentangled representation learning has been proven to be impossible without inductive bias [3], and supervised learning requires expensive labeled data. Nevertheless, weakly supervised and semi-supervised approaches have shown promising performance recently [4]. In weak supervision, one has access to pairs of data that share some subsets of generative factors, and in semi-supervision, a limited set of data with known ground-truth generative factors is available.

Project description: Despite several attempts to learn disentangled and interpretable representations for sequential data, including speech, they are still far from achieving satisfactory performance. Furthermore, the majority of the existing works assume independence on the latent dimensions, whereas in real-world applications, subsets of latent codes (generative factors) might be correlated. On the other hand, human perception is inherently multimodal, yet there is only limited research on multimodal disentangled representation learning. Specifically, the visual modality, e.g jaw and lip movements, which provides complementary information about speech [5], has not been properly investigated in this context, neither as a stand-alone signal (e.g. for lip-reading tasks) nor in combination with the acoustic modality (e.g. for audio-visual speech processing).
In this post-doc project, we are going to develop efficient frameworks to learn disentangled and interpretable representations for audio-visual speech. To achieve this objective, weakly/semi-supervised approaches along with flexible (hierarchically structured) latent variable models will be investigated. We will also explore transfer learning techniques to enable learning with the help of a second dataset (synthetic or real) that has ground truth annotations for generative factors. We will build our models using probabilistic generative modeling, in particular, flow-based latent variable models in combination with adversarial

training, which have proven effective in representation learning and distribution modeling. We will use the TCD-TIMIT [6] and RAVDESS [7] audio-visual speech corpora.

References:

[1] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives," IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 35(8): 1798–1828, 2013.
[2] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. "Interpretable machine learning: Fundamental principles and 10 grand challenges," arXiv preprint arXiv:2103.11251, 2021.
[3] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Scholkopf, and Olivier Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in International Conference on Machine Learning, pp. 4114–4124, 2019.
[4] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole, "Weakly supervised disentanglement with guarantees," in International Conference on Learning Representations (ICLR), 2020.
[5] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 28, pp. 1788 –1800, 2020.
[6] N. Harte and E. Gillen, "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," IEEE Transactions on Multimedia, vol.17, no.5, pp.603-615, May 2015.
[7] S. R Livingstone and F. A Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PloS one 13, 5 (2018).

Requirements & skills:

- Ph.D. degree in the field of speech/audio processing, computer vision, machine learning, or in a related field,
- Ability to work independently as well as in a team,
- Solid programming skills (Python, PyTorch),
- Decent level of written and spoken English.

Facilities:

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.) Social, cultural, and sports events and activities
- Access to vocational training
- Social security coverage

Salary: 2653€ gross/month
Application & more information: https://jobs.inria.fr/public/classic/fr/offres/2021-03550