Master's Internships 2025/2026

A limited number of positions for a Master's internship are available this year. These internships should be for A MINIMUM OF 5-6 MONTHS. Please send your application by email (CV + cover letter required) to ***syntheticlearner@gmail.com*** **(unless another email is specified).** ***Indicate in the email which topic(s)[1] you are interested in.***

Subject 1: Self-supervised speech representation learning on synthetic data (Maxime Poli)

Subject 2: Probing and benchmarking robustness of self-supervised speech representations (Maxime Poli)

Subject 3: Emergent Communication for Graph-Structured Information: From One-Hot Vectors to Topological Structures (Jean-Baptiste Sevestre)

Subject 4: On edge-device Voice Type Classification (Tarek Kunze, Théo Charlot)

[1] For the latest list of topics, see https://cognitive-ml.fr/docs/2026_CoML_Internship_projects.pdf

S1: Self-supervised speech representation learning on synthetic data

Recent advances in Self-supervised Speech Representation Learning (SSL) [1,2,3] have enabled the development of label-free representations that are valuable for various downstream tasks [4]. Those representations encode linguistic information directly extractable, would it be paralinguistics [5], phones [6] or word segment information [6,7]. In particular, recent models have representations that discriminate phonemes and triphones extremely well, even though the representations are still not invariant to the context [8].

This raises the following question: why is such linguistic information so easily accessible? Is this inherent to the masked representation learning objective? Or is this an artifact of training on speech gathered from audiobooks? This can be investigated by training models in the same way but on different datasets generated synthetically in a controlled way. Previous works [9,10] have shown that SSL models trained on natural noise still have representations that can discriminate phonemes, and that training only on synthetic audio from a synthesizer is possible and can match training on natural sounds [12].

The internship will go through the full pipeline of training SSL models: building the datasets, training the models and evaluating them thoroughly. The datasets generated will go from random noise, to natural noise, animal sounds, synthesized speech, and natural speech. There can also be specific controls, such as a given phonemic contrast, that can be inserted or not in the dataset, and then evaluated downstream. Finally, comprehensive evaluations will be done by probing for phoneme and word information, and by comparing those to results on downstream tasks [4].

Junior supervision: *Maxime Poli*

Senior supervision: *Emmanuel Dupoux*

References

[1] Mohamed, A., Lee, H., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T. N., & Watanabe, S. (2022). Self-Supervised Speech Representation Learning : A Review. IEEE Journal of Selected Topics in Signal Processing.

[2] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems.

[3] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[4] Yang, S., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., & Lee, H. (2021). SUPERB : Speech Processing Universal PERformance Benchmark. Interspeech.

[5] Y. Li, Y. Mohamied, P. Bell and C. Lai Exploration of a Self-Supervised Speech Model: A Study on Emotional Corpora 2022 IEEE Spoken Language Technology Workshop (SLT).

[6] Pasad, A., Chou, J.-C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 914-921.

[7] Pasad, A., Chien, C.-M., Settle, S., & Livescu, K (2024); What Do Self-Supervised Speech Models Know About Words?. Transactions of the Association for Computational Linguistics

[8] Poli M., Chemla E., Dupoux E.. 2024. Improving Spoken Language Modeling with Phoneme Classification: A Simple Fine-tuning Approach. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing

[9] Millet, J., & Dunbar, E. (2022). Do self-supervised speech models develop human-like perception biases? *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[10] Poli, M., Schatz, T., Dupoux, E., & Lavechin, M. (2024). Modeling the initial state of early phonetic learning in infants. Language Development Research.

[12] Cherep M., Singh N. (2024) Contrastive Learning from Synthetic Audio Doppelgängers

S2: Probing and benchmarking robustness of self-supervised speech representations

This project starts from the same premise as Project S1: with self-supervised speech representation learning, we now have access to models that build contextual representations powerful enough for downstream tasks. Probing them reveals that they encode phonemic information at a first level. But is that all there is? Do they encode higher-order information about language, such as lexical or syntactic information? This project aims to develop a better understanding of the representation space of self-supervised representation models (wav2vec 2.0 [1], HuBERT [2], etc.). There are two main axes of work.

The first dimension will be to investigate what is directly extractable beyond phonemic information, whether that includes finer phonetic details and articulatory features, or alternatively word-level [3] and syntactic information. Previous work has found orthogonality between speaker and phonetic information [4, 5] and has explored information completeness and accessibility in discretized representations [6]. We will employ evaluation methods including ABX discrimination tasks [7,8], mean average precision (MAP) over words [9], and information-theoretic metrics [2,6].

The second direction of work would be to understand how robust these representations are to variations in input stimuli, in terms of speaker and acoustic conditions. Previous work has already focused on building models that are more invariant [10,11], but there is a need for a deeper understanding of the existing variability. We want both to understand the structure of the representation space and to find a way to represent the model's uncertainty. This will require building an evaluation framework using methods such as Monte Carlo Dropout [12].

This internship topic does not involve pretraining speech models and will focus more on building careful evaluation protocols in order to contribute to a deeper understanding of self-supervised learning.

Junior supervision: *Maxime Poli*
Senior supervision: *Emmanuel Dupoux*

References

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems.
[2] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing.
[3] Pasad, A., Chien, C.-M., Settle, S., & Livescu, K (2024); What Do Self-Supervised Speech Models Know About Words?. Transactions of the Association for Computational Linguistics
[4] Liu, O. D., Tang, H., & Goldwater, S. (2023). Self-supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces.
[5] Mohamed, M., Liu, O. D., Tang, H., & Goldwater, S. (2024). Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations
[6] Yeh, S.-L., & Tang, H. (2024). Estimating the completeness of discrete speech units. 2024 IEEE SLT
[7] Schatz, T. (2016). ABX-Discriminability Measures and Applications
[8] Poli, M., Chemla, E., & Dupoux, E. (2025). fastabx : A library for efficient computation of ABX discriminability
[9] Carlin, M. A., Thomas, S., Jansen, A., & Hermansky, H. (2011). Rapid evaluation of speech representations for spoken term discovery.
[10] Chang, H.-J., Liu, A. H., & Glass, J. (2023). Self-supervised Fine-tuning for Improved Content Representations by Speaker-invariant Clustering. INTERSPEECH 2023
[11] Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.-I., Cox, D., Hasegawa-Johnson, M., & Chang, S. (2022). ContentVec : An Improved Self-Supervised Speech Representation by Disentangling Speakers. ICML
[12] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation : Representing Model Uncertainty in Deep Learning. Proceedings of The 33rd International Conference on Machine Learning ICML

S3: Emergent Communication for Graph-Structured Information: From One-Hot Vectors to Topological Structures

Context and Motivation

Current paradigms in emergent communication research have made significant progress in understanding how artificial agents can develop communication protocols [1]. However, these approaches remain constrained by simplistic semantic representations—typically using discrete categorical spaces that are small-scale, explicitly hand-coded, and structurally limited. This stands in stark contrast to human language, which operates on complex semantic structures (likely graph-based), vast representational capacity, and implicit semantic spaces that themselves emerge through use.

This internship focuses on systematically extending emergent communication beyond these limitations by investigating structural complexity. The goal is to address fundamental questions about how communication shapes reasoning and how agents can coordinate on complex conceptual representations rather than merely exchanging low-level perceptual features.

Research Question

How can emergent communication protocols effectively encode and transmit complex structural information, such as graphs, and what linguistic properties emerge when the semantic space has a non-linear topology? [2]

This question is motivated by the fundamental challenge of linearization: how can multi-dimensional, graph-structured information be mapped onto a sequential communication channel? This mirrors the problem faced by human speakers when converting complex conceptual structures into linear utterances.

Objectives & Methodology

We will develop a communication framework where agents must transmit graph-structured information through a linear communication channel. This introduces the challenge of mapping a multi-dimensional structure onto a sequential message. Building on the approach of [3] and [4], we will design tasks requiring agents to reconstruct graph topologies from messages.

Junior supervision: *Jean-Baptiste Sevestre*

Senior supervision: *Emmanuel Dupoux*

References

[1] Jannik Peters, Constantin Waubert de Puiseau, Hasan Tercan, Arya Gopikrishnan, Gustavo Adolpho Lucas De Carvalho, Christian Bitter, and Tobias Meisen. A survey on emergent language. *arXiv preprint arXiv:2409.02645*, 2024.
[2] Willem JM Levelt. The speaker's linearization problem. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 295(1077):305–315, 1981.
[3] Agnieszka Słowik, Abhinav Gupta, William L Hamilton, Mateja Jamnik, and Sean B Holden. Towards graph representation learning in emergent communication. *arXiv preprint arXiv:2001.09063*, 2020.
[4] Agnieszka Słowik, Abhinav Gupta, William L Hamilton, Mateja Jamnik, Sean B Holden, and Christopher Pal. Structural inductive biases in emergent communication. *arXiv preprint arXiv:2002.01335*, 2020.

S4: On edge-device Voice Type Classification

The Voice Type Classifier [1,3] is a multi-label speech classification model based on BabyHuBERT [1,2], an in-house speech representation model trained on a large corpus of child-centered naturalistic audio data [4,5].

Researchers studying early childhood language acquisition rely on such models to extract key information from large corpora of sensitive data. However, most do not have access to computing clusters with high-end GPUs, limiting their ability to process files efficiently.

The goal of this internship is to develop a suite of models that can run efficiently on consumer hardware (laptops and small computers).

The internship will be divided into two phases:

- **Phase 1 – Inference Optimization:** You will survey the model compression literature and improve the inference efficiency of an already trained model. This will involve techniques such as quantization, pruning, and evaluating different inference engines.
- **Phase 2 – Compact Model Training**: This more exploratory phase will focus on training a new suite of smaller models. You will investigate knowledge-distillation [6,7] techniques to transfer knowledge from the original model to more compact architectures, as well as parameter-efficient fine-tuning [8] methods to reduce computational requirements while preserving performance.

This internship offers the opportunity to explore multiple facets of model compression and contribute to research with real-world impact for the child development research community. This internship is suitable for a Master 1 or 2 student.

Junior supervision: *Tarek Kunze, Théo Charlot*

Senior supervision: *Emmanuel Dupoux*, Alejandrina Cristia, Marvin Lavechin

References

[1] Charlot, T., Kunze, T., Poli, M., Cristia, A., Dupoux, E., & Lavechin, M. (2025). BabyHuBERT: Multilingual Self-Supervised Learning for Segmenting Speakers in Child-Centered Long-Form Recordings. arXiv [Eess.AS]. Retrieved from http://arxiv.org/abs/2509.15001

[2] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[3] Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., Cristia, A. (2020) An Open-Source Voice Type Classifier for Child-Centered Daylong Recordings. Proc. Interspeech 2020, 3072-3076, doi: 10.21437/Interspeech.2020-1690

[4]Li, J., Lavechin, M., Fan, X., McElwain, N. L., Cristia, A., Garcia-Perera, P., & Hasegawa-Johnson, M. (2025). Automated Analysis of Naturalistic Recordings in Early Childhood: Applications, Challenges, and Opportunities. arXiv [Eess.AS]. Retrieved from http://arxiv.org/abs/2509.18235

[5] Lucas Gautheron, Marvin Lavechin, Rachid Riad, Camila Scaff, Alejandrina Cristia. Longform recordings : Opportunities and challenges. *LIFT 2020 - 2èmes journées scientifiques du Groupement de Recherche "Linguistique informatique, formelle et de terrain"*, Dec 2020, Montrouge / Virtual, France. pp.64-71. ⟨hal-03047153⟩

[6] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv [Stat.ML]. Retrieved from http://arxiv.org/abs/1503.02531

[7] Jang, K., Kim, S., Yun, S.-Y., Kim, H. (2023) Recycle-and-Distill: Universal Compression Strategy for Transformer-based Speech SSL Models with Attention Map Reusing and Masking Distillation. Proc. Interspeech 2023, 316-320, doi: 10.21437/Interspeech.2023-1329

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., … Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2106.09685