RESEARCH CENTRE

Inria Centre
at Université Grenoble Alpes

IN PARTNERSHIP WITH:

Université de Grenoble Alpes

2023

ACTIVITY REPORT

Project-Team

ROBOTLEARN

Learning, perception and control for social robots

DOMAIN

Perception, Cognition and Interaction

THEME

Vision, perception and multimedia interpretation

Inria

Contents

Project-Team ROBOTLEARN 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3

3 Research program 4
3.1 Deep probabilistic models . . . 4
3.2 Human behavior understanding . . . 5
3.3 Learning and control for social robots . . . 7

4 Application domains 8

5 Social and environmental responsibility 9
5.1 Impact of research results . . . 9

6 Highlights of the year 11
6.1 Team Evaluation . . . 11
6.2 Keynote Speaker at ACM Multimedia 2023 . . . 11
6.3 PhD Defences . . . 11
6.3.1 Wen Guo . . . 11
6.3.2 Xiaoyu Bie . . . 12
6.3.3 Louis Airale . . . 12
6.4 Commitments . . . 12

7 New software, platforms, open data 12
7.1 New software . . . 12
7.1.1 xi_learning . . . 12
7.1.2 Social MPC . . . 13
7.1.3 2D Social Simulator . . . 13
7.1.4 dvae-speech . . . 13
7.1.5 exputils . . . 14
7.1.6 MixDVAE . . . 14
7.1.7 Light-DVAE . . . 14
7.1.8 DDGM-SE . . . 15

8 New results 15
8.1 Deep Probabilistic Models . . . 15
8.1.1 Unsupervised speech enhancement with deep dynamical generative speech and noise models . . . 15
8.1.2 Speech Modeling with a Hierarchical Transformer Dynamical VAE . . . 16
8.1.3 Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation . . . 16
8.2 Human Behavior Understanding . . . 16
8.2.1 Continual Attentive Fusion for Incremental Learning in Semantic Segmentation . . . 16
8.2.2 Semi-supervised learning made simple with self-supervised clustering . . . 17
8.2.3 Motion-DVAE: Unsupervised learning for fast human motion denoising . . . 17
8.2.4 Learning and controlling the source-filter representation of speech with a variational autoencoder . . . 17
8.2.5 Expression-preserving face frontalization improves visually assisted speech processing 18
8.2.6 Audio-Visual Speaker Diarization in the Framework of Multi-User Human-Robot Interaction . . . 18
8.2.7 Back to MLP: A Simple Baseline for Human Motion Prediction . . . 18
8.3 Learning and Control for Social Robots . . . 19
8.3.1 Variational Meta Reinforcement Learning for Social Robotics . . . 19

8.3.2 Successor Feature Representations . . . 19

9 Partnerships and cooperations **20**
9.1 International initiatives . . . 20
9.1.1 Visits to international teams . . . 20
9.2 European initiatives . . . 20
9.2.1 H2020 projects . . . 20
9.3 National initiatives . . . 20
9.3.1 ANR JCJC Project ML3RI . . . 20
9.3.2 ANR MIAI Chair . . . 21

10 Dissemination **21**
10.1 Promoting scientific activities . . . 21
10.1.1 Scientific events: organisation . . . 21
10.1.2 Scientific events: selection . . . 21
10.1.3 Journal . . . 21
10.1.4 Invited talks . . . 21
10.1.5 Leadership within the scientific community . . . 22
10.2 Teaching - Supervision - Juries . . . 22
10.2.1 Teaching . . . 22
10.2.2 Supervision . . . 22
10.2.3 Juries . . . 22

11 Scientific production **22**
11.1 Major publications . . . 22
11.2 Publications of the year . . . 24
11.3 Cited publications . . . 25

Project-Team ROBOTLEARN

Creation of the Project-Team: 2021 July 01

Keywords

Computer sciences and digital sciences

A5.4.2. – Activity recognition

A5.4.5. – Object tracking and motion analysis

A5.7.3. – Speech

A5.7.4. – Analysis

A5.10.2. – Perception

A5.10.4. – Robot control

A5.10.5. – Robot interaction (with the environment, humans, other robots)

A5.10.7. – Learning

A9.2. – Machine learning

A9.3. – Signal analysis

A9.5. – Robotics

Other research topics and application domains

B2. – Health

B5.6. – Robotic systems

1 Team members, visitors, external collaborators

Research Scientists

- Xavier Alameda Pineda [Team leader, INRIA, Researcher, HDR]
- Patrice Horaud [INRIA, Emeritus, HDR]
- Chris Reinke [INRIA, Starting Research Position]

PhD Students

- Louis Airale [LIG, from Apr 2023 until Oct 2023]
- Louis Airale [INRIA, until Mar 2023]
- Anand Ballou [UGA]
- Xiaoyu Bie [UGA]
- Jordan Cosio [UGA, from Nov 2023]
- Wen Guo [POLE EMPLOI, from Aug 2023 until Oct 2023]
- Wen Guo [UGA, until Jun 2023]
- Gaetan Lepage [INRIA]
- Xiaoyu Lin [INRIA]

Technical Staff

- Alex Auternaud [INRIA, Engineer]
- Kirubakaran Ramamoorthy [INRIA, Engineer, from Jul 2023]
- Victor Sanchez [INRIA, Engineer, from Mar 2023]

Interns and Apprentices

- Ghazi Shazan Ahmad [INRIA, Intern, from Dec 2023]
- Andres Bermeo Marinelli [UNIV TRIESTE, from Jul 2023]
- Andres Bermeo Marinelli [INRIA, Intern, until Jul 2023]
- Ahmad Ghalawinji [INRIA, Intern, from Feb 2023 until Aug 2023]
- Daniel Jost [INRIA, Intern, from Nov 2023]
- Estelle Long-Merle [UGA, Intern, from Nov 2023]
- Victor Sanchez [INRIA, Intern, until Feb 2023]

Administrative Assistant

- Nathalie Gillot [INRIA]

Visiting Scientists

- Pietro Astolfi [UNIV TRENTE]
- Thomas De Min [UNIV TRENTE, from Feb 2023 until May 2023]
- Lorenzo Vaquero Otal [Universidad Santiago de Compostela]

External Collaborators

- Christian Dondrup [UNIV HERIOT-WATT]
- Laurent Girin [GRENOBLE INP, HDR]
- Tomas Pajdla [UNIV CTU]
- Elisa Ricci [UNIV TRENTE]
- Pini Tandeitnik [UNIV BAR - ILAN]
- Yihong Xu [VAELO.AI]

2 Overall objectives

In recent years, social robots have been introduced into public spaces, such as museums, airports, commercial malls, banks, show-rooms, schools, universities, hospitals, and retirement homes, to mention a few examples. In addition to classical robotic skills such as navigating in complex environments, grasping and manipulating objects, i.e. *physical interactions*, social robots must be able to communicate with people and to adopt appropriate behavior. Welcoming newcomers, providing various pieces of information, and entertaining groups of people are typical services that social robots are expected to provide in the near future.

Nevertheless, today's state-of-the-art in robotics is not well-suited to fulfill these needs, and there are two main bottlenecks: (i) robots are limited to a handful of simple scenarios which leads to (ii) social robots not being well accepted by a large percentage of users. While there are research programs and projects which have tackled some of these challenges, existing commercially available robots cannot (or only to a very limited extent) recognize individual behaviors (e.g. facial expressions, hand- and body-gestures, head- and eye-gaze) or group behaviors (e.g. who looks at whom, who speaks to whom, who needs robot assistance, etc.). They do not have the ability to take social (or non-verbal) signals into account while they are engaged in spoken dialogue and they cannot connect the dialogue with the persons and objects that are physically present in their surroundings. We would like to develop robots that are responsible for their perception, and act to enhance the quality of the signals they receive, instead of asking the users to adapt their behavior to the robotic platform.

The scientific ambition of ROBOTLEARN is to train robots to acquire the capacity to **look**, **listen**, **learn**, **move** and **speak** in a socially acceptable manner. We identify three main objectives:

1. Develop deep probabilistic models and methods that allow the fusion of audio and visual data, possibly sequential, recorded with cameras and microphones, and in particular with sensors onboard of robots.
2. Increase the performance of human behaviour understanding using deep probabilistic models and jointly exploiting auditory and visual information.
3. Learn robot-action policies that are socially acceptable and that enable robots to better perceive humans and the physical environment.

ROBOTLEARN stands at the cross-roads of several fields: computer vision, audio signal processing, speech technology, statistical learning, deep learning, and robotics. In partnership with several companies (e.g. PAL Robotics and ERM Automatismes Industriels), the technological objective is to launch a brand new generation of robots that are flexible enough to adapt to the needs of the users, and not the other

way around. The experimental objective is to validate the scientific and technological progress in the real world. Furthermore, we believe that ROBOTLEARN will contribute with tools and methods able to process robotic data (perception and action signals) in such a way that connections with more abstract representations (semantics, knowledge) are possible. The developments needed to discover and use such connections could be addressed through collaborations. Similarly, aspects related to robot deployment in the consumer world, such as ethics and acceptability will be addressed in collaboration, for instance, with the Broca day-care hospital in Paris.

From a methodological perspective, the challenge is at least three-fold. First, to reduce the amount of human intervention needed to adapt the designed learning models in a new environment. We aim to further develop strategies based on unsupervised learning and unsupervised domain adaptation, within the framework of deep probabilistic modling with latent variables [49]. Second, to successfully exploit auditory and visual data for human behavior understanding. For instance by developing mechanisms that manage to model and learn the complementarity between sounds and images [9]. Third, by developping reinforcement learning algorithms that can transfer previous knowledge to future tasks and environments. One potential way forward is to anchor the learning into key features that can be hand-crafted or learned [27].

3 Research program

ROBOTLEARN will be structured in three research axes, allowing to develop socially intelligent robots. First, on deep probabilistic models, which include the large family of deep neural network architectures, the large family of probabilistic models, and their intersection. Briefly, we will investigate how to jointly exploit the representation power of deep network together with the flexibility of probabilistic models. A well-known example of such combination are variational autoencoders. Deep probabilistic models are the methodological backbone of the proposed projet, and set the foundations of the two other research axes. Second, we will develop methods for the automatic understanding of human behavior from both auditory and visual data. To this aim we will design our algorithms to exploit the complementary nature of these two modalities, and adapt their inference and on-line update procedures to the computational resources available when operating with robotic platforms. Third, we will investigate models and tools allowing a robot to automatically learn the optimal social action policies. In other words, learn to select the best actions according to the social environment. Importantly, these action policies should also allow us to improve the robotic perception, in case this is needed to better understand the ongoing interaction. We believe that these two research axes, grounded on deep and probabilistic models, will ultimately enable us to train robots to acquire social intelligence, meaning, as discussed in the introduction, the capacity to look, listen, learn, move and speak.

3.1 Deep probabilistic models

A large number of perception and interaction processes require temporal modeling. Consider for example the task of extracting a clean speech signal from visual and audio data. Both modalities live in high-dimensional observation spaces and one challenge is to extract low-dimensional embeddings that encode information in a compact way and to update it over time. These high-dimensional to low-dimensional mappings are nonlinear in the general case. Moreover, audio and visual data are corrupted by various perturbations, e.g. by the presence of background noise which is mixed up with the speech signal uttered by a person of interest, or by head movements that overlap with lip movements. Finally, for robotics applications, the available data is scarce, and datasets captured in other settings can only serve as proxies, thus requiring either adaptation [54] or the use of unsupervised models [42]. Therefore, the problem is manyfold: to extract low-dimensional compact representations from high-dimensional inputs, to disregard useless data in order to retain information that is relevant for the task at hand, to update and maintain reliable information over time, and to do so in without (or with very few) annotated data from the robot.

This class of problems can be addressed in the framework of state-space models (SSMs). In their most general form, SSMs are stochastic nonlinear systems with latent variables. Such a system is composed of a state equation, that describes the dynamics of the latent (or state) variables, and M observation

equations (an observation equation for each sensorial modality m) that predict observations from the state of the system, namely:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{v}_t \qquad \mathbf{y}_t^m = g_m(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{w}_t^m, \forall m \in \{1 \dots M\}, \tag{1}$$

where the latent vector $\mathbf{x} \in \mathbb{R}^L$ evolves according to a nonlinear stationary Markov dynamic model driven by the observed control variable $\mathbf{u}$ and corrupted by the noise $\mathbf{v}$. Similarly, the observed vectors $\mathbf{y}^m \in \mathbb{R}^{D_m}$ are modeled with nonlinear stationary functions of the current state and current input, affected by noise $\mathbf{w}^m$. Models of this kind have been examined for decades and their complexity increases from linear-Gaussian models to nonlinear and non-Gaussian ones. Interestingly, they can also be viewed in the framework of probabilistic graphical models to represent the conditional dependencies between the variables. The objective of an SSM is to infer the sequence of latent variables by computing the posterior distribution of the latent variable, conditioned by the sequence of observations, $p(\mathbf{x}_t|\mathbf{y}_{1:t})$.

When the two functions are linear, the model boils down to a linear dynamical system, that can be learned with an exact Expectation-Maximization (EM) algorithm. Beyond this simple case, non-linearity can be achieved via mixtures of K linear models or more general non-linear (e.g. deep neural) functions. Either case, learning and inference cannot be exact and must be approximated, either by using variational EM algorithms [41, 50, 43, 3], amortized variational inference [49, 24] or a combination of both techniques [26, 17].

We name the larger family of all these methods as Deep Probabilistic Models (DPMs), which form a backbone among the methodological foundations of ROBOTLEARN. Learning DPMs is challenging from the theoretical, methodological and computational points of view. Indeed, the problem of learning, for instance, deep generative Bayesian filters in the framework of nonlinear and non-Gaussian SSMs remains intractable and approximate solutions, that are both optimal from a theoretical point of view and efficient from a computational point of view, remain to be proposed. We plan to investigate both discriminative and generative deep recurrent Bayesian networks and to apply them to audio, visual and audio-visual processing tasks.

Exemplar application: deep probabilistic sequential modeling We have investigated a latent-variable generative model called mixture of dynamical variational autoencoders (MixDVAE) to model the dynamics of a system composed of multiple moving sources. A DVAE model is pre-trained on a single-source dataset to capture the source dynamics. Then, multiple instances of the pre-trained DVAE model are integrated into a multi-source mixture model with a discrete observation-to-source assignment latent variable. The posterior distributions of both the discrete observation-to-source assignment variable and the continuous DVAE variables representing the sources content/position are estimated using the variational expectation-maximization algorithm, leading to multi-source trajectories estimation. We illustrated the versatility of the proposed MixDVAE model on two tasks: a computer vision task, namely multi-object tracking, and an audio processing task, namely single-channel audio source separation. Consequently, this mixture models allows to mix different non-linear source models within the maximum likelihood umbrella and combine the model with other probabilistic models as well.

3.2 Human behavior understanding

Interactions between a robot and a group of people require human behavior understanding (HBU) methods. Consider for example the tasks of detecting eye-gaze and head-gaze and of tracking the gaze directions associated with a group of participants. This means that, in addition to gaze detection and gaze tracking, it is important to detect persons and to track them as well. Additionally, it is important to extract segments of speech, to associate these segments with persons and hence to be able to determine over time who looks to whom and who is the speaker and who are the listeners. The temporal and spatial fusion of visual and audio cues stands at the basis of understanding social roles and of building a multimodal conversational model.

Performing HBU tasks in complex, cluttered and noisy environments is challenging for several reasons: participants come in an out of the camera field of view, their photometric features, e.g. facial texture, clothing, orientation with respect to the camera, etc., vary drastically, even over short periods of time, people look at an object of interest (a person entering the room, a speaking person, a TV/computer

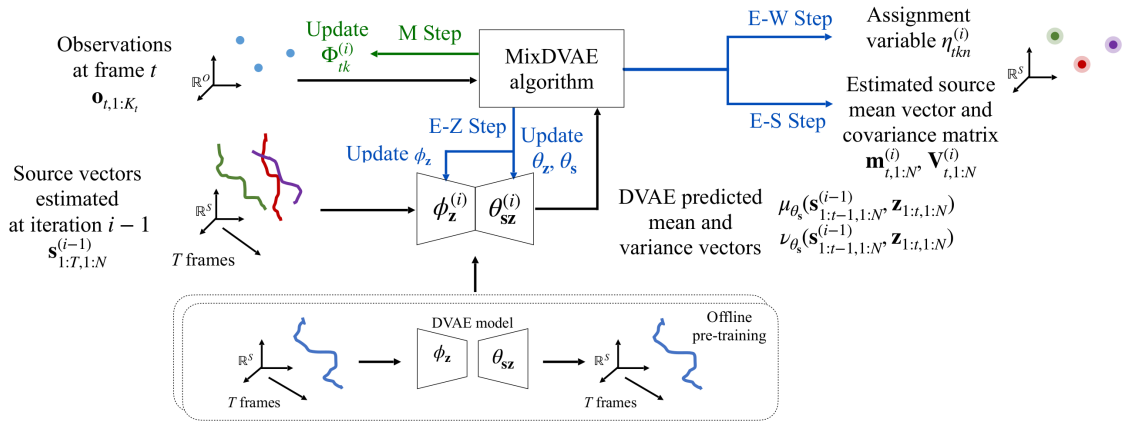

Figure 1: MixDVAE overall diagram.

screen, a wall painting, etc.) by turning their heads away from the camera, hence facial image analysis is difficult, small head movements are often associated with speech which perturbs both lip reading and head-gaze tracking, etc. Clearly, understanding multi-person human-robot interaction is complex because the person-to-person and person-to-object, in addition to person-to-robot, interactions must explicitly be taken into account.

We propose to perform audio-visual HBU by taking explicitly into account the complementary nature of these two modalities. Differently from one current trend in AV learning [40, 46, 48], we opt for unsupervised probabilitic methods that can (i) assign observations to persons without supervision, (ii) be combined with various probabilistic noise models and (iii) and fuse various cues depending on their availability in time (i.e. handle missing data). Indeed, in face-to-face communication, the robot must choose with who it should engage dialog, e.g. based on proximity, eye gaze, head movements, lip movements, facial expressions, etc., in addition to speech. Unlike in the single-user human-robot interaction case, it is crucial to associate temporal segments of speech to participants, referred to as speech diarization. Under such scenarios, speech signals are perturbed by noise, reverberation and competing audio sources, hence speech localization and speech enhancement methods must be used in conjunction with speech recognition.

It is also necessary to perform some kind of adaptation to the distribution of the particular data at hand, e.g. collected with robot sensors. If these data are available in advance, off-line adaptation can be done, otherwise the adaptation needs to be performed on-line or at run time. Such strategies will be useful given the particular experimental conditions of practical human-robot interaction scenarios. Either way we will need some sort of on-line learning to perform final adaptation. On-line learning based on deep neural networks is far from being well understood. We plan to thoroughly study the incorporation of on-line learning into both Bayesian and discriminative deep networks. In the practical case of interaction, real-time processing is crucial. Therefore, a compromise must be found between the size of the network, its discriminative power and the computational cost of the learning and prediction algorithms. Clearly, there is no single solution given the large variety of problems and scenarios that are encountered in practice.

Exemplar application: expression-preserving face frontalization Face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed one. We proposed a frontalization methodology that preserves non-rigid facial deformations in order to boost the performance of visually assisted speech communication. The method alternates between the estimation of (i) the rigid transformation (scale, rotation, and translation) and (ii) the non-rigid deformation between an arbitrarily-viewed face and a face model. The method has two important merits: it can deal with non-Gaussian errors in the data and it incorporates a dynamical face deformation model. For that purpose, we used the generalized Student t-distribution in combination with a linear dynamic system in order to account for both rigid head motions and time-varying facial deformations caused by speech production. We proposed to use the

zero-mean normalized cross-correlation (ZNCC) score to evaluate the ability of the method to preserve facial expressions. We showed that the method, when incorporated into deep learning pipelines, namely lip reading and speech enhancement, improves word recognition and speech intelligibility scores by a considerable margin.

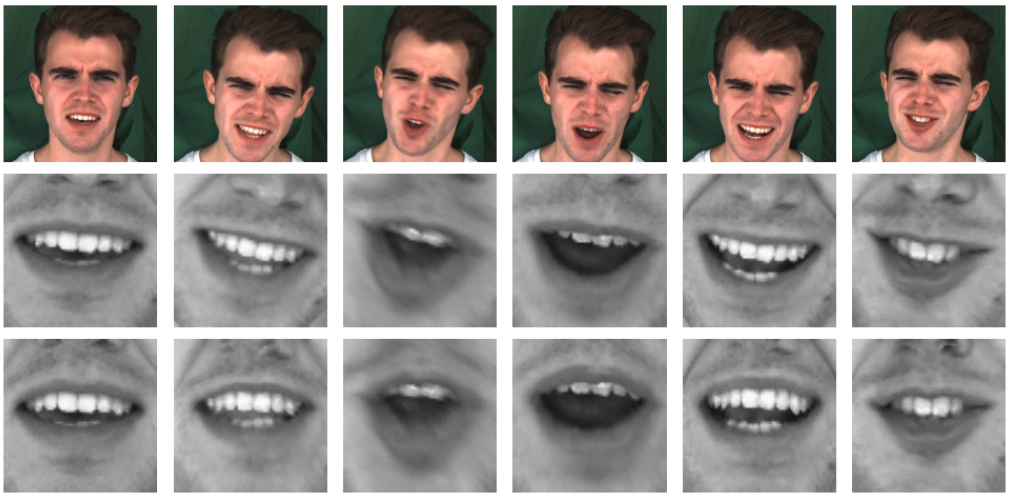

Figure 2: Some results of the proposed expression-preserving face frontalization method.

3.3 Learning and control for social robots

Traditionally, research on human-robot interaction focused on single-person scenarios also called dyadic interactions. However, over the past decade several studies were devoted to various aspects of *multi-party* interactions, meaning situations in which a robot interacts with a group of two or more people [51]. This line of research is much more challenging because of two main reasons. First, the behavioral cues of each individual and of the group need to be faithfully extracted (and assigned to each individual). Second, the behavioral dynamics of groups of people can be pushed by the presence of the robot towards competition [45] or even bullying [44]. This is why some studies restrict the experimental conditions to very controlled collaborative scenarios, often lead by the robot, such as quiz-like game playing [53] or very specific robot roles [47]. Intuitively, constraining the scenario also reduces the gesture variabilty and the overall interaction dynamics, leading to methods and algorithms with questionable generalisation to free and natural social multi-party interactions.

Whenever a robot participates in such multi-party interactions, it must perform *social actions*. Such robot social actions are typically associated with the need to perceive a person or a group of persons in an optimal way as well as to take appropriate decisions such as to safely move towards a selected group, to pop into a conversation or to answer a question. Therefore, one can distinguish between two types of robot social actions: (i) *physical actions* which correspond to synthesizing appropriate motions using the robot actuators (motors), possibly within a sensorimotor loop, so as to enhance perception and maintain a natural interaction and (ii) *spoken actions* which correspond to synthesizing appropriate speech utterances by a spoken dialog system. In ROBOTLEARN we will focus on the former, and integrate the latter via collaborations with research groups having with established expertise in speech technologies.

In this regard we face three problems. First, given the complexity of the environment and the inherent limitations of the robot's perception capabilities, e.g. limited camera field of view, cluttered spaces, complex acoustic conditions, etc., the robot will only have access to a partial representation of the environment, and up to a certain degree of accuracy. Second, for learning purposes, there is no easy way to annotate which are the best actions the robot must choose given a situation: supervised methods are therefore not an option. Third, since the robot cannot learn from scratch by random exploration in a new environment, standard model-free RL approaches cannot be used. Some sort of previous knowledge

on the environment or a similar one should be exploited. Finally, given that the robot moves within a populated environment, it is desirable to have the capability to enforce certain constrains, thus limiting the range of possible robot actions.

Building algorithms to endow robots with autonomous decision taking is not straightforward. Two relatively distinct paradigms are available the literature. First, one can devise customized strategies based on techniques such as *robot motion planning* combined with *sensor-based robot control*. These techniques lack generalization, in particular when the robot acts in complex, dynamic and unconstrained environments. Second, one can let the robot devise its own strategies based on *reinforcement learning* (RL) – a machine learning paradigm in which “agents" learn by themselves by trial and error to achieve successful strategies[52]. It is very difficult, however, to enforce any kind of soft- or hard-constraint within this framework. We will showcase these two scientific streams with one group of techniques for each one: *model predictive control* (MPC) and Q-learning, *deep Q-networks* (DQNs), more precisely. These two techniques are promising. Moreover, they are well documented in the robotics and machine learning. Nevertheless, combining them is extremely challenging.

An additional challenge, independent from the learning and control combination foreseen, is the data distribution gap between the simulations and the real-world. Meta-learning, or the ability to learn how to learn, can provide partial answers to this problem. Indeed, developing machine learning methods able to understand how the learning is achieved can be used to extend this learning to a new task and speed up the learning process on the new task. Recent developments proposed meta-learning strategies specifically conceived for reinforcement learning, leading to Meta-RL methods. One promising trend in Meta-RL is to have a probabilistic formulation involving SSMs and VAEs, i.e. hence sharing the methodology based on dynamical variational autoencoders described before. Very importantly, we are not aware of any studies able to combine Meta-RL with MPC to handle the constraints, and within a unified formulation. From a methodological perspective, this is an important challenge we face in the next few years.

Exemplar application: transfering poilicies via successor feature representations Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor Representations (SR) and their extension Successor Features (SF) are prominent transfer mechanisms in domains where reward functions change between tasks. They reevaluate the expected return of previously learned policies in a new target task to transfer their knowledge. The SF framework extended SR by linearly decomposing rewards into successor features and a reward weight vector allowing their application in high-dimensional tasks. But this came with the cost of having a linear relationship between reward functions and successor features, limiting its application to tasks where such a linear relationship exists. We proposed a novel formulation of SR based on learning the cumulative discounted probability of successor features, called Successor Feature Representations (SFR). Crucially, SFR allows to reevaluate the expected return of policies for general reward functions. We introduced different SFR variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on SFR with function approximation demonstrate its advantage over SF not only for general reward functions, but also in the case of linearly decomposable reward functions.

4 Application domains

For the last decades, there has been an increasing interest in robots that cooperate and communicate with people. As already mentioned, we are interested *Socially Assistive Robots* (SARs) that can communicate with people and that are perceived as social entities. So far, the humanoid robots developed to fill this role are mainly used as research platforms for human-robot collaboration and interaction and their prices, if at all commercially available, are in the 6-digit-euro category, e.g. 250,000€ for the iCub robot and Romeo humanoid robots, developed by the Italian Institute of Technology and SoftBank Robotics Europe, respectively, as well as the REEM-C and TALOS robots from PAL Robotics. A notable exception being the NAO robot which is a humanoid (legged) robot, available at an affordable price. Apart from humanoid robots, there are also several companion robots manufactured in Europe and available at a much lower price (in the range 10,000–30,000 €) that address the SAR market. For example, the Kompaï, the TIAGo, and the Pepper robots are wheeled indoor robotic platforms. The user interacts with these robots via touch screen and voice commands. The robots manage shopping lists, remember appointments, play

music, and respond to simple requests. These affordable robots (Kompaï, TIAGo, NAO, and Pepper) rapidly became the platforms of choice for many researchers in cognitive robotics and in HRI, and they have been used by many EU projects, e.g. HUMAVIPS, EARS, VHIA, and ENRICHEME.

When interacting, these robots rely on a few selected modalities. The voice interface of this category of robots, e.g. Kompaï, NAO, and Pepper, is based on speech recognition similar to speech technologies used by smart phones and table-top devices, e.g. Google Home. *Their audio hardware architecture and software packages are designed to handle single-user face-to-face spoken dialogue based on keyword spotting, but they can neither perform multiple sound-source analysis, fuse audio and visual information for more advanced multi-modal/multi-party interactions, nor hold a conversation that exceeds a couple of turns and that is out of very narrow predefined domain.*

To the best of our knowledge, the only notable efforts to overcome some of the limitations mentioned above are the FP7 EARS and H2020 MuMMER projects. The EARS project's aim was to redesign the microphone-array architecture of the commercially available humanoid robot NAO, and to build a robot head prototype that can support software based on advanced multi-channel audio signal processing. The EARS partners were able to successfully demonstrate the usefulness of this microphone array for speech-signal noise reduction, dereverberation, and multiple-speaker localisation. Moreover, the recent IEEE-AASP Challenge on Acoustic Source Localisation and Tracking (LOCATA) comprises a dataset that uses this microphone array. The design of NAO imposed severe constraints on the physical integration of the microphones and associated hardware. Consequently and in spite of the scientific and practical promises of this design, SoftBank Robotics has not integrated this technology into their commercially available robots NAO and Pepper. In order to overcome problems arising from human-robot interaction in unconstrained environments and open-domain dialogue on the Pepper robot, the H2020 MuMMER project aimed to deploy an entertaining and helpful robot assistant to a shopping mall. While they had initial success with short deployments of the robot to the mall, they were not specifically addressing the issues arising from multi-party interaction: Pepper's audio hardware/software design cannot locate and separate several simultaneously emitting speech sources.

To conclude, *current robotic platforms available in the consumer market, i.e. with large-scale deployment potential, are neither equipped with the adequate hardware nor endowed with the appropriate software required for multi-party social interactions in real-world environments.*

In the light of the above discussion, the partners of the H2020 SPRING project decided to build a robot prototype well suited for socially assistive tasks and shared by the SPRING partners as well as by other EU projects. We participated to the specifications of the ARI robot prototype (shown on the right), designed, developed and manufactured by PAL Robotics, an industrial partner of the SPRING project. ARI is a ROS-enabled, non-holonomic, differential-drive wheeled robot, equipped with a pan and tilt head, with both color and depth cameras and with a microphone array that embeds the latest audio signal processing technologies. Seven ARI robot units were delivered to the SPRING partners in April 2021.

We are committed to implement our algorithms and associated software packages onto this advanced robotic platform, from low-level control to high-level perception, interaction and planning tasks, such that the robot has a socially-aware behaviour while it safely navigates in an ever changing environment. We will experiment in environments of increasing complexity, e.g. our robotic lab, the Inria Grenoble cafeteria and Login exhibition, as well as the Broca hospital in Paris. The expertise that the team's engineers and researchers have acquired for the last decade would be crucial for present and future robotic developments and experiments.

5 Social and environmental responsibility

5.1 Impact of research results

Our line of research on developing unsupervised learning methods exploiting audio-visual data to understand social scenes and to learn to interact within is very interesting and challenging, and has large economical and societal impact. Economical impact since the auditory and visual sensors are the most common one, and we can find (many of) them in almost every smartphone in the market. Beyond telephones, manufacturers designing new systems meant for human use, should take into account the need for verbal interaction, and hence for audio-visual perception. A clear example of this potential is

Figure 3: The ARI robot from PAL Robotics.

the transfer of our technology to a real robotic platform, for evaluation within a day-care hospital (DCH). This is possible thanks to the H2020 SPRING EU project, that assesses the interest of social robotics in the non-medical phases of a regular day for elder patients in a DCH. We are evaluating the performance of our methods for AV speaker tracking, AV speech enhancement, and AV sound source separation, for future technology transfer to the robot manufacturer. This is the first step toward a robot that can be part of the social environment of the DCH, helping to reduce patient and companion stress, at the same time as being a useful tool for the medical personnel. We are confident that developing robust AV perception and action capabilities for robots and autonomous systems, will make them more suitable for environments populated with humans.

6 Highlights of the year

6.1 Team Evaluation

Our team was evaluated this year and received very positive feedback. In the "Main achievements" section the reviewers said:

- The work of the RobotLearn team is well-received by the community as testified by some milestone publications at top venues. The review on DVAEs currently has 129 citations [...] These records indicate that the team has been significantly contributing to the progress of the field over the years.
- The results obtained by the team are significant and are expected to have a considerable impact in broader computer vision and signal processing fields. [...] The relevance of the results produced is significant for the domains covered by the team and are the outcome of an appropriate risk taking.
- The first two topics (axis) build upon the previous experience available in the team and were explored in depth and breadth. The the third one is a recent activity line [...] very promising results. In this respect, the team appears to be capable to consolidate and reinforce its existing expertise, while still opening new lines of investigation.

In the "International Standing and Collaborations in the Field" section, they mention:

- The team has a good number of fruitful national and international collaborations [...]
- RobotLearn has produced significant results [...] that have given the team a substantial visibility [...] complemented by the contribution of the team to international activities.
- The achievements above suggest that the team [...] recognized as leading by other research groups. All observable indicators suggest that [...] RobotLearn is likely to become one of the key-actors in its area of expertise.

6.2 Keynote Speaker at ACM Multimedia 2023

Xavier Alameda-Pineda was invited to give a Keynote Talk at ACM Multimedia 2023, in Ottawa, on the topic of "Variational Audio-Visual Representation Learning".

6.3 PhD Defences

6.3.1 Wen Guo

Wen defended her PhD on "Multi-person Pose Understanding in Complex Physical Interactions." Understanding the pose and motion of humans in 3D space has undergone enormous progress in recent years. The majority of studies in multi-person scenarios view individuals as separate instances, neglecting the importance of interaction information. However, in numerous everyday contexts, people always engage with one another, and it is essential to consider pose instances jointly as the pose of an individual is influenced by the poses of their interactees.In this context, this thesis aims to develop deep learning techniques to understand human pose and motion in complex interactions and leverage interaction information to enhance the performance of human pose estimation and human motion prediction.Our

investigation encompasses modeling and learning interaction information, leveraging this information to refine human pose and motion estimation, and addressing the issue of insufficient 3D interacting human datasets.To overcome these challenges, we undertake the following steps: (1) we verified the feasibility of considering person interaction on the task of 3D human pose estimation from a single RGB image, by modeling and learning the interaction information with a deep network (PI-Net) on publicly available datasets; (2) we collected and released a new dataset for extreme interacting poses (ExPI dataset) to study person interaction; (3) observing poses as temporal sequences, we study the task of collaborative motion prediction and propose a model with cross-interaction attention (XIA), using interaction information as guidance to improve multi-person motion prediction; (4) rethinking the task of human motion prediction, we further propose a simple and effective baseline model for human motion prediction, which works not only on single-person motion prediction but also on multi-person scenarios. The code of our work is publicly available.

6.3.2 Xiaoyu Bie

Xiaoyu defended his PhD on "Dynamical Variational Autoencoders for Multimedia Processing". The overall goal of the PhD project is to develop machine learning methodologies for robust automatic perception of multi-person conversational scenes. In a nutshell, 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. 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 hybrid probabilistic/deep learning models.

6.3.3 Louis Airale

Louis defended his PhD on "Adversarial learning methods for the generation of human interaction data." The objective of this thesis work is to explore new deep generative model architectures for diverse human interaction data generation tasks. The applications for such systems are various: social robotics, animation or entertainment, but always pertain to building more natural interactive systems between humans and machines. Owing to their astonishing performances on a wide range of applications, deep generative models offer an ideal framework to address this task. In return, one can learn how to improve the training of such models by adjusting them to tackle the challenges and constraints posed by human interaction data generation. In this thesis, we consider three generation tasks, corresponding to as many target modalities or conditioning signals. Interactions are first modeled as sequences of discrete, high-level actions simultaneously achieved by a free number of participants. Then, we consider the continuous facial dynamics of a conversing individual and attempt to produce realistic animations from a single reference frame in the facial landmark domain. Finally, we address the task of co-speech talking face generation, where the aim is to correlate the output head and lips motion with an input speech signal. Interestingly, similar deep generative models based on autoregressive adversarial networks provide state-of-the-art results on these otherwise slightly related tasks.

6.4 Commitments

Xavier Alameda-Pineda has been appointed General co-Chair of ACM Multimedia 2026, that will be hosted at New Jersey Institute of Technology, in Newark.

7 New software, platforms, open data

7.1 New software

7.1.1 xi_learning

Name: Successor Feature Representation Learning

Keywords: Reinforcement learning, Transfer Learning

Functional Description: Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor features (SF) are a prominent transfer mechanism in domains where the reward function changes between tasks. They reevaluate the expected return of previously learned policies in a new target task and to transfer their knowledge. A limiting factor of the SF framework is its assumption that rewards linearly decompose into successor features and a reward weight vector. We propose a novel SF mechanism, ξ-learning, based on learning the cumulative discounted probability of successor features. Crucially, ξ-learning allows to reevaluate the expected return of policies for general reward functions. We introduce two ξ-learning variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on ξ-learning with function approximation demonstrate the prominent advantage of ξ-learning over available mechanisms not only for general reward functions, but also in the case of linearly decomposable reward functions.

URL: https://gitlab.inria.fr/robotlearn/xi_learning

Authors: Chris Reinke, Xavier Alameda Pineda

Contact: Chris Reinke

7.1.2 Social MPC

Functional Description: A library for controlling a social robot. This library allows a non-holonomic robot to navigate in a crowded environment using model predictive control and social force models. This library has been developed for the SPRING project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871245.

The main components of this library are: - A module to determine optimal positioning of a robot in a group, using methods from the litterature. - A navigation component to compute optimal paths - The main module, implementing a model predictive controller using the Jax library to determine optimal commands to steer the robot

Authors: Alex Auternaud, Timothee Wintz, Chris Reinke

Contact: Alex Auternaud

7.1.3 2D Social Simulator

Keywords: Simulator, Robotics

Functional Description: A python based simulator using Box2D allowing a robot to interact with people. This software enables: - The configuration of a scene with physical obstacles and people populating a room - The simulation of the motion of a robot in this space - Social force models for the behaviour of people, groups between themselves and in reaction to the motion of the robot

Rendering is done using PyGame and is optional (headless mode is possible).

A gym environment is provided for reinforcement learning.

URL: https://gitlab.inria.fr/spring/wp6_robot_behavior/2D_Simulator

Authors: Alex Auternaud, Timothee Wintz, Chris Reinke

Contact: Alex Auternaud

7.1.4 dvae-speech

Name: dynamic variational auto-encoder for speech re-synthesis

Keywords: Variational Autoencoder, Deep learning, Pytorch, Speech Synthesis

Functional Description: It can be considered a library for speech community, to use different dynamic VAE models for speech re-synthesis (potentially for other speech application)

URL: https://github.com/XiaoyuBIE1994/DVAE-speech

Publication: hal-02926215

Authors: Xiaoyu Bie, Xavier Alameda Pineda, Laurent Girin

Contact: Xavier Alameda Pineda

7.1.5 exputils

Name: experiment utilities

Keywords: Python, Toolbox, Computer Science

Functional Description: Experiment Utilities (exputils) contains various tools for the management of scientific experiments and their experimental data. It is especially designed to handle experimental repetitions, including to run different repetitions, to effectively store and load data for them, and to visualize their results.

Main features: Easy definition of default configurations using nested python dictionaries. Setup of experimental configuration parameters using an ODF file. Running of experiments and their repetitions in parallel. Logging of experimental data (numpy, json). Loading and filtering of experimental data. Interactive Jupyter widgets to load, select and plot data as line, box and bar plots.

URL: https://gitlab.inria.fr/creinke/exputils

Authors: Chris Reinke, Gaetan Lepage

Contact: Chris Reinke

7.1.6 MixDVAE

Name: Source code for the article "Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation"

Keywords: Variational Autoencoder, Finite mixture

Functional Description: In this paper, we propose a latent-variable generative model called mixture of dynamical variational autoencoders (MixDVAE) to model the dynamics of a system composed of multiple moving sources. A DVAE model is pre-trained on a single-source dataset to capture the source dynamics. Then, multiple instances of the pre-trained DVAE model are integrated into a multi-source mixture model with a discrete observation-to-source assignment latent variable. The posterior distributions of both the discrete observation-to-source assignment variable and the continuous DVAE variables representing the sources content/position are estimated using a variational expectation-maximization algorithm, leading to multi-source trajectories estimation. We illustrate the versatility of the proposed MixDVAE model on two tasks: a computer vision task, namely multi-object tracking, and an audio processing task, namely single-channel audio source separation. Experimental results show that the proposed method works well on these two tasks, and outperforms several baseline methods.

Contact: Xiaoyu Lin

7.1.7 Light-DVAE

Keyword: Variational Autoencoder

Functional Description: The dynamical variational autoencoders (DVAEs) are a family of latent-variable deep generative models that extends the VAE to model a sequence of observed data and a corresponding sequence of latent vectors. In almost all the DVAEs of the literature, the temporal dependencies within each sequence and across the two sequences are modeled with recurrent neural networks. In this paper, we propose to model speech signals with the Hierarchical Transformer DVAE (HiT-DVAE), which is a DVAE with two levels of latent variable (sequence-wise and frame-wise) and in which the temporal dependencies are implemented with the Transformer architecture. We show that HiT-DVAE outperforms several other DVAEs for speech spectrogram modeling, while enabling a simpler training procedure, revealing its high potential for downstream low-level speech processing tasks such as speech enhancement.

URL: https://gitlab.inria.fr/robotlearn/light-dvae

Contact: Xiaoyu Lin

7.1.8 DDGM-SE

Keywords: Speech processing, Generative Models

Functional Description: This work builds on a previous work on unsupervised speech enhancement using a dynamical variational autoencoder (DVAE) as the clean speech model and non-negative matrix factorization (NMF) as the noise model. We propose to replace the NMF noise model with a deep dynamical generative model (DDGM) depending either on the DVAE latent variables, or on the noisy observations, or on both. This DDGM can be trained in three configurations: noise-agnostic, noise-dependent and noise adaptation after noise-dependent training. Experimental results show that the proposed method achieves competitive performance compared to state-of-the-art unsupervised speech enhancement methods, while the noise-dependent training configuration yields a much more time-efficient inference process.

URL: https://gitlab.inria.fr/robotlearn/light-dvae

Contact: Xiaoyu Lin

8 New results

The new results listed below are organised by research axis.

8.1 Deep Probabilistic Models

8.1.1 Unsupervised speech enhancement with deep dynamical generative speech and noise models

Participants: Xiaoyu Lin, Laurent Girin, Xavier Alameda-Pineda.

This work builds on a previous work on unsupervised speech enhancement using a dynamical variational autoencoder (DVAE) as the clean speech model and non-negative matrix factorization (NMF) as the noise model. We propose to replace the NMF noise model with a deep dynamical generative model (DDGM) depending either on the DVAE latent variables, or on the noisy observations, or on both. This DDGM can be trained in three configurations: noise-agnostic, noise-dependent and noise adaptation after noise-dependent training. Experimental results show that the proposed method achieves competitive performance compared to state-of-the-art unsupervised speech enhancement methods, while the noise-dependent training configuration yields a much more time-efficient inference process.

8.1.2 Speech Modeling with a Hierarchical Transformer Dynamical VAE

Participants: Xiaoyu Lin, Xiaoyu Bie, Laurent Girin, Xavier Alameda-Pineda.

The dynamical variational autoencoders (DVAEs) are a family of latent-variable deep generative models that extends the VAE to model a sequence of observed data and a corresponding sequence of latent vectors. In almost all the DVAEs of the literature, the temporal dependencies within each sequence and across the two sequences are modeled with recurrent neural networks. In this paper, we propose to model speech signals with the Hierarchical Transformer DVAE (HiT-DVAE), which is a DVAE with two levels of latent variable (sequence-wise and frame-wise) and in which the temporal dependencies are implemented with the Transformer architecture. We show that HiT-DVAE outperforms several other DVAEs for speech spectrogram modeling, while enabling a simpler training procedure, revealing its high potential for downstream low-level speech processing tasks such as speech enhancement.

8.1.3 Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation

Participants: Xiaoyu Lin, Laurent Girin, Xavier Alameda Pineda.

In this paper, we present an unsupervised probabilistic model and associated estimation algorithm for multi-object tracking (MOT) based on a dynamical variational autoencoder (DVAE), called DVAE-UMOT. The DVAE is a latent-variable deep generative model that can be seen as an extension of the variational autoencoder for the modeling of temporal sequences. It is included in DVAE-UMOT to model the objects' dynamics, after being pre-trained on an unlabeled synthetic dataset of single-object trajectories. Then the distributions and parameters of DVAE-UMOT are estimated on each multi-object sequence to track using the principles of variational inference: Definition of an approximate posterior distribution of the latent variables and maximization of the corresponding evidence lower bound of the data likehood function. DVAE-UMOT is shown experimentally to compete well with and even surpass the performance of two state-of-the-art probabilistic MOT models. Code and data are publicly available.

8.2 Human Behavior Understanding

8.2.1 Continual Attentive Fusion for Incremental Learning in Semantic Segmentation

Participants: Elisa Ricci, Xavier Alameda-Pineda.

Over the past years, semantic segmentation, as many other tasks in computer vision, benefited from the progress in deep neural networks, resulting in significantly improved performance. However, deep architectures trained with gradient-based techniques suffer from catastrophic forgetting, which is the tendency to forget previously learned knowledge while learning new tasks. Aiming at devising strategies to counteract this effect, incremental learning approaches have gained popularity over the past years. However, the first incremental learning methods for semantic segmentation appeared only recently. While effective, these approaches do not account for a crucial aspect in pixel-level dense prediction problems, i.e. the role of attention mechanisms. To fill this gap, in this paper we introduce a novel attentive feature distillation approach to mitigate catastrophic forgetting while accounting for semantic spatial- and channel-level dependencies. Furthermore, we propose a continual attentive fusion structure, which takes advantage of the attention learned from the new and the old tasks while learning features for the new task. Finally, we also introduce a novel strategy to account for the background class in the distillation loss, thus preventing biased predictions. We demonstrate the effectiveness of our approach with an extensive evaluation on Pascal-VOC 2012 and ADE20K, setting a new state of the art.

8.2.2 Semi-supervised learning made simple with self-supervised clustering

Participants: Pietro Astolfi, Elisa Ricci, Xavier Alameda-Pineda.

Self-supervised models have been shown to produce comparable or better visual representations than their supervised counterparts when trained offline on unlabeled data at scale. However, their efficacy is catastrophically reduced in a Continual Learning (CL) scenario where data is presented to the model sequentially. In this paper, we show that self-supervised loss functions can be seamlessly converted into distillation mechanisms for CL by adding a predictor network that maps the current state of the representations to their past state. This enables us to devise a framework for Continual self-supervised visual representation Learning that (i) significantly improves the quality of the learned representations, (ii) is compatible with several state-of-the-art self-supervised objectives, and (iii) needs little to no hyperparameter tuning. We demonstrate the effectiveness of our approach empirically by training six popular self-supervised models in various CL settings.

8.2.3 Motion-DVAE: Unsupervised learning for fast human motion denoising

Participants: Xavier Alameda Pineda.

Pose and motion priors are crucial for recovering realistic and accurate human motion from noisy observations. Substantial progress has been made on pose and shape estimation from images, and recent works showed impressive results using priors to refine frame-wise predictions. However, a lot of motion priors only model transitions between consecutive poses and are used in time-consuming optimization procedures, which is problematic for many applications requiring real-time motion capture. We introduce Motion-DVAE, a motion prior to capture the short-term dependencies of human motion. As part of the dynamical variational autoencoder (DVAE) models family, Motion-DVAE combines the generative capability of VAE models and the temporal modeling of recurrent architectures. Together with Motion-DVAE, we introduce an unsupervised learned denoising method unifying regression- and optimization-based approaches in a single framework for real-time 3D human pose estimation. Experiments show that the proposed approach reaches competitive performance with state-of-the-art methods while being much faster.

8.2.4 Learning and controlling the source-filter representation of speech with a variational autoencoder

Participants: Laurent Girin, Xavier Alameda Pineda.

Understanding and controlling latent representations in deep generative models is a challenging yet important problem for analyzing, transforming and generating various types of data. In speech processing, inspiring from the anatomical mechanisms of phonation, the source-filter model considers that speech signals are produced from a few independent and physically meaningful continuous latent factors, among which the fundamental frequency f_0 and the formants are of primary importance. In this work, we start from a variational autoencoder (VAE) trained in an unsupervised manner on a large dataset of unlabeled natural speech signals, and we show that the source-filter model of speech production naturally arises as orthogonal subspaces of the VAE latent space. Using only a few seconds of labeled speech signals generated with an artificial speech synthesizer, we propose a method to identify the latent subspaces encoding f_0 and the first three formant frequencies, we show that these subspaces are orthogonal, and based on this orthogonality, we develop a method to accurately and independently control the source-filter speech factors within the latent subspaces. Without requiring additional information such as text or human-labeled data, this results in a deep generative model of speech spectrograms that is conditioned

on f_0 and the formant frequencies, and which is applied to the transformation speech signals. Finally, we also propose a robust f_0 estimation method that exploits the projection of a speech signal onto the learned latent subspace associated with f_0 .

8.2.5 Expression-preserving face frontalization improves visually assisted speech processing

Participants: Radu Horaud, Xavier Alameda Pineda.

Face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed one. The main contribution of this paper is a frontalization methodology that preserves non-rigid facial deformations in order to boost the performance of visually assisted speech communication. The method alternates between the estimation of (i) the rigid transformation (scale, rotation, and translation) and (ii) the non-rigid deformation between an arbitrarily-viewed face and a face model. The method has two important merits: it can deal with non-Gaussian errors in the data and it incorporates a dynamical face deformation model. For that purpose, we use the generalized Student t-distribution in combination with a linear dynamic system in order to account for both rigid head motions and time-varying facial deformations caused by speech production. We propose to use the zero-mean normalized cross-correlation (ZNCC) score to evaluate the ability of the method to preserve facial expressions. The method is thoroughly evaluated and compared with several state of the art methods, either based on traditional geometric models or on deep learning. Moreover, we show that the method, when incorporated into deep learning pipelines, namely lip reading and speech enhancement, improves word recognition and speech intelligibilty scores by a considerable margin. Supplemental material is accessible here.

8.2.6 Audio-Visual Speaker Diarization in the Framework of Multi-User Human-Robot Interaction

Participants: Radu Horaud.

The speaker diarization task answers the question “who is speaking at a given time?”. It represents valuable information for scene analysis in a domain such as robotics. In this paper, we introduce a temporal audiovisual fusion model for multiusers speaker diarization, with low computing requirement, a good robustness and an absence of training phase. The proposed method identifies the dominant speakers and tracks them over time by measuring the spatial coincidence between sound locations and visual presence. The model is generative, parameters are estimated online, and does not require training. Its effectiveness was assessed using two datasets, a public one and one collected in-house with the Pepper humanoid robot.

8.2.7 Back to MLP: A Simple Baseline for Human Motion Prediction

Participants: Wen Guo, Xavier Alameda Pineda.

This paper tackles the problem of human motion prediction, consisting in forecasting future body poses from historically observed sequences. State-of-the-art approaches provide good results, however, they rely on deep learning architectures of arbitrary complexity, such as Recurrent Neural Networks (RNN), Transformers or Graph Convolutional Networks (GCN), typically requiring multiple training stages and more than 2 million parameters. In this paper, we show that, after combining with a series of standard practices, such as applying Discrete Cosine Transform (DCT), predicting residual displacement of joints and optimizing velocity as an auxiliary loss, a light-weight network based on multi-layer perceptrons (MLPs) with only 0.14 million parameters can surpass the state-of-the-art performance. An exhaustive evaluation on the Human3.6M, AMASS, and 3DPW datasets shows that our method, named siMLPe, consistently outperforms all other approaches. We hope that our simple method could serve as a strong

baseline for the community and allow re-thinking of the human motion prediction problem. The code is publicly available here.

8.3 Learning and Control for Social Robots

8.3.1 Variational Meta Reinforcement Learning for Social Robotics

Participants: Anand Ballou, Xavier Alameda Pineda, Chris Reinke.

With the increasing presence of robots in our every-day environments, improving their social skills is of utmost importance. Nonetheless, social robotics still faces many challenges. One bottleneck is that robotic behaviors need to be often adapted as social norms depend strongly on the environment. For example, a robot should navigate more carefully around patients in a hospital compared to workers in an office. In this work, we investigate meta-reinforcement learning (meta-RL) as a potential solution. Here, robot behaviors are learned via reinforcement learning where a reward function needs to be chosen so that the robot learns an appropriate behavior for a given environment. We propose to use a variational meta-RL procedure that quickly adapts the robots' behavior to new reward functions. As a result, given a new environment different reward functions can be quickly evaluated and an appropriate one selected. The procedure learns a vectorized representation for reward functions and a meta-policy that can be conditioned on such a representation. Given observations from a new reward function, the procedure identifies its representation and conditions the meta-policy to it. While investigating the procedures' capabilities, we realized that it suffers from posterior collapse where only a subset of the dimensions in the representation encode useful information resulting in a reduced performance. Our second contribution, a radial basis function (RBF) layer, partially mitigates this negative effect. The RBF layer lifts the representation to a higher dimensional space, which is more easily exploitable for the meta-policy. We demonstrate the interest of the RBF layer and the usage of meta-RL for social robotics on four robotic simulation tasks.

8.3.2 Successor Feature Representations

Participants: Chris Reinke, Xavier Alameda Pineda.

Transfer in Reinforcement Learning aims to improve learning performance on target tasks using knowledge from experienced source tasks. Successor Representations (SR) and their extension Successor Features (SF) are prominent transfer mechanisms in domains where reward functions change between tasks. They reevaluate the expected return of previously learned policies in a new target task to transfer their knowledge. The SF framework extended SR by linearly decomposing rewards into successor features and a reward weight vector allowing their application in high-dimensional tasks. But this came with the cost of having a linear relationship between reward functions and successor features, limiting its application to such tasks. We propose a novel formulation of SR based on learning the cumulative discounted probability of successor features, called Successor Feature Representations (SFR). Crucially, SFR allows to reevaluate the expected return of policies for general reward functions. We introduce different SFR variations, prove its convergence, and provide a guarantee on its transfer performance. Experimental evaluations based on SFR with function approximation demonstrate its advantage over SF not only for general reward functions but also in the case of linearly decomposable reward functions.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Visits to international teams

Xavier Alameda-Pineda

Visited institution: International Laboratory on Learning Systems

Country: Montréal, Canada

Dates: Nov 2nd - Nov 9th

Context of the visit: Discussing potential scientific collaborations and common interests.

Mobility program/type of mobility: Research stay.

9.2 European initiatives

9.2.1 H2020 projects

Participants: Alex Auternaud, Gaetan Lepage, Chris Reinke, Soraya Arias, Nicolas Turro, Radu Horaud, Xavier Alameda-Pineda.

Started on Januray 1st, 2020 and finalising on May 31st, 2024, SPRING is a research and innovation action (RIA) with eight partners: Inria Grenoble (coordinator), Università degli Studi di Trento, Czech Technical University Prague, Heriot-Watt University Edinburgh, Bar-Ilan University Tel Aviv, ERM Automatismes Industriels Carpentras, PAL Robotics Barcelona, and Hôpital Broca Paris.. The main objective of SPRING (Socially Pertinent Robots in Gerontological Healthcare) is the development of socially assistive robots with the capacity of performing multimodal multiple-person interaction and open-domain dialogue. In more detail:

- The scientific objective of SPRING is to develop a novel paradigm and novel concept of socially-aware robots, and to conceive innovative methods and algorithms for computer vision, audio processing, sensor-based control, and spoken dialog systems based on modern statistical- and deep-learning to ground the required social robot skills.
- The technological objective of SPRING is to create and launch a brand new generation of robots that are flexible enough to adapt to the needs of the users, and not the other way around.
- The experimental objective of SPRING is twofold: to validate the technology based on HRI experiments in a gerontology hospital, and to assess its acceptability by patients and medical staff.

Website

9.3 National initiatives

9.3.1 ANR JCJC Project ML3RI

Participants: Xiaoyu Lin, Xavier Alameda-Pineda.

Starting on March 1st 2020 and finalising on February 28th 2024, ML3RI is an ANR JCJC that has been awarded to Xavier Alameda-Pineda. Multi-person robot interactionin the wild (i.e. unconstrained and using only the robot's resources) is nowadays unachievable because of the lack of suitable machine perception and decision-taking models. *Multi-Modal Multi-person Low-Level Learning models for Robot Interaction* (ML3RI) has the ambition to develop the capacity to understand and react to low-level

behavioral cues, which is crucial for autonomous robot communication. The main scientific impact of ML3RI is to develop new learning methods and algorithms, thus opening the door to study multi-party conversations with robots. In addition, the project supports open and reproducible research.
Website

9.3.2 ANR MIAI Chair

Participants: Xiaoyu Bie, Anand Ballou, Radu Horaud, Xavier Alameda-Pineda.

The overall goal of the MIAI chair "Audio-visual machine perception & interaction for robots" it to enable socially-aware robot behavior for interactions with humans. Emphasis on unsupervised and weakly supervised learning with audio-visual data, Bayesian inference, deep learning, and reinforcement learning. Challenging proof-of-concept demonstrators. We aim to develop robots that explore populated spaces, understand human behavior, engage multimodal dialog with several users, etc. These tasks require audio and visual cues (e.g. clean speech signals, eye-gaze, head-gaze, facial expressions, lip movements, head movements, hand and body gestures) to be robustly retrieved from the raw sensor data. These features cannot be reliably extracted with a static robot that listens, looks and communicates with people from a distance, because of acoustic reverberation and noise, overlapping audio sources, bad lighting, limited image resolution, narrow camera field of view, visual occlusions, etc. We will investigate audio and visual perception and communication, e.g. face-to-face dialog: the robot should learn how to collect clean data (e.g. frontal faces, signals with high speech-to-noise ratios) and how to react appropriately to human verbal and non-verbal solicitations. We plan to demonstrate these skills with a companion robot that assists and entertains the elderly in healthcare facilities.
Website

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

Area Chair: Xavier Alameda-Pineda was Area Chair of ACM Multimedia 2023, IEEE WASPAA 2023, and ICRA 2024.

10.1.2 Scientific events: selection

Reviewer: The members of the team reviewed for top-tier conferences such as: IEEE IROS 2023, IEEE WACV 2023, IEEE/CVF CVPR 2024, ICRA 2024, and ACM Multimedia 2023, among others.

10.1.3 Journal

Member of the editorial boards: During 2023, Xavier Alameda-Pineda was Associated Editor of four top-tier journals: Computer Vision and Image Understanding, ACM Transactions on Mutimedia Tools, Applications and IEEE Transactions on Multimedia and ACM Transactions on Intelligent Systems and Technology.

Reviewing activities: The members of the team reviewed for top-tier journals such as: IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Trasnactions on Multimedia, IEEE Transactions on Audio, Speech and Language Processing and Applied Intelligence.

10.1.4 Invited talks

Xavier Alameda Pineda was invited for a Keynote Talk at ACM Multimedia 2023 in Ottawa, to discuss recent work on variational representation audio-visual learning.

10.1.5 Leadership within the scientific community

Xavier Alameda-Pineda is member of the Steering Committe of ACM Multimedia (2022-2026) and of the IEEE Audio and Acoustic Signal Processing Technical Committee (2022-2024)

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

In 2023, Xavier Alameda-Pineda was involved and responsible in teaching two coursers at Masters 2 level:

- Learning, Probabilities and Causality - at Master of Science in Industrial and Applied Mathematics
- Advanced Machine Learning, Applications to Vision, Audio and Text - at Master of Science in Informatics at Grenoble

10.2.2 Supervision

Xavier Alameda-Pineda is (co-)supervising Anand Ballou, Louis Airale (defended), Xiaoyu Bie (defended), Xiaoyu Lin, Wen Guo (defended), and Jordan Cosio.

10.2.3 Juries

In 2023, Xavier Alameda-Pineda participated to the following PhD committees as a reviewer (Remi Rigal and Stephanie Tan).

11 Scientific production

11.1 Major publications

[1] L. Airale, D. Vaufreydaz and X. Alameda-Pineda. 'SocialInteractionGAN: Multi-person Interaction Sequence Generation'. In: *IEEE Transactions on Affective Computing* (11th May 2022). DOI: 10.1109/TAFFC.2022.3171719. URL: https://hal.inria.fr/hal-03163467.

[2] Y. Ban, X. Alameda-Pineda, C. Evers and R. Horaud. 'Tracking Multiple Audio Sources with the Von Mises Distribution and Variational EM'. In: *IEEE Signal Processing Letters* 26.6 (1st June 2019), pp. 798–802. DOI: 10.1109/LSP.2019.2908376. URL: https://hal.inria.fr/hal-01969050.

[3] Y. Ban, X. Alameda-Pineda, L. Girin and R. Horaud. 'Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.5 (1st May 2021), pp. 1761–1776. DOI: 10.1109/TPAMI.2019.2953020. URL: https://hal.inria.fr/hal-01950866.

[4] X. Bie, S. Leglaive, X. Alameda-Pineda and L. Girin. 'Unsupervised Speech Enhancement using Dynamical Variational Autoencoders'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 30 (16th Sept. 2022), pp. 2993–3007. DOI: 10.1109/TASLP.2022.3207349. URL: https://hal.inria.fr/hal-03295630.

[5] G. Delorme, Y. Xu, S. Lathuilière, R. Horaud and X. Alameda-Pineda. 'CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification'. In: ICPR 2020 - 25th International Conference on Pattern Recognition. Milano, Italy: IEEE, 2021, pp. 4428–4435. DOI: 10.1109/ICPR48806.2021.9412431. URL: https://hal.inria.fr/hal-02882285.

[6] G. Evangelidis and R. Horaud. 'Joint Alignment of Multiple Point Sets with Batch and Incremental Expectation-Maximization'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6 (1st June 2018), pp. 1397–1410. DOI: 10.1109/TPAMI.2017.2717829. URL: https://hal.inria.fr/hal-01413414.

[7] I. Gebru, S. Ba, X. Li and R. Horaud. 'Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.5 (2nd July 2018), pp. 1086–1099. DOI: 10.1109/TPAMI.2017.2648793. URL: https://hal.inria.fr/hal-01413403.

[8] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber and X. Alameda-Pineda. 'Dynamical Variational Autoencoders: A Comprehensive Review'. In: *Foundations and Trends in Machine Learning* 15.1-2 (2nd Dec. 2021), pp. 1–175. DOI: 10.1561/2200000089. URL: https://hal.inria.fr/hal-02926215.

[9] Z. Kang, M. Sadeghi, R. Horaud and X. Alameda-Pineda. 'Expression-preserving face frontalization improves visually assisted speech processing'. In: *International Journal of Computer Vision* (12th Jan. 2023). DOI: 10.1007/s11263-022-01742-1. URL: https://hal.science/hal-03902610.

[10] S. Lathuilière, B. Massé, P. Mesejo and R. Horaud. 'Neural Network Based Reinforcement Learning for Audio-Visual Gaze Control in Human-Robot Interaction'. In: *Pattern Recognition Letters* 118 (1st Feb. 2019), pp. 61–71. DOI: 10.1016/j.patrec.2018.05.023. URL: https://hal.inria.fr/hal-01643775.

[11] S. Lathuilière, P. Mesejo, X. Alameda-Pineda and R. Horaud. 'A Comprehensive Analysis of Deep Regression'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (1st Sept. 2020), pp. 2065–2081. DOI: 10.1109/TPAMI.2019.2910523. URL: https://hal.inria.fr/hal-01754839.

[12] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud. 'Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments'. In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (8th Mar. 2019), pp. 88–103. DOI: 10.1109/JSTSP.2019.2903472. URL: https://hal.inria.fr/hal-01851985.

[13] X. Li, S. Gannot, L. Girin and R. Horaud. 'Multichannel Identification and Nonnegative Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26.10 (21st May 2018), pp. 1755–1768. DOI: 10.1109/TASLP.2018.2839362. URL: https://hal.inria.fr/hal-01645749.

[14] X. Li, L. Girin, S. Gannot and R. Horaud. 'Multichannel Speech Separation and Enhancement Using the Convolutive Transfer Function'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 27.3 (1st Mar. 2019), pp. 645–659. DOI: 10.1109/TASLP.2019.2892412. URL: https://hal.inria.fr/hal-01799809.

[15] X. Li, S. Leglaive, L. Girin and R. Horaud. 'Audio-noise Power Spectral Density Estimation Using Long Short-term Memory'. In: *IEEE Signal Processing Letters* 26.6 (June 2019), pp. 918–922. DOI: 10.1109/LSP.2019.2911879. URL: https://hal.inria.fr/hal-02100059.

[16] B. Massé, S. Ba and R. Horaud. 'Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.11 (1st Nov. 2018), pp. 2711–2724. DOI: 10.1109/TPAMI.2017.2782819. URL: https://hal.inria.fr/hal-01511414.

[17] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin and R. Horaud. 'Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 28 (30th May 2020), pp. 1788–1800. DOI: 10.1109/TASLP.2020.3000593. URL: https://hal.inria.fr/hal-02364900.

[18] A. Siarohin, G. Zen, C. Majtanovic, X. Alameda-Pineda, E. Ricci and N. Sebe. 'Increasing Image Memorability with Neural Style Transfer'. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 15.2 (1st June 2019). DOI: 10.1145/3311781. URL: https://hal.inria.fr/hal-01858389.

[19] D. Xu, X. Alameda-Pineda, W. Ouyang, E. Ricci, X. Wang and N. Sebe. 'Probabilistic Graph Attention Network with Conditional Kernels for Pixel-Wise Prediction'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.5 (1st May 2022), pp. 2673–2688. DOI: 10.1109/TPAMI.2020.3043781. URL: https://hal.inria.fr/hal-03328687.

[20] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus and X. Alameda-Pineda. 'TransCenter: Transformers With Dense Representations for Multiple-Object Tracking'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (28th Nov. 2022), pp. 1–16. DOI: 10.1109/TPAMI.2022.3225078. URL: https://hal.inria.fr/hal-03906940.

[21] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, M. Nabi, X. Alameda-Pineda and E. Ricci. 'Uncertainty-aware Contrastive Distillation for Incremental Semantic Segmentation'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (31st Mar. 2022), pp. 1–14. DOI: 10.1109/TPAMI.2022.3163806. URL: https://hal.inria.fr/hal-03908664.

[22] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, H. Tang, X. Alameda-Pineda and E. Ricci. 'Continual Attentive Fusion for Incremental Learning in Semantic Segmentation'. In: *IEEE Transactions on Multimedia* (14th Apr. 2022). DOI: 10.1109/TMM.2022.3167555. URL: https://hal.inria.fr/hal-03626393.

11.2 Publications of the year

International journals

[23] L. Airale, X. Alameda-Pineda, S. Lathuilière and D. Vaufreydaz. 'Autoregressive GAN for Semantic Unconditional Head Motion Generation'. In: *ACM Transactions on Multimedia Computing, Communications and Applications* (2024). URL: https://inria.hal.science/hal-03833759.

[24] A. Ballou, X. Alameda-Pineda and C. Reinke. 'Variational Meta Reinforcement Learning for Social Robotics'. In: *Applied Intelligence* (6th Sept. 2023), pp. 1–16. DOI: 10.1007/s10489-023-04691-5. URL: https://inria.hal.science/hal-03908505.

[25] Z. Kang, M. Sadeghi, R. Horaud and X. Alameda-Pineda. 'Expression-preserving face frontalization improves visually assisted speech processing'. In: *International Journal of Computer Vision* 131.5 (1st May 2023), pp. 1122–1140. DOI: 10.1007/s11263-022-01742-1. URL: https://hal.science/hal-03902610.

[26] X. Lin, L. Girin and X. Alameda-Pineda. 'Mixture of Dynamical Variational Autoencoders for Multi-Source Trajectory Modeling and Separation'. In: *Transactions on Machine Learning Research Journal* (2024). URL: https://inria.hal.science/hal-03584014.

[27] C. Reinke and X. Alameda-Pineda. 'Successor Feature Representations'. In: *Transactions on Machine Learning Research Journal* (May 2023), pp. 1–35. URL: https://inria.hal.science/hal-03426870.

[28] M. Sadeghi, X. Alameda-Pineda and R. Horaud. 'Unsupervised Performance Analysis of 3D Face Alignment with a Statistically Robust Confidence Test'. In: *Neurocomputing* 564 (Jan. 2024). DOI: 10.1016/j.neucom.2023.126941. URL: https://hal.science/hal-04265797.

[29] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda and R. Séguier. 'A Multimodal Dynamical Variational Autoencoder for Audiovisual Speech Representation Learning'. In: *Neural Networks* (2024). URL: https://inria.hal.science/hal-04132316.

[30] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda and R. Séguier. 'Learning and controlling the source-filter representation of speech with a variational autoencoder'. In: *Speech Communication* 148 (Mar. 2023), pp. 53–65. DOI: 10.1016/j.specom.2023.02.005. URL: https://hal.science/hal-03650569.

[31] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, H. Tang, X. Alameda-Pineda and E. Ricci. 'Continual Attentive Fusion for Incremental Learning in Semantic Segmentation'. In: *IEEE Transactions on Multimedia* (Nov. 2023), pp. 1–13. DOI: 10.1109/TMM.2022.3167555. URL: https://inria.hal.science/hal-03626393.

International peer-reviewed conferences

[32] T. Dhaussy, B. Jabaian, F. Lefèvre and R. Horaud. 'Audio-Visual Speaker Diarization in the Framework of Multi-User Human-Robot Interaction'. In: ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Ixia-Ialyssos, Greece: IEEE, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096295. URL: https://hal.science/hal-04140076.

[33] G. Fiche, S. Leglaive, X. Alameda-Pineda and R. Séguier. 'Motion-DVAE: Unsupervised learning for fast human motion denoising'. In: ACM SIGGRAPH Conference on Motion, Interaction and Games (ACM MIG). Rennes, France, 2023. DOI: 10.1145/3623264.3624454. URL: https://inria.hal.science/hal-04132314.

[34] E. Fini, P. Astolfi, K. Alahari, X. Alameda-Pineda, J. Mairal, M. Nabi and E. Ricci. 'Semi-supervised learning made simple with self-supervised clustering'. In: CVPR 2023 – IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023, pp. 1–11. URL: https://inria.hal.science/hal-04073630.

[35] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda and F. Moreno-Noguer. 'Back to MLP: A Simple Baseline for Human Motion Prediction'. In: WACV 2023 - IEEE Winter Conference on Applications of Computer Vision. Waikoloa, United States: IEEE, 2023, pp. 1–11. URL: https://inria.hal.science/hal-03906936.

[36] X. Lin, X. Bie, S. Leglaive, L. Girin and X. Alameda-Pineda. 'Speech Modeling with a Hierarchical Transformer Dynamical VAE'. In: ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes, Greece: IEEE, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096751. URL: https://inria.hal.science/hal-04132313.

[37] X. Lin, S. Leglaive, L. Girin and X. Alameda-Pineda. 'Unsupervised speech enhancement with deep dynamical generative speech and noise models'. In: Interspeech 2023 - 24th Annual Conference of the International Speech Communication Association. Dublin, Ireland, Aug. 2023, pp. 1–5. URL: https://inria.hal.science/hal-04132312.

Reports & preprints

[38] L. Airale, D. Vaufreydaz and X. Alameda-Pineda. *A Comprehensive Multi-scale Approach for Speech and Dynamics Synchrony in Talking Head Generation*. 3rd July 2023. URL: https://hal.science/hal-04149083.

[39] A. Golmakani, M. Sadeghi, X. Alameda-Pineda and R. Serizel. *A weighted-variance variational autoencoder model for speech enhancement*. 20th Sept. 2023. URL: https://inria.hal.science/hal-03833827.

11.3 Cited publications

[40] T. Afouras, A. Owens, J. S. Chung and A. Zisserman. 'Self-supervised learning of audio-visual objects from video'. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer. 2020, pp. 208–224.

[41] S. Ba, X. Alameda-Pineda, A. Xompero and R. Horaud. 'An On-line Variational Bayesian Model for Multi-Person Tracking from Cluttered Scenes'. In: *Computer Vision and Image Understanding* 153 (Dec. 2016), pp. 64–76. DOI: 10.1016/j.cviu.2016.07.006. URL: https://hal.inria.fr/hal-01349763.

[42] Y. Ban, X. Alameda-Pineda, F. Badeig, S. Ba and R. Horaud. 'Tracking a Varying Number of People with a Visually-Controlled Robotic Head'. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vancouver, Canada: IEEE, Sept. 2017, pp. 4144–4151. DOI: 10.1109/IROS.2017.8206274. URL: https://hal.inria.fr/hal-01542987.

[43] Y. Ban, X. Alameda-Pineda, C. Evers and R. Horaud. 'Tracking Multiple Audio Sources with the Von Mises Distribution and Variational EM'. In: *IEEE Signal Processing Letters* 26.6 (June 2019), pp. 798–802. DOI: 10.1109/LSP.2019.2908376. URL: https://hal.inria.fr/hal-01969050.

[44] D. Brščić, H. Kidokoro, Y. Suehiro and T. Kanda. 'Escaping from children's abuse of social robots'. In: *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*. 2015, pp. 59–66.

[45] W.-L. Chang, J. P. White, J. Park, A. Holm and S. Šabanović. 'The effect of group size on people's attitudes and cooperative behaviors toward robots in interactive gameplay'. In: *RO-MAN International Symposium on Robot and Human Interactive Communication*. IEEE. 2012, pp. 845–850.

[46] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson and K. Grauman. 'Soundspaces: Audio-visual navigation in 3d environments'. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer. 2020, pp. 17–36.

[47] M. E. Foster, A. Gaschler and M. Giuliani. 'Automatically classifying user engagement for dynamic multi-party human–robot interaction'. In: *International Journal of Social Robotics* 9.5 (2017), pp. 659–674.

[48] R. Gao and K. Grauman. 'Visualvoice: Audio-visual speech separation with cross-modal consistency'. In: *IEEE/CVF CVPR*. 2021.

[49] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber and X. Alameda-Pineda. 'Dynamical Variational Autoencoders: A Comprehensive Review'. In: *Foundations and Trends in Machine Learning* 15.1-2 (Dec. 2021), pp. 1–175. DOI: 10.1561/2200000089. URL: https://inria.hal.science/hal-02926215.

[50] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda and R. Horaud. 'Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments'. In: *IEEE Journal of Selected Topics in Signal Processing* 13.1 (Mar. 2019), pp. 88–103. DOI: 10.1109/JSTSP.2019.2903472. URL: https://hal.inria.fr/hal-01851985.

[51] S. Sebo, B. Stoll, B. Scassellati and M. F. Jung. 'Robots in groups and teams: a literature review'. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–36.

[52] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[53] M. Żarkowski. 'Multi-party turn-taking in repeated human–robot interactions: an interdisciplinary evaluation'. In: *International Journal of Social Robotics* 11.5 (2019), pp. 693–707.

[54] J. Zhang, L. Tai, P. Yun, Y. Xiong, M. Liu, J. Boedecker and W. Burgard. 'Vr-goggles for robots: Real-to-sim domain adaptation for visual control'. In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 1148–1155.