RESEARCH CENTRE

Paris

2021 ACTIVITY REPORT

Team COML

Cognitive Machine Learning

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions)

IN COLLABORATION WITH: Laboratoire de sciences cognitives et psycholinguistique

DOMAIN

Perception, Cognition and Interaction

THEME

Language, Speech and Audio

Contents

Team COML

Creation of the Team: 2017 May 04

Keywords

Computer sciences and digital sciences

A2.5.1. – Software Architecture & Design

A2.5.4. – Software Maintenance & Evolution

A2.5.5. – Software testing

A3.4.2. – Unsupervised learning

A3.4.5. – Bayesian methods

A3.4.6. – Neural networks

A3.4.8. – Deep learning

A5.7. – Audio modeling and processing

A5.7.1. – Sound

A5.7.3. – Speech

A5.7.4. – Analysis

A5.8. – Natural language processing

A6.3.3. – Data processing

A9.2. – Machine learning

A9.3. – Signal analysis

A9.4. – Natural language processing

A9.7. – AI algorithmics

Other research topics and application domains

B1.2. – Neuroscience and cognitive science

B1.2.2. – Cognitive science

B2.2.6. – Neurodegenerative diseases

B2.5.2. – Cognitive disabilities

B9.6.1. – Psychology

B9.6.8. – Linguistics

B9.8. – Reproducibility

B9.10. – Privacy

1 Team members, visitors, external collaborators

Research Scientist

- Justine Cassell [Inria, Advanced Research Position]

Faculty Member

- Emmanuel Dupoux [Team leader, École des hautes études en sciences sociales, Professor, HDR]

Post-Doctoral Fellows

- William Havard [École Normale Supérieure de Paris, from Oct 2021]
- Thomas Janssoone [Inria, from May 2021]
- Paul Michel [École Normale Supérieure de Paris, from Sep 2021]

PhD Students

- Alafate Abulimiti [Inria]
- Robin Algayres [École Normale Supérieure de Paris]
- Rahma Chaabouni [École Normale Supérieure de Paris, until Sep 2021]
- Maureen De Seyssel [École Normale Supérieure de Paris]
- Marvin Lavechin [École Normale Supérieure de Paris]
- Juliette Millet [École Normale Supérieure de Paris]
- Yann Raphalen [Inria]
- Rachid Riad [École Normale Supérieure de Paris]
- Mathieu Rita [Inria, from Feb 2021]

Technical Staff

- Mathieu Bernard [Inria, Engineer]
- Xuan Nga Cao [École des hautes études en sciences sociales, Engineer]
- Nicolas Hamilakis [École Normale Supérieure de Paris, Engineer]
- Manel Khentout [École Normale Supérieure de Paris, Engineer]
- Marianne Metais [École Normale Supérieure de Paris, Engineer]
- Biswesh Mohapatra [Inria, Engineer, from Sep 2021]
- Robin San Roman [École Normale Supérieure de Paris, Engineer, from Nov 2021]
- Valentin Taillandier [École Normale Supérieure de Paris, Engineer, Dec 2021]
- Hadrien Titeux [École Normale Supérieure de Paris, Engineer]
- Gwendal Virlet [École Normale Supérieure de Paris, Engineer]

Interns and Apprentices

- Deniz Baran Aslan [Inria, from Feb 2021 until Apr 2021]
- Louis Bard [Inria, from Apr 2021 until Sep 2021]
- Victoria Brami [École Normale Supérieure de Paris, from Mar 2021 until Jul 2021]
- Hazal Celik Burle [Inria, until Jul 2021]
- Leopold Favre [École Normale Supérieure de Paris, from May 2021 until Sep 2021]
- Gustav Grimberg [Inria, from Jun 2021]
- Adoracion Guzman Garcia [Inria, from Feb 2021 until Sep 2021]
- Hugo Laurençon [Inria, from Apr 2021 until Sep 2021]
- Luca Leisten [Inria, from Jul 2021 until Sep 2021]
- Clement Nguyen [École Normale Supérieure de Paris, from Apr 2021 until Sep 2021]
- Patricia Roze [École Normale Supérieure de Paris, from Feb 2021 until Jun 2021]
- Sarah Said [Inria, from May 2021 until Sep 2021]
- Andrea Santos Revilla [École Normale Supérieure de Paris, from Jul 2021]
- Aliah Zewail [Inria, from Jun 2021 until Aug 2021]

Administrative Assistants

- Meriem Guemair [Inria]
- Catherine Urban [École Normale Supérieure de Paris, until Oct 2021]

Visiting Scientist

- Gustav Grimberg [École Normale Supérieure de Paris, from Feb 2021 until Jun 2021]

External Collaborators

- Ewan Dunbar [University of Toronto]
- Abdellah Fourtassi [Univ de Provence, from Jun 2021 until Aug 2021]

2 Overall objectives

Brain-inspired machine learning algorithms combined with big data have recently reached spectacular results, equalling or beating humans on specific high level tasks (e.g. the game of go). However, there are still a lot of domains in which even humans infants outperform machines: unsupervised learning of rules and language, common sense reasoning, and more generally, cognitive flexibility (the ability to quickly transfer competence from one domain to another one).

The aim of the Cognitive Computing team is to *reverse engineer* such human abilities, i.e., to construct effective and scalable algorithms which perform as well (or better) than humans, when provided with similar data, study their mathematical and algorithmic properties and test their empirical validity as models of humans by comparing their output with behavioral and neuroscientific data. The expected results are more adaptable and autonomous machine learning algorithm for complex tasks, and quantitative models of cognitive processes which can used to predict human developmental and processing data. Most of the work is focused on speech and language and common sense reasoning.

3 Research program

3.1 Background

In recent years, Artificial Intelligence (AI) has achieved important landmarks in matching or surpassing human level performance on a number of high level tasks (playing chess and go, driving cars, categorizing picture, etc., [30, 34, 39, 29, 36]). These strong advances were obtained by deploying on large amounts of data, massively parallel learning architectures with simple brain-inspired 'neuronal' elements. However, humans brains still outperform machines in several key areas (language, social interactions, common sense reasoning, motor skills), and are more flexible : Whereas machines require extensive expert knowledge and massive training for each particular application, humans learn autonomously over several time scales: over the developmental scale (months), humans infants acquire cognitive skills with noisy data and little or no expert feedback (weakly/unsupervised learning)[1]; over the short time scale (minutes, seconds), humans combine previously acquired skills to solve new tasks and apply rules systematically to draw inferences on the basis of extremely scarce data (learning to learn, domain adaptation, one- or zero-shot learning) [32].

The general aim of CoML, following the roadmap described in [1], is to bridge the gap in cognitive flexibility between humans and machines learning in language processing and common sense reasoning by reverse engineering how young children between 1 and 4 years of age learn from their environment. We conduct work along two axes: the first one, which we called *Developmental AI* is focused on building infant inspired machine learning algorithms. The second axis is devoted to using the developed algorithms to conduct *quantitative studies* of how infant learn across diverse environments.

3.2 Weakly/Unsupervised Learning

Much of standard machine learning is construed as regression or classification problems (mapping input data to expert-provided labels). Human infants rarely learn in this fashion, at least before going to school: they learn language, social cognition, and common sense autonomously (without expert labels) and when adults provide feedback, it is ambiguous and noisy and cannot be taken as a gold standard. Modeling or mimicking such achievement requires deploying unsupervised or weakly supervised algorithms which are less well known than their supervised counterparts.

We take inspiration from infant's landmarks during their first years of life: they are able to learn acoustic models, a lexicon, and susbtantive elements of language models and world models from raw sensory inputs. Building on previous work [3, 7, 11], we use DNN and Bayesian architectures to model the emergence of linguistic representations without supervision. Our focus is to establish how the labels in supervised settings can be replaced by weaker signals coming either from multi-modal input or from hierarchically organised linguistic levels.

At the level of phonetic representations, we study how cross-modal information (lips and self feedback from articulation) can supplement top-down lexical information in a weakly supervised setting. We use Siamese architectures or Deep CCA algorithms to combine the different views. We study how an attentional framework and uncertainty estimation can flexibly combine these informations in order to adapt to situations where one view is selectively degraded.

At the level of lexical representations, we study how audio/visual parallel information (ie. descriptions of images or activities) can help in segmenting and clustering word forms, and vice versa, help in deriving useful visual features. To achieve this, we will use architectures deployed in image captioning or sequence to sequence translation [37].

At the level of semantic and conceptual representations, we study how it is possible to learn elements of the laws of physics through the observation of videos (object permanence, solidity, spatio-temporal continuity, inertia, etc.), and how objects and relations between objects are mapped onto language.

3.3 Evaluating Machine Intelligence

Increasingly, complicated machine learning systems are being incorporated into real-life applications (e.g. self-driving cars, personal assistants), even though they cannot be formally verified, guaranteed

statistically, nor even explained. In these cases, a well defined *empirical approach* to evaluation can offer interesting insights into the functioning and offer some control over these algorithms.

Several approaches exist to evaluate the 'cognitive' abilities of machines, from the subjective comparison of human and machine performance [38] to application-specific metrics (e.g., in speech, word error rate). A recent idea consist in evaluating an AI system in terms of it's *abilities* [31] , i.e., functional components within a more global cognitive architecture [35]. Psychophysical testing can offer batteries of tests using simple tasks that are easy to understand by humans or animals (e.g, judging whether two stimuli are same or different, or judging whether one stimulus is 'typical') which can be made selective to a specific component and to rare but difficult or adversarial cases. Evaluations of learning rate, domain adaptation and transfer learning are simple applications of these measures. Psychophysically inspired tests have been proposed for unsupervised speech and language learning [10], [33].

3.4 Documenting human learning

Infants learn their first language in a spontaneous fashion, across a lot of variation in amount of speech and the nature of the infant/adult interaction. In some linguistic communities, adults barely address infants until they can themselves speak. Despite these large variations in quantity and content, language learning proceeds at similar paces. Documenting such resilience is an essential step in understanding the nature of the learning algorithms used by human infants. Hence, we propose to collect and/or analyse large datasets of inputs to infants and correlate this with outcome measure (phonetic learning, vocabulary growth, syntactic learning, etc.).

4 Application domains

4.1 Speech processing for underresourced languages

We plan to apply our algorithms for the unsupervised discovery of speech units to problems relevant to language documentation and the construction of speech processing pipelines for underresourced languages.

4.2 Tools for the analysis of naturalistic speech corpora

Daylong recordings of speech in the wild gives rise a to number of specific analysis difficulties. We plan to use our expertise in speech processing to develop tools for performing signal processing and helping annotation of such resources for the purpose of phonetic or linguistic analysis.

5 Social and environmental responsibility

5.1 Footprint of research activities

The footprint of the CoML team due to travel was close to zero since, because of the sanitary conditions, all conferences were attended via video conference. The compute footprint was that of our 4 GPU cluster, used on average 30% of the time, to which we should add the compute of our accounts at Jean Zay (we could not access the data at the time of the report).

5.2 Impact of research results

Our fundamental work in unsupervised learning algorithms are very early stage and have up to now no known environmental/societal application. Our applicative work is dedicated to develop spoken language annotation and analysis tools for researchers, which should help conduct research in clinical and developmental areas (Health and Well Being, and early Education).

6 Highlights of the year

In 2021, we published two PNAS papers highlighting the use of machine learning to model human cognition (the first one on modeling phonetic learning in infants [22], the second one on human color systems [21]). We defined a new benchmark of metrics to evaluate language modeling from raw audio (Zero Ressource Speech Challenge 2021, [19]), and, in collaboration with Meta platforms, open sourced gSLM, the first generative spoken language model trained without any text or labels [13], and VoxPopuli, the largest multilingual public speech dataset ever (400K hours in 23 language) [23]. These high profile publications have already gathered a total of 83 citations by the end of 2021.

7 New software and platforms

7.1 New software

7.1.1 shennong

Keywords: Speech processing, Python, Information extraction, Audio signal processing

Functional Description: Shennong is a Python library which implement the most used methods for speech features extraction. Features extraction is the first step of every speech processing pipeline.

Shennong provides the following functionalities: - implementation of the main methods from state of the art (including pre and post processing) - exhaustive documentation and tests - usage from a Python API or a command line tool - simple and coherent interface

News of the Year: New processors for Vocal Tract Length Normalization and pitch extraction.

URL: https://docs.cognitive-ml.fr/shennong

Contact: Mathieu Bernard

7.1.2 phonemizer

Keyword: Text

Functional Description: * Conversion of a text into its phonemic representation * Wrapper on speech synthesis programs espeak and festival

News of the Year: Support for SAMPA phonetic alphabet with the new espeak-sampa backend. A lot of improvments and bug fixes.

URL: https://github.com/bootphon/phonemizer

Contact: Mathieu Bernard

7.1.3 TDE

Name: Term Discovery Evaluation

Keywords: NLP, Speech recognition, Speech

Scientific Description: This toolbox allows the user to judge of the quality of a word discovery algorithm. It evaluates the algorithms on these criteria : - Boundary : efficiency of the algorithm to found the actual boundaries of the words - Group : efficiency of the algorithm to group similar words - Token/Type: efficiency of the algorithm to find all words from the corpus (types), and to find all occurences (token) of these words. - NED : Mean of the edit distance across all the word pairs found by the algorithm - Coverage : efficiency of the algorithm to find every discoverable phone in the corpus

Functional Description: Toolbox to evaluate algorithms that segment speech into words. It allows the user to evaluate the efficiency of algorithms to segment speech into words, and create clusters of similar words.

News of the Year: Complete rewrite (optimization and bugfixes)

URL: https://github.com/bootphon/tdev2

Contact: Emmanuel Dupoux

7.1.4 wordseg

Name: wordseg

Keywords: Segmentation, Text, NLP

Functional Description: * Provides a collection of tools for text based word segmentation. * Covers the whole segmentation pipeline: data preprocessing, algorithms, evaluation and descriptive statistics. * Implements 6 segmentation algorithms and a baseline * Available as a Python library and command-line tools

News of the Year: New functionalities for cross-validation.

URL: https://wordseg.readthedocs.io

Contact: Mathieu Bernard

Partner: ENS Paris

7.1.5 abkhazia

Keywords: Speech recognition, Speech-text alignment

Functional Description: The Abkhazia sofware makes it easy to obtain simple baselines for supervised ASR (using Kaldi) and ABX tasks (using ABXpy) on the large corpora of speech recordings typically used in speech engineering, linguistics or cognitive science research.

URL: https://github.com/bootphon/abkhazia

Contact: Emmanuel Dupoux

7.1.6 ABXpy

Keywords: Evaluation, Speech recognition, Machine learning

Functional Description: The ABX package gives a performance score to speech recognition systems by measuring their capacity to discriminate linguistic contrasts (accents, phonemes, speakers, etc...)

URL: https://github.com/bootphon/ABXpy

Contact: Emmanuel Dupoux

7.1.7 abnet3

Keywords: Artificial intelligence, Speech processing, Deep learning, Unsupervised learning

Functional Description: Siamese network for unsupervised speech representation learning

URL: https://github.com/bootphon/abnet3

Contact: Emmanuel Dupoux

7.1.8 h5features

Keyword: File format

Functional Description: The h5features python package provides easy to use and efficient storage of large features data on the HDF5 binary file format.

URL: https://github.com/bootphon/h5features

Contact: Emmanuel Dupoux

7.1.9 intphys

Name: IntPhys: A Benchmark for Visual Intuitive Physics Reasoning

Keywords: Competition, Physical simulation, Artificial intelligence, Video Game

Functional Description: The intphys benchmark can be applied to any vision system, engineered, or trained, provided it can output a scalar when presented with a video clip, which should correspond to how physically plausible the video clip is. Our test set contains well matched videos of possible versus impossible events, and the metric consists in measuring how well the vision system can tell apart the possible from the impossible events..

URL: http://www.intphys.com

Contact: Mathieu Bernard

7.1.10 Seshat

Name: Seshat Audio Annotation Platform

Keywords: Audio, Speech, Web Application, Speech-text alignment

Functional Description: A web application to ease audio annotation campaigns, while also enabling the campaign manager to ensure that all annotations stick to a predefined format.

URL: https://github.com/bootphon/seshat

Contact: Hadrien Titeux

Partner: ENS Paris

7.1.11 pyGammaAgreement

Name: pyGammaAgreement

Keywords: Reliability, Measures

Functional Description: Python library for measuring inter and intra annotator reliability for annotation sequences

URL: https://github.com/bootphon/pygamma-agreement

Contact: Emmanuel Dupoux

8 New results

8.1 Unsupervised learning

Humans learn to speak and to perceive the world in a largely self-supervised fashion. Yet, most of machine learning is still devoted to supervised algorithms that rely on abundant quantities of human labelled data. We have used humans as sources of inspiration for developing novel machine learning algorithms in order to push the field towards self-supervised learning.

- In [13], we introduced Generative Spoken Language Modeling, the task of learning the acoustic and linguistic characteristics of a language from raw audio (no text, no labels), and a set of metrics to automatically evaluate the learned representations at acoustic and linguistic levels for both encoding and generation. We set up baseline systems consisting of a discrete speech encoder (returning pseudo-text units), a generative language model (trained on pseudo-text), and a speech decoder (generating a waveform from pseudo-text) all trained without supervision and validate the proposed metrics with human evaluation. Across 3 speech encoders (CPC, wav2vec 2.0, HuBERT), we find that the number of discrete units (50, 100, or 200) matters in a task-dependent and encoder-dependent way, and that some combinations approach text-based systems.

- In [20], we proposed using self-supervised discrete representations for the task of speech resynthesis. To generate disentangled representation, we separately extract low-bitrate representations for speech content, prosodic information, and speaker identity. This allows to synthesize speech in a controllable manner. We analyze various state-of-the-art, self-supervised representation learning methods and shed light on the advantages of each method while considering reconstruction quality and disentanglement properties. Specifically, we evaluate the F0 reconstruction, speaker identification performance (for both resynthesis and voice conversion), recordings' intelligibility, and overall quality using subjective human evaluation. Lastly, we demonstrate how these representations can be used for an ultra-lightweight speech codec. Using the obtained representations, we can get to a rate of 365 bits per second while providing better speech quality than the baseline methods. Audio samples can be found under the following link: speechbot.github.io/resynthesis.

- To reach human performance on complex tasks, a key ability for artificial systems is to understand physical interactions between objects, and predict future outcomes of a situation. This ability, often referred to as intuitive physics, has recently received attention and several methods were pro-posed to learn these physical rules from video sequences. Yet, most of these methods are restricted to the case where no, or only limited, occlusions occur. In [26], we propose a probabilistic formulation of learning intuitive physics in 3D scenes with significant inter-object occlusions. In our formulation, object positions are modeled as latent variables enabling the reconstruction of the scene. We then propose a series of approximations that make this problem tractable. Object proposals are linked across frames using a combination of a recurrent interaction network,modeling the physics in object space, and a compositional renderer, modeling the way in which objects project onto pixel space. We demonstrate significant improvements over state-of-the-art in the intuitive physics benchmark of Riochet et al.(2018). We apply our method to a second dataset with increasing levels of occlusions, showing it realistically predicts segmentation masks up to 30frames in the future. Finally, we also show results on predicting motion of objects in real videos

8.2 Datasets and Benchmarks

Self-supervised learning is a relatively new field of research. The CoML team contributes to the research by building benchmarks and organizing challenges to enable comparison between systems on a single set of metrics and cumulative progress across laboratories. The specificity of our approach is that we base our metrics on human psycholinguistics and psychophysics, enablig direct human - machine comparisons.

- In [19], we present the Zero Resource Speech Challenge 2021, which asks participants to learn a language model directly from audio, without any text or labels. The challenge is based on the Libri-light dataset, which provides up to 60k hours of audio from English audio books without any associated text. We provide a pipeline baseline system consisting on an encoder based on

contrastive predictive coding (CPC), a quantizer (k-means) and a standard language model (BERT or LSTM). The metrics evaluate the learned representations at the acoustic (ABX discrimination), lexical (spot-the-word), syntactic (acceptability judgment) and semantic levels (similarity judgment). We present an overview of the eight submitted systems from four groups and discuss the main results.

- In [23], we introduce VoxPopuli, a large-scale multilingual corpus providing 400K hours of unlabeled speech data in 23 languages. It is the largest open data to date for unsupervised representation learning as well as semisupervised learning. VoxPopuli also contains 1.8K hours of transcribed speeches in 15 languages and their aligned oral interpretations into 15 target languages totaling 17.3K hours. We provide speech recognition (ASR) baselines and validate the versatility of VoxPopuli unlabeled data in semisupervised ASR and speech-to-text translation under challenging out-of-domain settings.

- In order to reach human performance on complex visual tasks, artificial systems need to incorporate a significant amount of understanding of the world in terms of macroscopic objects, movements, forces, etc. Inspired by work on intuitive physics in infants, we propose in [16] an evaluation framework which diagnoses how much a given system understands about physics by testing whether it can tell apart well matched videos of possible versus impossible events. The test requires systems to compute a physical plausibility score over an entire video. It is free of bias and can test a range of specific physical reasoning skills. We then describe the first release of a benchmark dataset aimed at learning intuitive physics in an unsupervised way, using videos constructed with a game engine. We describe two Deep Neural Network baseline systems trained with a future frame prediction objective and tested on the possible versus impossible discrimination task. The analysis of their results compared to human data gives novel insights in the potentials and limitations of next frame prediction architectures.

8.3 Language emergence in communicative agents

In this research topic, which was the focus of Rahma Chaabouni's PhD thesis [24], which was taken up by the MSR funded PhD of Mathieu Rita, we study the inductive biases of neural systems by presenting them with few or no data.

- Words categorize the semantic fields they refer to in ways that maximize communication accuracy while minimizing complexity. Focusing on the well-studied color domain, we show that artificial neural networks trained with deep-learning techniques to play a discrimination game develop communication systems whose distribution on the accuracy/complexity plane closely matches that of human languages. The observed variation among emergent color-naming systems is explained by different degrees of discriminative need, of the sort that might also characterize different human communities. Like human languages, emergent systems show a preference for relatively low-complexity solutions, even at the cost of imperfect communication. We demonstrate next that the nature of the emergent systems crucially depends on communication being discrete (as is human word usage). When continuous message passing is allowed, emergent systems become more complex and eventually less efficient. Our study suggests that efficient semantic categorization is a general property of discrete communication systems, not limited to human language. It suggests moreover that it is exactly the discrete nature of such systems that, acting as a bottleneck, pushes them toward low complexity and optimal efficiency.

8.4 Evaluation of AI algorithms

Machine learning algorithms are typically evaluated in terms of end-to-end tasks, but it is very often difficult to get a grasp of how they achieve these tasks, what could be their break point, and more generally, how they would compare to the algorithms used by humans to do the same tasks. This is especially true of Deep Learning systems which are particularly opaque. The team develops evaluation/interpretation methods based on psycholinguistic/linguistic/neuroscience criteria, and deploy them for systematic comparison of systems.

- Deep Learning models have become potential candidates for auditory neuroscience research, thanks to their recent successes on a variety of auditory tasks. Yet, these models often lack interpretability to fully understand the exact computations that have been performed. In [15], we propose a parametrized neural network layer, that computes specific spectro-temporal modulations based on Gabor kernels (Learnable STRFs) and that is fully interpretable. We evaluated the predictive capabilities of this layer on Speech Activity Detection, Speaker Verification, Urban Sound Classification and Zebra Finch Call Type Classification. We found out that models based on this learnable parametrized neural network are on par for all tasks with the different toplines, and obtain the best performance for Speech Activity Detection. As this layer is fully interpretable, we used quantitative measures to describe the distribution of the learned spectro-temporal modulations. The filters adapted to each task and focused mostly modulation on low temporal and spectral modulations. The analyses show that the filters learned on human speech have similar spectro-temporal parameters as the ones measured directly in the human auditory cortex. Finally, equipped with the Sinkhorn distance to com- pare the learned STRFs distributions, we observed that the tasks organized in a meaningful way: the human vocalizations tasks closer to each other and bird vocalizations far away from human vocalizations and urban sounds tasks.

8.5 Simulation of language learning in infants

Supervised learning algorithms provide very interesting quantitative models of the early phases of language acquisition. When fed with realistic input, they can generate predictions that can be compared with available developmental behavioral data.

- Before they even speak, infants become attuned to the sounds of the language(s) they hear, processing native phonetic contrasts more easily than non-native ones. For example, between 6-8 months and 10-12 months, infants learning American English get better at distinguishing English [r] and [l], as in 'rock' vs 'lock', relative to infants learning Japanese. Influential accounts of this early phonetic learning phenomenon initially proposed that infants group sounds into native vowel- and consonant-like phonetic categories—like [r] and [l] in English—through a statistical clustering mechanism dubbed 'distributional learning'. The feasibility of this mechanism for learning phonetic categories has been challenged, however. In [22], we demonstrate that a distributional learning algorithm operating on naturalistic speech can predict early phonetic learning as observed in Japanese and American English infants, suggesting that infants might learn through distributional learning after all. We further show, however, that contrary to the original distributional learning proposal, our model learns units too brief and too fine-grained acoustically to correspond to phonetic categories. This challenges the influential idea that what infants learn are phonetic categories. More broadly, our work introduces a novel mechanism-driven approach to the study of early phonetic learning, together with a quantitative modeling framework that can handle realistic input. This allows, for the first time, accounts of early phonetic learning to be linked to concrete, systematic predictions regarding infants' attunement.

- Theories and data on language acquisition suggest a range of cues are used, ranging from information on structure found in the linguistic signal itself, to information gleaned from the environmental context or through social interaction. In [17], we propose a blueprint for computational models of the early language learner (SCALa, for Socio-Computational Architecture of Language Acquisition) that makes explicit the connection between the kinds of information available to the social learner and the computational mechanisms required to extract language-relevant information and learn from it. SCALa integrates a range of views on language acquisition, further allowing us to make precise recommendations for future large-scale empirical research.

8.6 Quantitative studies of human learning and processing

Evidently, infants are acquiring their language based on whatever linguistic input is available around them. The extent of variation that can be found across languages, cultures and socio-economic background provides strong constraints (lower bounds on data, higher bounds on noise, and variation and ambiguity) for language learning algorithms. Vice-versa, aging adults, or patients with neurological impairements

show degradation in speech and language patterns which can be used to diagnose or predict the progress of the impairement.

- Nous présentons une implémentation libre (open-source) en Python de la mesure Gamma (γ) pour l'accord inter/intra-annotateurs.
- A prominent hypothesis holds that by speaking to infants in infant-directed speech (IDS) as opposed to adult-directed speech (ADS), parents help them learn phonetic categories. Specifically, two characteristics of IDS have been claimed to facilitate learning: hyperarticulation, which makes the categories more separable and variability, which makes the generalization more robust. Here, we test the separability and robustness of vowel category learning on acoustic representations of speech uttered by Japanese adults in either ADS, IDS (addressed to 18-24 month olds) or read speech (RS). Separability is determined by means of a distance measure computed between the five short vowel categories of Japanese, while robustness is assessed by testing the ability of six different machine learning algorithms trained to classify vowels to generalize on stimuli spoken by a novel speaker in ADS. Using two different speech representations, we find that hyperarticulated speech, in the case of RS, can yield better separability, and that increased between-speaker variability in ADS, can yield, for some algorithms, more robust categories. However, these conclusions do not apply to IDS, which turned out to yield neither more separable nor more robust categories compared to ADS inputs. We discuss the usefulness of machine learning algorithms run on real data to test hypotheses about the functional role of IDS.

8.7 Test of the psychological validity of AI algorithms.

In this section, we focus on the utilisation of machine learning algorithms of speech and language processing to derive testable quantitative predictions in humans (adults or infants).

-

8.8 Applications and tools for researchers

Some of CoMLs' activity is to produce speech and language technology tools that facilitate research into language development or clinical applications.

-

8.9 Interactive AI

- Curiosity is a vital metacognitive skill in educational contexts, leading to creativity, and a love of learning. And while many school systems increasingly undercut curiosity by teaching to the test, teachers are increasingly interested in how to evoke curiosity in their students to prepare them for a world in which lifelong learning and reskilling will be more and more important. One aspect of curiosity that has received little attention, however, is the role of peers in eliciting curiosity. We present what we believe to be the first theoretical framework that articulates an integrated socio-cognitive account of curiosity that ties observable behaviors in peers to underlying curiosity states. We make a bipartite distinction between individual and interpersonal functions that contribute to curiosity, and multimodal behaviors that fulfill these functions. We validate the proposed framework by leveraging a longitudinal latent variable modeling approach. Findings confirm a positive predictive relationship between the latent variables of individual and interpersonal functions and curiosity, with the interpersonal functions exercising a comparatively stronger influence. Prominent behavioral realizations of these functions are also discovered in a data-driven manner. We instantiate the proposed theoretical framework in a set of strategies and tactics that can be incorporated into learning technologies to indicate, evoke, and scaffold curiosity. This work is a step towards designing learning technologies that can recognize and evoke moment-by-moment curiosity during learning in social contexts and towards a more complete multimodal learning analytics. The underlying rationale is applicable more generally for developing computer support for other metacognitive and socio-emotional skills.

- Interaction takes place not only on the propositional level but also on the social level. In this paper, we consider rapport as an important social phenomenon in interaction. Motivated by data from the tutoring domain, we hypothesize that (i) off-task episodes are triggered by a low level of rapport and (ii) such episodes are means of raising the level of rapport. We sketch a planning model that allows off-task episodes to be triggered by (low) rapport level, which we apply to two simple examples.

9 Bilateral contracts and grants with industry

- **Facebook AI Research Grant** (2021, PI: E. Dupoux, 350K€) - Unrestricted Gift - The aim is to help the development of machine learning tools geared towards the psycholinguistic research community.

10 Partnerships and cooperations

10.1 International research visitors

10.1.1 Visits of international scientists

Inria International Chair

IIC SMOLENSKY Paul

Name of the chair: *Tensor Product Representations*

Institution of origin: *Johns Hopkins University*

Country: *USA*

Dates: From Sun Jan 01 2017 to Fri Dec 31 2021

Title: Contribute to the development of a Fifth Generation of Artificial Intelligence: AI-5

Summary: The aim is to integrate symbolic and neural network computation for modeling reasoning and, especially, grammar in the human mind/brain. The work is formal and computational, with emerging applications to neuroscience and applied natural language processing.

Other international visits to the team

Ewan Dunbar

Status: Researcher

Institution of origin: University of Toronto

Country: Canada

Dates: Jan 01, 2021 to Dec 31, 2021

Context of the visit: Develop linguistically and psycholinguistically evaluation methods for neural models of speech and language processing.

Mobility program/type of mobility: research stay and collaboration

10.2 National initiatives

10.2.1 ANR

- **ANR GEOMPHON**. (2018-2021; coordinating PI : E. Dunbar; 299K€) - Study the effects of typologically common properties of linguistic sound systems on speech perception, human learning, and machine learning applied to speech.

10.3 Regional initiatives

- **SESAME** (Région Ile de France) (2021-2031; coordinating PI: E. Dupoux, 400k€). The echolalia platform. Digital platform for annotating audio/video language data and analyzing them with AI systems. Applications: language development, sign language, language pathology.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- organization of the ZR Challenge 2021 (Interspeech challeng; E Dupoux, E. Dunbar main organizers)
- organization of the multimodal ZR Challenge 2021 (Neurips challenge; E. Dupoux, E. Dunbar, main organizers)
- co-organizer of blackbox NLP 2021 (E. Dupoux, co-organizer)

11.1.2 Scientific events: selection

Member of the conference program committees

- SIGdial Workshop on Discourse and Dialogue (J. Cassell)
- Semantics and Pragmatics of Dialogue (J. Cassell)
- Computer Animation and Social Agents (J. Cassell)
- International Society of Gesture Studies Annual Conference (J. Cassell)

11.1.3 Journal

Member of the editorial boards

- Member, Editorial Board, Interaction Studies: J. Cassell

11.1.4 Invited talks

E. Dupoux :

- Invited talk at CIFAR LMB on Self Supervised Speech learning
- Invited talk at College de France. Colloquium on representation of language in brain and machines (June 24-25, 2021)

J. Cassell :

- Distinguished Lecture in Computer Science, Stockholm, Sweden, October 2021
- Distinguished Lecture in Cognitive Science, University of Lund, Sweden, October 2021
- Oberlander Memorial Lecture, University of Edinburgh, May 2021
- Lily Endowment Colloquium on the Ethics of Relating Digitally, February 2021
- Keynote : Les Robots, Nouveaux Partenaires des Soins Tendres, Grenoble, France (Nov.)
- Keynote Conférence Nationale en Intelligence Artificielle (FSIA : CNIA), Bordeaux, France (June).
- Opening Keynote, OECD International Conference on AI in Work, Innovation, Productivity and Skills, Paris, France (February).
- Keynote: Human-Computer Interaction, Gangwon-do, Korea (Jan).

11.1.5 Research administration

- Member of the coordinating committee, Carnot Cognition: E. Dupoux

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

E. Dupoux is co-director of the Cognitive Engineering track in the Cognitive Science Master (ENS, EHESS, Paris V).

- Master : E. Dupoux (with B. Sagot, ALMANACH, N. Zeghidour & R. Riad, COML), "Algorithms for speech and language processing", 30h, M2, (MVA), ENS Cachan, France
- Doctorat : E. Dupoux, "Computational models of cognitive development", 32 h, Séminaire EHESS, Paris France

J. Cassell is professor of language technologies and human-computer interaction (ENS)

- Master: J. Cassell "Conversation among People and Bots", fall semester, M2 (Cogmaster), ENS-EHESS-UP
- Guest lectures/seminars in classes: J. Cassell: EPFL, University of Pennsylvania, UE Cognition Sociale de la Sorbonne, Seminaire Doctoral Littérature et Culture d'Enfance at ENS-Afreloce, Ethics & Societal Impact of AI at MBZUAI (MBZ University of AI in Abu Dhabi), Sociologie du Numérique.

11.2.2 Supervision

- Alafate Abulimiti, PhD thesis (J. Cassell)
- Yann Raphalen, PhD thesis (J. Cassell)
- Robin Algayres, PhD thesis (E. Dupoux, B. Sagot)
- Rahma Chaabouni, PhD thesis (E. Dupoux, M. Baroni)
- Maureen De Seyssel, PhD thesis (E. Dupoux)
- Marvin Lavechin, PhD thesis (E. Dupoux, A. Cristia)
- Rachid Riad, PhD thesis (E. Dupoux, A-C. Bachoud-Levi)
- Mathieu Rita, PhD thesis (E. Dupoux, O. Pietquin)

11.2.3 Juries

- Tenure/Track hirhin committee (ENS, S. Mascarenas): E. Dupoux
- PhD Jury: Tanvi Dinkar (Telecom Paris), Comités de Suivi : Pierre-Louis Guhur (Inria) : J. Cassell

11.3 Popularization

11.3.1 Internal or external responsibilities

J. Cassell:

- Member (by order of the Prime Minister), Conseil National du Numérique
- Presidente Jury, Choose France Recrutement Senior Faculty in AI, Inria
- Jury, Choose France Junior Recrutement
- Jury, Marie Curie Postdoctoral Fellowships, Ile de France

- Jury, Chaire Blaise Pascale, Ile de France
- Advisory Board, ILCB (Institute for Language, Communication and the Brain), Université d'Aix-Marseille
- Advisory Board, Computer Science Department, University of the People
- Advisory Board, NSF Project on " Developing Conversational Videos to Support Children's STEM Learning and Engagement"
- Member, Editorial Board, Interaction Studies

11.3.2 Interventions

J. Cassell :

- Global Tech Thinkers Dinner (private dinner hosted by President Macron at the Hôtel de la Marine), November 2021
- GPT-3 and Conversational Agents (journée d'études Tesaco), November 2021
- OECD Conference on the Role of AI in the Productivity of Science, November 2021
- Fête de La Science, Lycée Louis le Grand, Paris, October 2021
- Hexagone Scène Nationale Arts Science Panel on IA et Langage, June 2021
- Adobe Horizons Business, Futur du Travail Table Ronde, June 2021
- USI (Unexpected Sources of Inspiration) Conference, June 2021
- Soka University Future of Education Panel, June 2021
- WikiMedia Panel, l'Avenir de l'Education en Ligne, May 2021
- Cognivence, Le Forum des Sciences Cognitives, April 2021

12 Scientific production

12.1 Major publications

[1] E. Dupoux. 'Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner'. In: *Cognition* (2018).

[2] A. Fourtassi and E. Dupoux. 'A Rudimentary Lexicon and Semantics Help Bootstrap Phoneme Acquisition'. In: *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*. Baltimore, Maryland USA: Association for Computational Linguistics, June 2014, pp. 191–200. DOI: 10.3115/v1/W14-1620.

[3] A. Fourtassi, T. Schatz, B. Varadarajan and E. Dupoux. 'Exploring the Relative Role of Bottom-up and Top-down Information in Phoneme Learning'. In: *Proceedings of the 52nd Annual meeting of the ACL*. Vol. 2. ACL. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 1–6. DOI: 10.3115/v1/P14-2001.

[4] Y. Hoshen, R. J. Weiss and K. W. Wilson. 'Speech acoustic modeling from raw multichannel waveforms'. In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 4624–4628.

[5] T. Linzen, E. Dupoux and Y. Goldberg. 'Assessing the ability of LSTMs to learn syntax-sensitive dependencies'. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535.

[6] T. Linzen, E. Dupoux and B. Spector. 'Quantificational features in distributional word representations'. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 2016, pp. 1–11. DOI: 10.18653/v1/S16-2001.

[7] A. Martin, S. Peperkamp and E. Dupoux. 'Learning Phonemes with a Proto-lexicon'. In: *Cognitive Science* 37 (2013), pp. 103–124. DOI: 10.1111/j.1551-6709.2012.01267.x.

[8] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville and Y. Bengio. 'SampleRNN: An unconditional end-to-end neural audio generation model'. In: *arXiv preprint arXiv:1612.07837* (2016).

[9] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson and O. Vinyals. 'Learning the speech front-end with raw waveform CLDNNs'. In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[10] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hynek and E. Dupoux. 'Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline'. In: *INTERSPEECH-2013*. International Speech Communication Association. Lyon, France, 2013, pp. 1781–1785.

[11] R. Thiollière, E. Dunbar, G. Synnaeve, M. Versteegh and E. Dupoux. 'A Hybrid Dynamic Time Warping-Deep Neural Network Architecture for Unsupervised Acoustic Modeling'. In: *INTERSPEECH-2015*. 2015, pp. 3179–3183.

[12] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu. 'Wavenet: A generative model for raw audio'. In: *CoRR abs/1609.03499* (2016).

12.2 Publications of the year

International journals

[13] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed and E. Dupoux. 'On Generative Spoken Language Modeling from Raw Audio'. In: *Transactions of the Association for Computational Linguistics* (1st Feb. 2021). URL: https://hal.inria.fr/hal-03329219.

[14] B. Ludusan, R. Mazuka and E. Dupoux. 'Does infant-directed speech help phonetic learning? A machine learning investigation'. In: *Cognitive Science* 45.5 (May 2021). DOI: 10.1111/cogs.12946. URL: https://hal.archives-ouvertes.fr/hal-03080098.

[15] R. Riad, J. Karadayi, A.-C. Bachoud-Lévi and E. Dupoux. 'Learning spectro-temporal representations of complex sounds with parameterized neural networks'. In: *Journal of the Acoustical Society of America* 150.1 (15th Mar. 2021), pp. 353–366. DOI: 10.1121/10.0005482. URL: https://hal.inria.fr/hal-03329261.

[16] R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard and E. Dupoux. 'IntPhys 2019: A Benchmark for Visual Intuitive Physics Understanding'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1st June 2021). URL: https://hal.archives-ouvertes.fr/hal-02274273.

[17] S. Tsuji, A. Cristia and E. Dupoux. 'SCALa: A blueprint for computational models of language acquisition in social context'. In: *Cognition* 213 (3rd June 2021), p. 104779. DOI: 10.1016/j.cognition.2021.104779. URL: https://hal.inria.fr/hal-03373586.

International peer-reviewed conferences

[18] A. Abulimiti, J. Cassell and J. Ginzburg. '"By the way, do you like Spider Man?" -Towards A Social Planning Model for Rapport'. In: SemDial 2021 - 25th Workshop on the Semantics and Pragmatics of Dialogue. Potsdam / Virtual, Germany, 20th Sept. 2021. URL: https://hal.inria.fr/hal-03536332.

[19] E. Dunbar, M. Bernard, N. Hamilakis, T. A. Nguyen, M. De Seyssel, P. Rozé, M. Rivière, E. Kharitonov and E. Dupoux. 'The Zero Resource Speech Challenge 2021: Spoken language modelling'. In: Interspeech 2021 - Conference of the International Speech Communication Association. Brno, Czech Republic, 30th Aug. 2021. DOI: 10.1109/TPAMI.2021.3083839. URL: https://hal.inria.fr/hal-03329301.

[20] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed and E. Dupoux. 'Speech Resynthesis from Discrete Disentangled Self-Supervised Representations'. In: INTERSPEECH 2021 - Annual Conference of the International Speech Communication Association. Brno, Czech Republic, 1st Apr. 2021. URL: https://hal.inria.fr/hal-03329245.

Edition (books, proceedings, special issue of a journal)

[21] R. Chaabouni, E. Kharitonov, E. Dupoux and M. Baroni. *Communicating artificial neural networks develop efficient color-naming systems*. Vol. 118. 12. National Academy of Sciences, 15th Mar. 2021. DOI: 10.1073/pnas.2016569118. URL: https://hal.inria.fr/hal-03329084.

[22] T. Schatz, N. H. Feldman, S. Goldwater, X. N. Cao and E. Dupoux. *Early phonetic learning without phonetic categories – Insights from large-scale simulations on realistic input*. Vol. 118. 7. National Academy of Sciences, 28th Jan. 2021, e2001844118. DOI: 10.1073/pnas.2001844118. URL: https://hal.archives-ouvertes.fr/hal-03070566.

[23] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino and E. Dupoux. *VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation*. 9th Aug. 2021. DOI: 10.18653/v1/2021.acl-long.80. URL: https://hal.inria.fr/hal-03329290.

Doctoral dissertations and habilitation theses

[24] R. Chaabouni. 'Emerging linguistic universals in communicating neural network agents'. Ecole doctorale cerveau-cognition comportement (ED3C), 17th Mar. 2021. URL: https://hal.inria.fr/tel-03536320.

[25] R. Riochet. 'Unsupervised Learning of Intuitive Physics from Videos'. Ecole Normale Superieure de Paris - ENS Paris, 30th June 2021. URL: https://hal.archives-ouvertes.fr/tel-03530321.

Reports & preprints

[26] R. Riochet, J. Sivic, I. Laptev and E. Dupoux. *Occlusion resistant learning of intuitive physics from videos*. 12th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03139755.

[27] T. Sinha, Z. Bai and J. Cassell. *A Novel Multimodal Approach for Studying the Dynamics of Curiosity in Small Group Learning*. 19th Jan. 2022. DOI: 10.35542/osf.io/rfxwg. URL: https://hal.inria.fr/hal-03536340.

[28] H. Titeux and R. Riad. *pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python*. 12th June 2021. DOI: 10.21105/joss.02989. URL: https://hal.archives-ouvertes.fr/hal-03144116.

12.3 Cited publications

[29] D. A. Ferrucci. 'Introduction to "this is watson"'. In: *IBM Journal of Research and Development* 56.3.4 (2012), pp. 1–1.

[30] K. He, X. Zhang, S. Ren and J. Sun. 'Delving deep into rectifiers: Surpassing human-level performance on imagenet classification'. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1026–1034.

[31] J. Hernández-Orallo, F. Martínez-Plumed, U. Schmid, M. Siebers and D. L. Dowe. 'Computer models solving intelligence test problems: Progress and implications'. In: *Artificial Intelligence* 230 (2016), pp. 74–107.

[32] B. M. Lake, T. D. Ullman, J. B. Tenenbaum and S. J. Gershman. 'Building machines that learn and think like people'. In: *arXiv preprint arXiv:1604.00289* (2016).

[33] T. Linzen, E. Dupoux and Y. Goldberg. 'Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies'. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535.

[34] C. Lu and X. Tang. 'Surpassing human-level face verification performance on LFW with GaussianFace'. In: *arXiv preprint arXiv:1404.3840* (2014).

[35] S. T. Mueller. 'A partial implementation of the BICA cognitive decathlon using the Psychology Experiment Building Language (PEBL)'. In: *International Journal of Machine Consciousness* 2.02 (2010), pp. 273–288.

[36] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis. 'Mastering the game of Go with deep neural networks and tree search'. In: *Nature* 529.7587 (2016), pp. 484–489.

[37] I. Sutskever, O. Vinyals and Q. V. Le. 'Sequence to sequence learning with neural networks'. In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

[38] A. M. Turing. 'Computing machinery and intelligence'. In: *Mind* 59.236 (1950), pp. 433–460.

[39] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig. 'Achieving human parity in conversational speech recognition'. In: *arXiv preprint arXiv:1610.05256* (2016).