A propos du centre ou de la direction fonctionnelle
Grenoble Rhône-Alpes Research Center groups together a few less than 800 people in 39 research teams and 8 research support departments.
Staff is localized on 5 campuses in Grenoble and Lyon, in close collaboration with labs, research and higher education institutions in Grenoble and Lyon, but also with the economic players in these areas.
Present in the fields of software, high-performance computing, Internet of things, image and data, but also simulation in oceanography and biology, it participates at the best level of international scientific achievements and collaborations in both Europe and the rest of the world.

Contexte et atouts du poste
The PhD work will take place at Inria Grenoble, in Montbonnot-Saint-Martin, in the Perceptron Team, headed by Radu Horaud. It will be supervised by Laurent Ginir (Professor Grenoble-INP) & Xavier Alameda-Pineda (Inria Research Scientist).

Mission confiée
The overall goal of the proposed PhD topic is to develop deep generative models for the automatic analysis of audio-visual temporal data. In the context of human-robot interaction, we want to automatically estimate how many people participate to a conversation, where they are, what they are saying, to whom, which gestures they perform, see [1,2]. The developed algorithms are expected to be implemented in a companion humanoid robot for Human-robot social interaction. In this PhD work, we will explore the development of deep probabilistic generative models [3,4].

Learning perception models in multi-person scenarios is challenging because we need to properly fuse multi-sensory (mainly audio-visual) data, efficiently solve the combinatorial observation-to-person assignment problem and account for a time-varying number of people. Therefore, we have to conceptualize parametric models that are able to faithfully and efficiently represent a scene, and develop and evaluate the associated parameter estimation algorithms. Importantly, for the sake of interpretation, the representation should be structured into the set of individual cues per person plus a set of collective cues. We will inspire on state-of-the-art techniques for visual person (body, face) detection and person description (appearance, pose, orientation), on the one side, and on speech processing (speech enhancement, speech and speaker automatic recognition), on the other side. Part of the informative features will be extracted using learnable parametric methods, e.g. deep neural networks (DNNs). We will need to investigate how to fine-tune these architectures to satisfy the goals of the project, and to adapt to the data distribution of multi-person conversational scenarios. Once these features are conceived and learned, we will be able to perform joint inference of individual and collective cues, and define and address the combinatorial assignment problem.

The fact that multiple features are extracted and can be assigned to multiple persons, together with the impact this assignment has on the temporal dynamics, leads to a combinatorial problem growing exponentially with time. A clear example of this is found in multi-person tracking [5] and in sound separation [6] for which we have explored generative/Bayesian probabilistic models and associated solutions based on variational inference. In the present PhD work, this kind of models will be combined with DNNs. The joint training of probabilistic and deep neural models is difficult and has to be done with care [7]. The PhD student will be expected to design the deep architectures able to extract cues from raw data design, to conceive their combination with probabilistic models, and to develop the optimization frameworks and algorithms able to soundly optimize for the overall set of parameters.

References

Compétences
Research Master's degree, or equivalent, in a discipline connected to signal and information processing, computer vision and machine learning. Experience in probabilistic models, specifically variational auto-encoders is highly welcome. A particular interest/experience in speech/audio processing, visual recognition, and/or multimodal fusion is a plus. Strong motivation for the research work. Ability to work both independently and to collaborate within a small team. Computer skills: MATLAB, Python, Deep Learning Toolkits (e.g. Keras, Pytorch).

Avantages
- 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

Rémunération
1st and 2nd year: 1,982 euros brut /month
3rd year: 2,085 euros brut / month