RESEARCH CENTRE

Paris

2020 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)

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]

PhD Students

- Alafate Abulimiti [Inria, from Oct 2020]
- Robin Algayres [École Normale Supérieure de Paris]
- Rahma Chaabouni [École Normale Supérieure de Paris]
- Maureen De Seyssel [École Normale Supérieure de Paris, from Oct 2020]
- Marvin Lavechin [École Normale Supérieure de Paris]
- Juliette Millet [École Normale Supérieure de Paris]
- Yann Raphalen [Inria, from Oct 2020]
- Rachid Riad [École Normale Supérieure de Paris]

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]
- Julien Karadayi [École Normale Supérieure de Paris, Engineer, until Oct 2020]
- Manel Khentout [École Normale Supérieure de Paris, Engineer]
- Hadrien Titeux [École Normale Supérieure de Paris, Engineer]
- Gwendal Virlet [École Normale Supérieure de Paris, Engineer, from Sep 2020]
- Mohamed Zaiem [Université de Paris, Engineer, until Aug 2020]

Interns and Apprentices

- Elias Aouad [École Normale Supérieure de Paris, from Mar 2020 until Jul 2020]
- Ruben Bouzbib [École Normale Supérieure de Paris, until Feb 2020]
- Maureen De Seyssel [École Normale Supérieure de Paris, until Apr 2020]
- Emma Ducos [École Normale Supérieure de Paris, from Apr 2020 until Oct 2020]
- Lucas Elbert [Inria, from Apr 2020 until Aug 2020]
- Louis Fournier [Université de Paris, from Apr 2020 until Sep 2020]
- Adel Nabli [École Normale Supérieure de Paris, from Jul 2020]
- Tu Anh Nguyen [Facebook, from Apr 2020]

- Maxime Poli [École Normale Supérieure de Paris, from Jun 2020]
- Tristan Ricoul [École Normale Supérieure de Paris, from Apr 2020 until Sep 2020]
- Mathieu Rita [École Normale Supérieure de Paris, from Apr 2020 until Aug 2020]

Administrative Assistants

- Meriem Guemair [Inria]
- Catherine Urban [École Normale Supérieure de Paris]

External Collaborator

- Ewan Dunbar [Université de Paris]

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., [41, 45, 50, 40, 47]). 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) [43].

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 [48].

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 [49] 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* [42] , i.e., functional components within a more global cognitive architecture [46]. 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], [44].

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 New software

5.1.1 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

5.1.2 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

5.1.3 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

5.1.4 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

6 New results

6.1 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 benchmarks and algorithms in order to push the field towards self-supervised learning.

- In [18], we present the results of the **Zero Resource Speech Challenge 2020** (special session at Interspeech 2020), which takes aims at learning speech representations from raw audio signals without any labels. The challenge combines the data sets and metrics from two previous benchmarks (2017 and 2019) and features two tasks which tap into two levels of speech representation. The first task is to discover low bit-rate subword representations that optimize the quality of speech synthesis; the second one is to discover word-like units from unsegmented raw speech. We present the results of the twenty submitted models and discuss the implications of the main findings for unsupervised speech learning.

- In [28], we introduce a new unsupervised task, **spoken language modeling**, which consists in the learning of linguistic representations from raw audio signals without any labels. The task is evaluated with a suite of 4 black-box, zero-shot metrics probing for the quality of the learned models at 4 linguistic levels: phonetics, lexicon, syntax and semantics. We present the results and analyses of a composite baseline made of the concatenation of three unsupervised systems: self-supervised contrastive representation learning (CPC), clustering (k-means) and language modeling (LSTM or BERT). The language models learn on the basis of the pseudo-text derived from clustering the learned representations. This simple pipeline shows better than chance performance on all four metrics, demonstrating the feasibility of spoken language modeling from raw speech. It also yields

worse performance compared to text-based 'topline' systems trained on the same data, delineating the space to be explored by more sophisticated end-to-end models. This task and baseline is to be part of the **Zero Resource Speech Benchmark 2021** (Interspeech 2021).

- Cross-lingual and multilingual training of Automatic Speech Recognition (ASR) has been extensively investigated in the supervised setting. This assumes the existence of a parallel corpus of speech and orthographic transcriptions. Recently, contrastive predictive coding (CPC) algorithms have been proposed to pretrain ASR systems with unlabelled data. In [34] we investigate whether **unsupervised pretraining transfers well across languages**. We show that a slight modification of the CPC pretraining extracts features that transfer well to other languages, being on par or even outperforming supervised pretraining. This shows the potential of unsupervised methods for languages with few linguistic resources.

- Recent work on unsupervised contrastive learning of speech representation has shown promising results, but so far has mostly been applied to clean, curated speech datasets. Can it also be used with unprepared audio data **"in the wild"**? In [33], we explore three potential problems in this setting: (i) presence of non-speech data, (ii) noisy or low quality speech data, and (iii) imbalance in speaker distribution. We show that on the Libri-light train set, which is itself a relatively clean speech-only dataset, these problems combined can already have a performance cost of up to 30% relative for the ABX score. We show that the first two problems can be alleviated by data filtering, with voice activity detection selecting speech segments, while perplexity of a model trained with clean data helping to discard entire files. We show that the third problem can be alleviated by learning a speaker embedding in the predictive branch of the model. We show that these techniques build more robust speech features that can be transferred to an ASR task in the low resource setting.

- Contrastive Predictive Coding (CPC), based on predicting future segments of speech based on past segments is emerging as a powerful algorithm for representation learning of speech signal. However, it still under-performs other methods on unsupervised evaluation benchmarks. In [24], we introduce **WavAugment**, a time-domain data augmentation library and find that applying augmentation in the past is generally more efficient and yields better performances than other methods. We find that a combination of pitch modification, additive noise and reverberation substantially increase the performance of CPC (relative improvement of 18-22%), beating the reference Libri-light results with 600 times less data. Using an out-of-domain dataset, time-domain data augmentation can push CPC to be on par with the state of the art on the Zero Speech Benchmark 2017. We also show that time-domain data augmentation consistently improves downstream limited-supervision phoneme classification tasks by a factor of 12-15% relative.

- In [23], we introduce **Libri-light**, a new collection of spoken English audio suitable for training speech recognition systems under limited or no supervision. It is derived from open-source audio books from the LibriVox project. It contains over 60K hours of audio , which is, to our knowledge, the largest freely-available corpus of speech. The audio has been segmented using voice activity detection and is tagged with SNR, speaker ID and genre descriptions. Additionally, we provide baseline systems and evaluation metrics working under three settings: (1) the zero resource/unsupervised setting (ABX), (2) the semi-supervised setting (PER, CER) and (3) the distant supervision setting (WER). Settings (2) and (3) use limited textual resources (10 minutes to 10 hours) aligned with the speech. Setting (3) uses large amounts of unaligned text. They are evaluated on the standard LibriSpeech dev and test sets for comparison with the supervised state-of-the-art. Index Terms-unsupervised and semi-supervised learning , distant supervision, dataset, zero-and low resource ASR.

6.2 Language emergence in communicative agents

In this relatively new research topic, which is currently the focus of Rahma Chaabouni's PhD thesis, we study the inductive biases of neural systems by presenting them with few or no data.

- Previous work has shown that artificial neural agents naturally develop surprisingly non-efficient codes. This is illustrated by the fact that in a referential game involving a speaker and a listener

neural networks optimizing accurate transmission over a discrete channel, the emergent messages fail to achieve an optimal length. Furthermore, frequent messages tend to be longer than infrequent ones, a pattern contrary to the **Zipf Law of Abbreviation** (ZLA) observed in all natural languages. In [32], we show that near-optimal and ZLA-compatible messages can emerge, but only if both the speaker and the listener are modified. We hence introduce a new communication system, "LazImpa", where the speaker is made increasingly lazy, i.e. avoids long messages, and the listener impatient, i.e., seeks to guess the intended content as soon as possible.

- Natural language allows us to refer to novel composite concepts by combining expressions denoting their parts according to systematic rules, a property known as **compositionality**. In [16], we study whether the language emerging in deep multi-agent simulations possesses a similar ability to refer to novel primitive combinations, and whether it accomplishes this feat by strategies akin to human-language compositionality. Equipped with new ways to measure compositionality in emergent languages inspired by disentan-glement in representation learning, we establish three main results. First, given sufficiently large input spaces, the emergent language will naturally develop the ability to refer to novel composite concepts. Second, there is no correlation between the degree of compositional-ity of an emergent language and its ability to generalize. Third, while compositionality is not necessary for generalization, it provides an advantage in terms of language transmission: The more compositional a language is, the more easily it will be picked up by new learners, even when the latter differ in architecture from the original agents. We conclude that compositionality does not arise from simple generalization pressure, but if an emergent language does chance upon it, it will be more likely to survive and thrive.

6.3 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 methods based on psycholinguistic/linguistic criteria, and deploy them for systematic comparison of systems.

- In [15], we study **spoken word embeddings**, which are fixed-size acoustic representations of variable-length audio sequences and we systematically compare two popular metrics for the quality of such embeddings: ABX discrimination and Mean Average Precision (MAP), on 5 languages across 17 embedding methods, ranging from supervised to fully unsu-pervised, and using different loss functions (autoencoders, cor-respondance autoencoders, siamese). Then we use the ABX and MAP to predict performances on a new downstream task: the unsupervised estimation of the frequencies of speech segments in a given corpus. We find that overall, ABX and MAP correlate with one another and with frequency estimation. However, substantial discrepancies appear in the fine-grained distinctions across languages and/or embedding methods. This makes it un-realistic at present to propose a task-independent silver bullet method for computing the intrinsic quality of speech embed-dings. There is a need for more detailed analysis of the metrics currently used to evaluate such embeddings.

- **Vector space models of words** have long been claimed to capture linguistic regularities as simple vector translations, but problems have been raised with this claim. In [20], we decompose and empirically analyze the classic arithmetic word analogy test, to motivate two new metrics that address the issues with the standard test, and which distinguish between class-wise offset concentration (similar directions between pairs of words drawn from different broad classes, such as France-London, China-Ottawa,. . .) and pairing consistency (the existence of a regular transformation between correctly-matched pairs such as France:Paris::China:Beijing). We show that, while the standard analogy test is flawed, several popular word embeddings do nevertheless encode linguistic regularities.

- **Reconstruction of articulatory trajectories** from the acoustic speech signal has been proposed for improving speech recognition and text-to-speech synthesis. However, to be useful in these settings,

articulatory reconstruction must be speaker independent. Furthermore, as most research focuses on single, small data sets with few speakers, robust articulatory reconstruction could profit from combining data sets. Standard evaluation measures such as root mean squared error and Pearson correlation are inappropriate for evaluating the speaker independence of models or the usefulness of combining data sets. In [29], we present a new evaluation for articulatory reconstruction which is independent of the articulatory data set used for training: the phone discrimination ABX task. We use the ABX measure to evaluate a bi-LSTM based model trained on three data sets (14 speakers), and show that it gives information complementary to standard measures, enabling us to evaluate the effects of data set merging, as well as the speaker independence of the model.

6.4 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.

- 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. In [36], 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.

- **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 [38] 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.

- **Disfluent speech** has been previously addressed from two main perspectives: the clinical perspective focusing on diagnostic, and the Natural Language Processing (NLP) perspective aiming at modeling these events and detect them for downstream tasks. In addition, previous works

often used different metrics depending on whether the input features are text or speech, making it difficult to compare the different contributions. In [30], we introduce a new evaluation framework for disfluency detection inspired by the clinical and NLP perspective together with the theory of performance from (Clark, 1996) which distinguishes between primary and collateral tracks. We introduce a novel forced-aligned disfluency dataset from a corpus of semi-directed interviews, and present baseline results directly comparing the performance of text-based features (word and span information) and speech-based (acoustic-prosodic information). Finally, we introduce new audio features inspired by the word-based span features. We show experimentally that using these features outperformed the baselines for speech-based predictions on the present dataset.

6.5 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).

- In [19], we explore the minimal knowledge a listener needs to compensate for **phonological assimilation, one kind of phonological process responsible for variation in speech.** We used standard automatic speech recognition models to represent English and French listeners. We found that, first, some types of models show language-specific assimilation patterns comparable to those shown by human listeners. Like English listeners, when trained on English, the models compensate more for place assimilation than for voicing assimilation, and like French listeners, the models show the opposite pattern when trained on French. Second, the models which best predict the human pattern use contextually-sensitive acoustic models and language models, which capture allophony and phonotactics, but do not make use of higher-level knowledge of a lexicon or word boundaries. Finally, some models overcompensate for assimilation, showing a (super-human) ability to recover the underlying form even in the absence of the triggering phonological context, pointing to an incomplete neutralization not exploited by human listeners.

- **The language discrimination process** in infants has been successfully modeled using i-vector based systems, with results replicating several experimental findings. Still, recent work found intriguing results regarding the difference between monolingual and mixed-language exposure on language discrimination tasks. In [17], we use two carefully designed datasets, with an additional "bilingual" condition on the i-vector model of language discrimination. Our results do not show any difference in the ability of discriminating languages between the three backgrounds, although we do replicate past observations that distant languages (English-Finnish) are easier to discriminate than close languages (English-German). We do, however, find a strong effect of background when testing for the ability of the learner to automatically sort sentences in language clusters: bilingual background being generally harder than mixed background (one speaker one language). Other analyses reveal that clustering is dominated by speakers information rather than by languages.

- Disease-modifying treatments are currently assessed in neurodegenerative diseases. **Huntington's Disease** represents a unique opportunity to design automatic sub-clinical markers, even in pre-manifest gene carriers. In [31] we investigated phonatory impairments as potential clinical markers and propose them for both diagnosis and gene carriers follow-up. We used two sets of features: Phonatory features and Modulation Power Spectrum Features. We found that phonation is not sufficient for the identification of sub-clinical disorders of premanifest gene carriers. According to our regression results, Phonatory features are suitable for the predictions of clinical performance in Huntington's Disease.

- In [27], we present **the Perceptimatic English Benchmark,** an open experimental benchmark for evaluating quantitative models of speech perception in English. The benchmark consists of ABX stimuli along with the responses of 91 American Englishspeaking listeners. The stimuli test discrimination of a large number of English and French phonemic contrasts. They are extracted directly from corpora of read speech, making them appropriate for evaluating statistical acoustic models (such as those used in automatic speech recognition) trained on typical speech data sets. We show that phone discrimination is correlated with several types of models, and give recommendations for

researchers seeking easily calculated norms of acoustic distance on experimental stimuli. We show that DeepSpeech, a standard English speech recognizer, is more specialized on English phoneme discrimination than English listeners, and is poorly correlated with their behaviour, even though it yields a low error on the decision task given to humans.

- In [26], we present a data set and methods to compare speech processing models and human behaviour on a phone discrimination task. We provide **Perceptimatic**, an open data set which consists of French and English speech stimuli, as well as the results of 91 English-and 93 French-speaking listeners. The stimuli test a wide range of French and English contrasts, and are extracted directly from corpora of natural running read speech, used for the 2017 Zero Resource Speech Challenge. We provide a method to compare humans' perceptual space with models' representational space, and we apply it to models previously submitted to the Challenge. We show that, unlike unsupervised models and supervised multilingual models, a standard supervised monolingual HMM-GMM phone recognition system, while good at discriminating phones, yields a representational space very different from that of human native listeners.

6.6 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.

- Spontaneous conversations in real-world settings such as those found in **child-centered recordings** have been shown to be amongst the most challenging audio files to process. Nevertheless, building speech processing models handling such a wide variety of conditions would be particularly useful for language acquisition studies in which researchers are interested in the quantity and quality of the speech that children hear and produce, as well as for early diagnosis and measuring effects of remediation. In [25], we present our approach to designing an open-source neural network to classify audio segments into vocalizations produced by the child wearing the recording device, vocalizations produced by other children, adult male speech, and adult female speech. To this end, we gathered diverse child-centered corpora which sums up to a total of 260 hours of recordings and covers 10 languages. Our model can be used as input for downstream tasks such as estimating the number of words produced by adult speakers, or the number of linguistic units produced by children. Our architecture combines SincNet filters with a stack of recurrent layers and outperforms by a large margin the state-of-the-art system, the Language ENvironment Analysis (LENA) that has been used in numerous child language studies.

- In [35], we introduce **Seshat, a new, simple and open-source software to efficiently manage annotations of speech corpora.** The Seshat software allows users to easily customise and manage annotations of large audio corpora while ensuring compliance with the formatting and naming conventions of the annotated output files. In addition, it includes procedures for checking the content of annotations following specific rules are implemented in personalised parsers. Finally, we propose a double-annotation mode, for which Seshat computes automatically an associated inter-annotator agreement with the γ measure taking into account the categorisation and segmentation discrepancies.

7 Bilateral contracts and grants with industry

- **Facebook AI Research Grant** (2020, PI: E. Dupoux, 350K€) - Unrestricted Gift - The aim is to help the development of machine learning tools geared towards the psycholinguistic research community.

8 Partnerships and cooperations

8.1 National initiatives

8.1.1 ANR

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

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

E. D.upoux and E. Dunbar organized the ZeroSpeech Challengee 2020 (Challenge and Special Session, Interspeech 2020)

9.2 Teaching - Supervision - Juries

9.2.1 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
- 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
- Master: E. Dunbar, "Phonology" , 36 h, Master Sciences du Langage, Paris Diderot
- Master: E. Dunbar, "Statistics", 28h, Master Sciences du Langage, Paris Diderot
- Licence 3: E. Dunbar, "Phonology", 36h, Licence Sciences du Langage, Paris Diderot
- Licence 3: E. Dunbar, "Experimental methods", 36h, Licence Sciences du Langage, Paris Diderot

9.2.2 Supervision

- PhD in progress : Rahma Chaabouni, Language learning in artificial agents, Sept 2017, co-advised E. Dupoux, M. Baroni (Facebook-CIFRE); to be defended in March 2021
- PhD in progress : Ronan Riochet, Learning models of intuitive physics, Sept 2017, co-advised E. Dupoux, I. Laptev, J. Sivic; to be defended in May 2021

- PhD in progress : Rachid Riad, "Speech technology for biomarkers in neurodegenerative diseases" , Sept 2018, co-advised E. Dupoux, A.-C. Bachoud-Levi
- PhD in progress : Robin Algayres "Audio word embeddings and word segmentation" , from Oct 2019, co-advised E. Dupoux, B. Sagot
- PhD in progress: Juliette Millet, "Modeling L2 Speech perception", from Sept 2018, advised E. Dunbar Bachoud-Lévi
- PhD in progress: Maureen de Seyssel, "Modeling bilingual language acquisition", from Sept 2020, co-advised E. Dupoux, G. Wisniewski

10 Scientific production

10.1 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).

10.2 Publications of the year

International journals

[13] J. Cassell. 'The ties that bind: Social Interaction in Conversational Agents'. In: *Réseaux* 220-221.2-3 (Mar. 2020), pp. 21–45. DOI: 10.3917/res.220.0021. URL: https://hal.archives-ouvertes.fr/hal-02985286.

[14] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stuker, P. Godard, M. Müller, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang and E. Dupoux. 'Speech technology for unwritten languages'. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* (13th Feb. 2020). DOI: 10.1109/TASLP.2020.2973896. URL: https://hal.inria.fr/hal-02480675.

International peer-reviewed conferences

[15] R. Algayres, M. S. Zaiem, B. Sagot and E. Dupoux. 'Evaluating the reliability of acoustic speech embeddings'. In: INTERSPEECH 2020 - Annual Conference of the International Speech Communication Association. Shanghai / Vitrtual, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02977539.

[16] R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux and M. Baroni. 'Compositionality and Generalization in Emergent Languages'. In: ACL 2020 - 8th annual meeting of the Association for Computational Linguistics. Seattle / Virtual, United States, 5th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02959466.

[17] M. De Seyssel and E. Dupoux. 'Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors'. In: CogSci 2020 - 42nd Annual Virtual Meeting of the Cognitive Science Society. Toronto / Virtual, Canada, 29th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02959451.

[18] E. Dunbar, J. Karadayi, M. Bernard, X.-N. Cao, R. Algayres, L. Ondel, L. Besacier, S. Sakti and E. Dupoux. 'The Zero Resource Speech Challenge 2020: Discovering discrete subword and word units'. In: Interspeech 2020 - Conference of the International Speech Communication Association. Shangai / Virtual, China, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02962224.

[19] B. '. Er Jiang, E. Dunbar, M. Sonderegger, M. Clayards and E. Dupoux. 'Modelling Perceptual Effects of Phonology with ASR Systems'. In: CogSci 2020 - 42nd Annual Virtual Meeting of the Cognitive Science Society. Virtual, France, 29th July 2020. URL: https://hal.archives-ouvertes.fr/hal-03070281.

[20] L. Fournier, E. Dupoux and E. Dunbar. 'Analogies minus analogy test: measuring regularities in word embeddings'. In: CoNLL 2020 - 24th Conference on Computational Natural Language Learning. Virtual, France, 19th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070260.

[21] P. García, J. Villalba, H. Bredin, J. Du, D. Castan, A. Cristia, L. Bullock, L. Guo, K. Okabe, P. S. Nidadavolu, S. Kataria, S. Chen, L. Galmant, M. Lavechin, L. Sun, M.-P. Gill, B. Ben-Yair, S. Abdoli, X. Wang, W. Bouaziz, H. Titeux, E. Dupoux, K. A. Lee and N. Dehak. 'Speaker detection in the wild: Lessons learned from JSALT 2019'. In: Odyssey 2020 The Speaker and Language Recognition Workshop. Tokyo, Japan, 1st Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02417632.

[22] L. Gautheron, M. Lavechin, R. Riad, C. Scaff and A. Cristia. 'Longform recordings : Opportunities and challenges'. In: *Actes des 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*. LIFT 2020 - 2èmes journées scientifiques du Groupement de Recherche "Linguistique informatique, formelle et de terrain". Montrouge / Virtual, France, 2020, pp. 64–71. URL: https://hal.archives-ouvertes.fr/hal-03047153.

[23] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fügen, T. Likhomanenko, G. Synnaeve, A. Joulin, M. I. Abdelrahman and E. Dupoux. 'LIBRI-LIGHT: a benchmark for asr with limited or no supervision'. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona / Virtual, Spain, 4th May 2020, pp. 7669–7673. DOI: 10.1109/ICASSP40776.2020.9052942. URL: https://hal.archives-ouvertes.fr/hal-02959460.

[24] E. Kharitonov, M. Rivière, G. Synnaeve, L. Wolf, P.-E. Mazaré, M. Douze and E. Dupoux. 'Data Augmenting Contrastive Learning of Speech Representations in the Time Domain'. In: SLT 2020 - IEEE Spoken Language Technology Workshop. Shenzhen / Virtual, China, 13th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070321.

[25] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux and A. Cristia. 'An open-source voice type classifier for child-centered daylong recordings'. In: Interspeech 2020 - Conference of the International Speech Communication Association. Shanghai / Virtual, China, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02989487.

[26] J. Millet and E. Dunbar. 'Perceptimatic: A human speech perception benchmark for unsupervised subword modelling'. In: Interspeech 2020 - 21st Annual Conference of the International Speech Communication Association. Proceedings of INTERSPEECH 2020, 21st Annual Conference of the International Speech Communication Association. Shanghai / Virtual, China: http://www.interspeech2020.org/, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-03087252.

[27] J. Millet and E. Dunbar. 'The Perceptimatic English Benchmark for Speech Perception Models'. In: CogSci 2020 - 42nd Annual Virtual Meeting of the Cognitive Science Society. Proceedings of the 42nd Annual Meeting of the Cognitive Science Society. Toronto / Virtual, Canada, 29th July 2020. URL: https://hal.archives-ouvertes.fr/hal-03087248.

[28] T. A. Nguyen, M. De Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar and E. Dupoux. 'The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling'. In: NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing. Virtuel, France, 6th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070362.

[29] M. Parrot, J. Millet and E. Dunbar. 'Independent and Automatic Evaluation of Speaker-Independent Acoustic-to-Articulatory Reconstruction'. In: Interspeech 2020 - 21st Annual Conference of the International Speech Communication Association. Proceedings of INTERSPEECH 2020, 21st Annual Conference of the International Speech Communication Association. Shanghai / Virtual, China: http://www.interspeech2020.org/, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-03087264.

[30] R. Riad, A.-C. Bachoud-Lévi, F. Rudzicz and E. Dupoux. 'Identification of primary and collateral tracks in stuttered speech'. In: LREC 2020 - 12th Conference on Language Resources and Evaluation. Marseille, France, 11th May 2020. URL: https://hal.archives-ouvertes.fr/hal-02959454.

[31] R. Riad, H. Titeux, L. Lemoine, J. Montillot, J. Hamet Bagnou, X. N. Cao, E. Dupoux and A.-C. Bachoud-Lévi. 'Vocal markers from sustained phonation in Huntington's Disease'. In: INTERSPEECH 2020 - Annual Conference of the International Speech Communication Association. Shanghai / Virtual, China, 25th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070388.

[32] M. Rita, R. Chaabouni and E. Dupoux. '"LazImpa": Lazy and Impatient neural agents learn to communicate efficiently'. In: CONLL 2020 - The SIGNLL Conference on Computational Natural Language Learning. Virtual, France, 5th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070404.

[33] M. Rivière and E. Dupoux. 'Towards unsupervised learning of speech features in the wild'. In: SLT 2020 : IEEE Spoken Language Technology Workshop. Shenzhen / Virtual, China, 13th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070411.

[34] M. Rivière, A. Joulin, P.-E. Mazaré and E. Dupoux. 'Unsupervised pretraining transfers well across languages'. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona / Virtual, Spain, 4th May 2020, pp. 7414–7418. DOI: 10.1109/ICASSP40776.2020.9054548. URL: https://hal.archives-ouvertes.fr/hal-02959418.

[35] H. Titeux, R. Riad, X.-N. Cao, N. Hamilakis, K. Madden, A. Cristia, A.-C. Bachoud-Lévi and E. Dupoux. 'Seshat: A tool for managing and verifying annotation campaigns of audio data'. In: LREC 2020 - 12th Language Resources and Evaluation Conference. Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France, 11th May 2020, pp. 6976–6982. URL: https://hal.archives-ouvertes.fr/hal-02496041.

Reports & preprints

[36] B. Ludusan, R. Mazuka and E. Dupoux. *Does infant-directed speech help phonetic learning? A machine learning investigation.* 17th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03080098.

[37] 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.

[38] 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.* 7th Aug. 2020. DOI: 10.31234/osf.io/fc4wh. URL: https://hal.archives-ouvertes.fr/hal-03070566.

[39] H. Titeux and R. Riad. *pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python.* 18th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03144116.

10.3 Cited publications

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

[41] 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.

[42] 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.

[43] 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).

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

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

[50] 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).