Activity Report 2018

Team COML

The Cognitive Machine Learning Team

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).

RESEARCH CENTER
Paris

THEME
Language, Speech and Audio

Table of contents

Team COML

Creation of the Team: 2017 May 04

Keywords:

Computer Science and Digital Science:

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
A9.6. - Decision support
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B1.2. - Neuroscience and cognitive science
B1.2.2. - Cognitive science

1. Team, Visitors, External Collaborators

Research Scientist

Xuan Nga Cao [Ecole des hautes études en sciences sociales, Researcher]

Faculty Member

Emmanuel Dupoux [Ecole des hautes études en sciences sociales, Professor, HDR]

PhD Students

Maria Julia Carbajal [Ecole Normale Supérieure Paris, until Sep 2018]
Rahma Chaabouni [Ecole Normale Supérieure Paris]
Adriana Carolina Guevara Rukoz [Ecole Normale Supérieure Paris, until Sep 2018]
Elin Larsen [Ecole Normale Supérieure Paris, until Dec 2018]
Rachid Riad [Ecole des hautes études en sciences sociales]
Neil Zeghidour [Facebook]

Technical staff

Mathieu Bernard [Inria]
Julien Karadayi [Ecole des hautes études en sciences sociales]
Catherine Urban [Ecole des hautes études en sciences sociales]

Interns

Diego Andai [Inria, until Apr 2018]
Erwan Simon [Inria, from Apr 2018 until Jul 2018]

Administrative Assistant
Chantal Chazelas [Inria]

2. Overall Objectives

2.1. 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., [28], [31], [36], [27], [33]). 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) [30].

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. We conduct work in three areas: weakly supervised and unsupervised algorithms, datasets and benchmarks, and machine intelligence evaluation.

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 [34].

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 [35] 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* [29] , i.e., functional components within a more global cognitive architecture [32]. 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], [6].

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. New Software and Platforms

5.1. 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.

- Contact: Emmanuel Dupoux
- URL: https://github.com/bootphon/abkhazia

5.2. TDE

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.

- Contact: Emmanuel Dupoux
- URL: https://github.com/bootphon/TDE

5.3. 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...)

- Contact: Emmanuel Dupoux
- URL: https://github.com/bootphon/ABXpy

5.4. 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.

- Contact: Emmanuel Dupoux
- URL: https://github.com/bootphon/h5features

6. New Results

6.1. Speech and Audio Processing from the Raw Waveform

State-of-the-art speech technology systems (e.g., ASR and TTS) rely on fixed, hand-crafted features such as mel-filterbanks to preprocess the waveform before the training pipeline. This is at odds with recent work in machine vision where hand-crafter features (SIFT, etc) have been succesfully replaced by features derived from raw pixels trained jointly with a downstream task. In this line of work, we explored how a similar approach could be undertaken for audio and speech processing.

- In [24], we train a bank of complex filters that operates at the level of the raw speech signal and feeds into a convolutional neural network for phone recognition. These time-domain filterbanks (TD-filterbanks) are initialized as an approximation of MFSC, and then fine-tuned jointly with the remaining convolutional network. We perform phone recognition experiments on TIMIT and show that for several architectures, models trained on TD-filterbanks consistently out-perform their counterparts trained on comparable MFSC. We get our best performance by learning all front-end steps, from pre-emphasis up to averaging. Finally, we observe that the filters at convergence have an asymmetric impulse response while preserving some analyticity.
- In [25], we study end-to-end systems trained directly from the raw waveform, building on two alternatives for trainable replacements of mel-filterbanks that use a convolutional architecture. The first one is inspired by gammatone filterbanks [4], [9], and the second one by the scattering transform [24]. We propose two modifications to these architectures and systematically compare them to mel-filterbanks, on the Wall Street Journal dataset. The first modification is the addition of an instance normalization layer, which greatly improves on the gammatone-based trainable filterbanks and speeds up the training of the scattering-based filterbanks. The second one relates to the low-pass filter used in these approaches. These modifications consistently improve performances for both approaches, and remove the need for a careful initialization in scattering-based trainable filterbanks. In particular, we show a consistent improvement in word error rate of the trainable filterbanks relatively to comparable mel-filterbanks. It is the first time end-to-end models trained from the raw signal significantly outperform mel-filterbanks on a large vocabulary task under clean recording conditions.
- Recent progress in deep learning for audio synthesis opens the way to models that directly produce the waveform, shifting away from the traditional paradigm of relying on vocoders or MIDI synthesizers. Despite their successes, current state-of-the-art neural audio synthesizers such as WaveNet and SampleRNN [12], [8] suffer from prohibitive training and inference times because they are based on autoregressive models that generate audio samples one at a time at a rate of 16kHz. In this work [26], we study the more computationally efficient alternative of generating the waveform frame-by-frame with large strides. We present SING, a lightweight neural audio synthesizer for the original task of generating musical notes given desired instrument, pitch and velocity. Our model is trained end-to-end to generate notes from nearly 1000 instruments with a single decoder, thanks to a new loss function that minimizes the distances between the log spectrograms of the generated and target waveforms. On the generalization task of synthesizing notes for pairs of pitch and instrument not seen during training, SING produces audio with significantly improved perceptual quality compared to a state-of-the-art autoencoder based on WaveNet [4] as measured by a Mean Opinion Score (MOS), and is about 32 times faster for training and 2,500 times faster for inference.

6.2. Development of cognitively inspired algorithms

Speech and language processing in humans infants and adults is particularly efficient. We use these as sources of inspiration for developing novel machine learning and speech technology algorithms. In this area, our results are as follows:

- In [22], we summarize the accomplishments of a multi-disciplinary 6-weeks workshop organized by E. Dupoux (PI) at Carnegy Mellon Univerrsity (Pittsburgh), funded through the Jelinek Memorial Summer Workshop Program of Johns Hopkins University. The workshop explored the computational and scientific issues surrounding the discovery of linguistic units (subwords and words) in a language without orthography. We studied the replacement of orthographic transcriptions by images and/or translated text in a well-resourced language to help unsupervised discovery from raw speech.
- Developing speech technologies for low-resource languages has become a very active research field over the last decade. Among others, Bayesian models have shown some promising results on artificial examples but still lack of in situ experiments. In [20] , we apply state-of-the-art Bayesian models to unsupervised Acoustic Unit Discovery (AUD) in a real low-resource language scenario. We also show that Bayesian models can naturally integrate information from other resourceful languages by means of informative prior leading to more consistent discovered units. Finally, discovered acoustic units are used, either as the 1-best sequence or as a lattice, to perform word segmentation. Word segmentation results show that this Bayesian approach clearly outperforms a Segmental-DTW baseline on the same corpus.
- Fixed-length embeddings of words are very useful for a variety of tasks in speech and language processing. In [19], we systematically explore two methods of computing fixed-length embeddings for variable-length sequences. We evaluate their susceptibility to phonetic and speaker-specific variability on English, a high resource language, and Xitsonga, a low resource language, using two evaluation metrics: ABX word discrimination and ROC-AUC on same-different phoneme n-grams. We show that a simple downsampling method supplemented with length information can be competitive with the variable-length input feature representation on both evaluations. Recurrent autoencoders trained without supervision can yield even better results at the expense of increased computational complexity.
- Recent studies have investigated siamese network architectures for learning invariant speech representations using same-different side information at the word level. In [21], we investigate systematically an often ignored component of siamese networks: the sampling procedure (how pairs of same vs. different tokens are selected). We show that sampling strategies taking into account Zipf's Law, the distribution of speakers and the proportions of same and different pairs of words significantly impact the performance of the network. In particular, we show that word frequency compression improves learning across a large range of variations in number of training pairs. This effect does not apply to the same extent to the fully unsupervised setting, where the pairs of same-different words are obtained by spoken term discovery. We apply these results to pairs of words discovered using an unsupervised algorithm and show an improvement on state-of-the-art in unsupervised representation learning using siamese networks.
- Unsupervised spoken term discovery is the task of finding recurrent acoustic patterns in speech without any annotations. Current approaches consists of two steps: (1) discovering similar patterns in speech, and (2) partitioning those pairs of acoustic tokens using graph clustering methods. In, [23] we propose a new approach for the first step. Previous systems used various approximation algorithms to make the search tractable on large amounts of data. Our approach is based on an optimized k-nearest neighbours (KNN) search coupled with a fixed word embedding algorithm. The results show that the KNN algorithm is robust across languages, consistently outperforms the DTW-based baseline, and is competitive with current state-of-the-art spoken term discovery systems.

6.3. 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).

- Two PhDs were defended this year. In [14], Adriana Guavara Rukoz presented a computational model of the perception of non-native speech contrasts based on standard ASR pipelines is presented. An adaptation of the model is proposed to account for forced-choice classification psycholinguistic

experiments and directly reproduced classical results. The general finding is that, suprisingly, the acoustic model part of a phone recognizer is sufficient to account for experimental data, even those apparently related to phonotactic properties of the native language. The 'language model' part does not improve the correlation with adult data (if anything, it degrades it). Yet the match between model and human is not perfect, and is was hypothetized that improvement in the acoustic model could help. In [13], Julia Maria Carbajal presented a study of the effect of multilingual exposure on language acquisition. She used a computational model of language separation based on i-vectors to reproduce some of the known effects of phonological distance on language discrimination in infants.

- In [16], we investigate whether infant-directed speech (IDS) facilitates lexical learning when compared to adult-directed speech (ADS). To study this, we compare the distinctiveness of the lexicon at two levels, acoustic and phonological, using a large database of spontaneous speech in Japanese. At the acoustic level we show that, as has been documented before for phonemes, the realizations of words are more variable and less discriminable in IDS. At the phonological level, we find that despite a slight increase in the number of phonological neighbors, the IDS lexicon contains more distinctive words (such as onomatopeias). Combining the acoustic and phonological metrics together in a global discrimination score, the two effects cancel each other out and the IDS lexicon winds up being as discriminable as its ADS counterpart. We discuss the implication of these findings for the view of IDS as hyperspeech, i.e., a register whose purpose is to facilitate language acquisition.
- Existing theories of cross-linguistic phonetic category perception agree that listeners perceive foreign sounds by mapping them onto their native phonetic categories. Yet, none of the available theories specify a way to compute this mapping. As a result, they cannot provide systematic quantitative predictions and remain mainly descriptive. Here [17], Automatic Speech Recognition (ASR) systems are used to provide a fully specified mapping between foreign and native sounds. This is shown to provide a quantitative model that can account for several empirically attested effects in human cross-linguistic phonetic category perception.
- Spectacular progress in the information processing sciences (machine learning, wearable sensors) promises to revolutionize the study of cognitive development. In [15], we analyse the conditions under which 'reverse engineering' language development, i.e., building an effective system that mimics infant's achievements, can contribute to our scientific understanding of early language development. We argue that, on the computational side, it is important to move from toy problems to the full complexity of the learning situation, and take as input as faithful reconstructions of the sensory signals available to infants as possible. On the data side, accessible but privacy-preserving repositories of home data have to be setup. On the psycholinguistic side, specific tests have to be constructed to benchmark humans and machines at different linguistic levels. We discuss the feasibility of this approach and present an overview of current results.

6.4. 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.

- In [18], we present BabyCloud, a platform for capturing, storing and analyzing daylong audio recordings and photographs of children's linguistic environments, for the purpose of studying infant's cognitive and linguistic development and interactions with the environment. The proposed platform connects two communities of users: families and academics, with strong innovation potential for each type of users. For families, the platform offers a novel functionality: the ability for parents to follow the development of their child on a daily basis through language and cognitive metrics (growth curves in number of words, verbal complexity, social skills, etc). For academic research, the platform provides a novel means for studying language and cognitive development at an unprecedented scale and level of detail. They will submit algorithms to the secure server which will only output anonymized aggregate statistics. Ultimately, BabyCloud aims at creating an ecosystem of third parties (public and private research labs...) gravitating around developmental data, entirely controlled by the party whose data originate from, i.e. families.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Grants with Industry

- Google Faculty Award - 100K€
- Facebook AI Research Grant - 350K€

8. Partnerships and Cooperations

8.1. Regional Initiatives

Collaboration with the Willow Team:

- co-advising with J. Sivic and I. Laptev of a PhD student: Ronan Riochet.
- construction of a naive physics benchmark (www.intphys.com)

8.2. National Initiatives

8.2.1. ANR

- Transatlantic Platform "Digging into Data". Title: "Analysis of Children's Language Experiences Around the World. (ACLEW)"; (coordinating PI : M. Soderstrom; Leader of tools development and co-PI : E. Dupoux), (2017–2020. 5 countries; Total budget: 1.4M€)

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

- Johns Hopkins University, Baltimore, USA: S. Kudanpur, H. Hermanksy
- RIKEN Institute, Tokyo, Japan: R. Mazuka

8.4. International Research Visitors

8.4.1. Visits of International Scientists

8.4.1.1. Internships

Internship of Diego Andai Castilla (partnership Inria-PUC-Inria Chile)

8.4.2. Visits to International Teams

8.4.2.1. Research Stays Abroad

- E. Dupoux Visiting Researcher at Facebook AI Research, Paris (Feb-Mar 2018)
- E. Dupoux Visiting Researcher at Google & DeepMind, London (April-July 2018)

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

- E. Dupoux Co-Program Chair of NIPS 2018 workshop on intuitive physics, Montreal.
- E. Dupoux Co-Program chair of of the LEGRAIN Conference on Learning in Humans and Machines, Ecole Normale Supérieure, 2018 (this conference had a scientific, an industrial and a general public track)

9.1.1.2. Member of the Organizing Committees

- Executive committee of SIGMORPHON (Association for Computational Linguistics Special Interest Group, http://www.sigmorphon.org/).
- Executive committee of DARCLE www.darcle.org.

9.1.2. Scientific Events Selection

9.1.2.1. Reviewer

Invited editor for international conferences: Interspeech, NIPS, ACL, etc. (around 5-10 papers per conferences, 2 conferences per year)

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

Member of the editorial board of: Mathématiques et Sciences Humaines, L'Année Psychologique, Frontiers in Psychology.

9.1.3.2. Reviewer - Reviewing Activities

Invited Reviewer for Frontiers in Psychology, Cognitive Science, Cognition, Transactions in Acoustics Signal Processing and Language, Speech Communication, etc. (around 4 papers per year)

9.1.4. Invited Talks

- Nov/29/2018, E. Dupoux, Invited Department Colloquium, Linguistics, U. Maryland: Reverse Engineering Language Acquisition
- Nov/21/2018, E. Dupoux, Invited Department Colloquium, LORIA, Nancy: Developmental AI
- Oct/17/2018, E. Dupoux, Invited Seminar, Departement Physics ENS& chaire Sciences des Données: Reverse Engineering Cognitive Development
- Jul/4/2018, E. Dupoux, Invited Seminar, PRAIRIE AI Summer School: Unsupervised Speech Technology
- Nov/23/2018, E. Dupoux, Invited Seminar, GDR "Cognitive Neurosciences of Development": What AI can bring to Cognitive Development (and vice versa)
- 2018, N. Zeghidour, invited Seminar, LORIA, Nancy: learning from raw waveforms
- 2018, N. Zeghidour, invited Talk, Legrain Conference on AI and Cognition, Paris: learning from raw waveforms

9.1.5. Scientific Expertise

E. Dupoux is invited expert for ERC, ANR, and other granting agencies, or tenure committees (around 2 per year).

9.1.6. Research Administration

E. Dupoux is on the Executive committee of the Foundation Cognition, the research programme IRIS-PSL "Sciences des Données et Données des Sciences", the industrial chair Almerys (2016-) and the collective organization DARCLE (www.darcle.org).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

- Master : E. Dupoux, "Theoretical Cognitive Science: Connections and symbols", 8h, M1/M2, PSL,Paris 5, Paris France
- Master : E. Dupoux (with B. Sagot, ALMANACH, N. Zeghidour & R. Riad, COML), "Algorithms for speech and language processing", 30h, M2, (MVA), ENS Cachan, France
- Master : E. Dupoux, "Cognitive Engineering", 80h, M2, ITI-PSL, Paris France
- Doctorat : E. Dupoux, "Computational models of cognitive development", 32 h, Séminaire EHESS, Paris France

9.2.2. Supervision

- PhD : Julia Maria Carbajal, Separation and acquisition of two languages in early childhood: a multidisciplinary approach, Ecole Normale Supérieure, sept 21, 2018, co-advised E. Dupoux, S. Peperkamp
- PhD : Adriana Rukoz Gevara, Decoding perceptual epenthesis: Experiments and Modelling., Ecole Normale Supérieure, oct 19, 2018, co-advised E. Dupoux, S. Peperkamp
- PhD in progress : Neil Zeghidour, Learning speech features from raw signals, Feb 2015, co-advised E. Dupoux, N. Usunier (Facebook-CIFRE)
- PhD in progress : Elin Larsen, Models of word learning in infants, Sept 2017, co-advised E. Dupoux, A. Cristia– abandon
- PhD in progress : Rama Chaabouni, Language learning in artificial agents, Sept 2017, co-advised E. Dupoux, M. Baroni (Facebook-CIFRE)
- PhD in progress : Ronan Riochet, Learning models of intuitive physics, Sept 2017, co-advised E. Dupoux, I. Laptev, J. Sivic
- PhD in progress : Rachid Riad, "Speech technology for biomarkers in neurodegenerative diseases" , Sept 2018, co-advised E. Dupoux, A.-C. Bachoud-Lévi

9.2.3. Juries

E. Dupoux participated in the PhD Jury of Andreux Mathieu, Nov 12, ENS, 2018.

9.3. Popularization

E. Dupoux talked in two general public conferences on speech technologies, one organized by the Institut Carnot Cognition (La Vilette, oct, 2018), one by the Institut IA in Toulouse (oct 2018), both with around 200 participants. He gave and/or organized smaller meetings geared towards enhancing contacts between industry and research in the general area of AI and Cognition (1 day and a half of scientific meetings between PSL and Facebook, seminar-style intervention with MSR, and with the CVT Athena). He co-chaired the conference Legrain on AI and Cognition which, besides the scientific track had a general public track and an industry track, which were both attended by 100-200 attendees (see http://olivierlegrain.ens.psl.eu/ia-et-cognition.html).

N. Zeghidour did a high level presentation of AI in Vivatech on the Facebook Stand (100 000 visitors). He presented the state of the art in ASR and TTS in the BNP Paribas-PRAIRIE Summer School with participants from the IT Industry. He co-redacted a 5 pages popularization article on neural networks and deep learning in the magazine "Tangent, the mathematical adventure" for high school students, 20,000 printed copies.

10. Bibliography

Major publications by the team in recent years

[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, 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, 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", Baltimore, Maryland, Association for Computational Linguistics, 2014, vol. 2, pp. 1-6 [*DOI :* 10.3115/V1/P14-2001]

[4] Y. HOSHEN, R. J. WEISS, 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, Y. GOLDBERG. *Assessing the ability of LSTMs to learn syntax-sensitive dependencies*, in "Transactions of the Association for Computational Linguistics", 2016, vol. 4, pp. 521-535

[6] T. LINZEN, E. DUPOUX, B. SPECTOR. *Quantificational features in distributional word representations*, in "Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics", 2016, pp. pages 1 – 1-11 [*DOI :* 10.18653/V1/S16-2001]

[7] A. MARTIN, S. PEPERKAMP, E. DUPOUX. *Learning Phonemes with a Proto-lexicon*, in "Cognitive Science", 2013, vol. 37, 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, 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, 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, E. DUPOUX. *Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline*, in "INTERSPEECH-2013", Lyon, France, International Speech Communication Association, 2013, pp. 1781-1785

[11] R. THIOLLIÈRE, E. DUNBAR, G. SYNNAEVE, M. VERSTEEGH, 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, K. KAVUKCUOGLU. *Wavenet: A generative model for raw audio*, in "CoRR abs/1609.03499", 2016

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] M. J. CARBAJAL. *Separation and acquisition of two languages in early childhood: A multidisciplinary approach*, Université de recherche Paris Sciences et Lettres , September 2018, https://hal.archives-ouvertes.fr/tel-01948483

[14] A. GUEVARA-RUKOZ. *Decoding perceptual vowel epenthesis: Experiments & Modelling*, Ecole Normale Supérieure (ENS), October 2018, https://hal.archives-ouvertes.fr/tel-01948548

Articles in International Peer-Reviewed Journals

[15] E. DUPOUX. *Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner*, in "Cognition", April 2018, vol. 173, pp. 43 - 59, https://arxiv.org/abs/1607.08723 [*DOI :* 10.1016/J.COGNITION.2017.11.008], https://hal.archives-ouvertes.fr/hal-01888694

[16] A. GUEVARA-RUKOZ, A. CRISTIA, B. LUDUSAN, R. THIOLLIÈRE, A. MARTIN, R. MAZUKA, E. DUPOUX. *Are Words Easier to Learn From Infant- Than Adult-Directed Speech? A Quantitative Corpus-Based Investigation*, in "Cognitive Science", July 2018, vol. 42, n^{o} 5, pp. 1586 - 1617 [*DOI :* 10.1111/COGS.12616], https://hal.archives-ouvertes.fr/hal-01888701

[17] T. SCHATZ, F. BACH, E. DUPOUX. *Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception*, in "Journal of the Acoustical Society of America", May 2018, vol. 143, n^{o} 5, pp. EL372 - EL378 [*DOI :* 10.1121/1.5037615], https://hal.archives-ouvertes.fr/hal-01888735

International Conferences with Proceedings

[18] X.-N. CAO, C. DAKHLIA, P. DEL CARMEN, M.-A. JAOUANI, M. OULD-ARBI, E. DUPOUX. *Baby Cloud, a technological platform for parents and researchers*, in "LREC 2018 - 11th edition of the Language Resources and Evaluation Conference", Miyazaki, Japan, Proceedings of LREC 2018, May 2018, https://hal.archives-ouvertes.fr/hal-01948107

[19] N. HOLZENBERGER, M. DU, J. KARADAYI, R. RIAD, E. DUPOUX. *Learning Word Embeddings: Unsupervised Methods for Fixed-size Representations of Variable-length Speech Segments*, in "Interspeech 2018", Hyderabad, India, Proceedings of Interspeech 2018, ISCA, September 2018 [*DOI :* 10.21437/INTERSPEECH.2018-2364], https://hal.archives-ouvertes.fr/hal-01888708

[20] L. ONDEL, P. GODARD, L. BESACIER, E. LARSEN, M. HASEGAWA-JOHNSON, O. SCHARENBORG, E. DUPOUX, L. BURGET, F. YVON, S. KHUDANPUR. *Bayesian Models for Unit Discovery on a Very Low Resource Language*, in "ICASSP 2018", Calgary, Alberta, Canada, Proceedings of ICASSP 2018, April 2018, https://arxiv.org/abs/1802.06053 - Accepted to ICASSP 2018, https://hal.archives-ouvertes.fr/hal-01888718

[21] R. RIAD, C. DANCETTE, J. KARADAYI, N. ZEGHIDOUR, T. SCHATZ, E. DUPOUX. *Sampling strategies in Siamese Networks for unsupervised speech representation learning*, in "Interspeech 2018", Hyderabad, India,

Proceedings of Interspeech 2018, September 2018, https://arxiv.org/abs/1804.11297 - Conference paper at Interspeech 2018, https://hal.archives-ouvertes.fr/hal-01888725

[22] O. SCHARENBORG, L. BESACIER, A. BLACK, M. HASEGAWA-JOHNSON, F. METZE, G. NEUBIG, S. STUKER, P. GODARD, M. MULLER, L. ONDEL, S. PALASKAR, P. ARTHUR, F. CIANNELLA, M. DU, E. LARSEN, D. MERKX, R. RIAD, L. WANG, E. DUPOUX. *Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking rosetta" JSALT 2017 workshop*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Alberta, Canada, April 2018, https://hal.archives-ouvertes.fr/hal-01709578

[23] A. THUAL, C. DANCETTE, J. KARADAYI, J. BENJUMEA, E. DUPOUX. *A K-nearest neighbours approach to unsupervised spoken term discovery*, in "IEEE Spoken Language Technology SLT-2018", Athènes, Greece, Proceedings of SLT 2018, December 2018, https://hal.archives-ouvertes.fr/hal-01947953

[24] N. ZEGHIDOUR, N. USUNIER, I. KOKKINOS, T. SCHATZ, G. SYNNAEVE, E. DUPOUX. *Learning Filterbanks from Raw Speech for Phoneme Recognition*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Alberta, Canada, Proceedings of ICASSP 2018, April 2018, https://arxiv.org/abs/1711.01161v2 - Accepted at ICASSP 2018, https://hal.archives-ouvertes.fr/hal-01888737

[25] N. ZEGHIDOUR, N. USUNIER, G. SYNNAEVE, R. COLLOBERT, E. DUPOUX. *End-to-End Speech Recognition From the Raw Waveform*, in "Interspeech 2018", Hyderabad, India, Proceedings of Interspeech 2018, September 2018, https://arxiv.org/abs/1806.07098 - Accepted for presentation at Interspeech 2018 [*DOI :* 10.21437/INTERSPEECH.2018-2414], https://hal.archives-ouvertes.fr/hal-01888739

Conferences without Proceedings

[26] A. DÉFOSSEZ, N. ZEGHIDOUR, N. USUNIER, L. BOTTOU, F. BACH. *SING: Symbol-to-Instrument Neural Generator*, in "Conference on Neural Information Processing Systems (NIPS)", Montréal, Canada, December 2018, https://arxiv.org/abs/1810.09785 , https://hal.archives-ouvertes.fr/hal-01899949

References in notes

[27] D. A. FERRUCCI. *Introduction to "this is watson"*, in "IBM Journal of Research and Development", 2012, vol. 56, n^{o} 3.4, pp. 1–1

[28] K. HE, X. ZHANG, S. REN, 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

[29] J. HERNÁNDEZ-ORALLO, F. MARTÍNEZ-PLUMED, U. SCHMID, M. SIEBERS, D. L. DOWE. *Computer models solving intelligence test problems: Progress and implications*, in "Artificial Intelligence", 2016, vol. 230, pp. 74–107

[30] B. M. LAKE, T. D. ULLMAN, J. B. TENENBAUM, S. J. GERSHMAN. *Building machines that learn and think like people*, in "arXiv preprint arXiv:1604.00289", 2016

[31] C. LU, X. TANG. *Surpassing human-level face verification performance on LFW with GaussianFace*, in "arXiv preprint arXiv:1404.3840", 2014

[32] S. T. MUELLER. *A partial implementation of the BICA cognitive decathlon using the Psychology Experiment Building Language (PEBL)*, in "International Journal of Machine Consciousness", 2010, vol. 2, n^{o} 02, pp. 273–288

[33] 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, D. HASSABIS. *Mastering the game of Go with deep neural networks and tree search*, in "Nature", 2016, vol. 529, n^{o} 7587, pp. 484–489

[34] I. SUTSKEVER, O. VINYALS, Q. V. LE. *Sequence to sequence learning with neural networks*, in "Advances in neural information processing systems", 2014, pp. 3104–3112

[35] A. M. TURING. *Computing machinery and intelligence*, in "Mind", 1950, vol. 59, n^{o} 236, pp. 433–460

[36] W. XIONG, J. DROPPO, X. HUANG, F. SEIDE, M. SELTZER, A. STOLCKE, D. YU, G. ZWEIG. *Achieving human parity in conversational speech recognition*, in "arXiv preprint arXiv:1610.05256", 2016