Master's Internships 2024/2025

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) you are interested in.***

S1: Benchmarking visual language models on child language inputs

Recent advances in visual language models (VLMs) in tasks like object recognition have demonstrated the potential to simulate how caregivers respond to child language learners in visually grounded settings. However, most multimodal benchmarks typically assume relevant and well-formed linguistic inputs [1], often neglecting real-world language variations such as child production. This limitation presents challenges for testing the robustness and semantic **flexibility** of VLMs for linguistically underspecified inputs [2].

This project aims to benchmark VLMs' ability to handle child-like linguistic inputs that include ungrammatical or underspecified references (e.g., *"Doggie run!"* describes the scene that a dog is chasing a ball). The key tasks include: (i) extracting linguistic features of child production from transcripts of child-adult interactions(e.g.CHILDES [3]) using advanced NLP tools (e.g. grammaticality [4]) (ii) designing a multimodal dataset that pairs visual scenes with child-like linguistic descriptions based on an existing image-captioning dataset (e.g. [5]) ; (iii) evaluating the performance of state-of-the-art VLMs (e.g., LLaVA [6]; CogVLM [7]; GPT-4 [8]) using a variety of multi-modal tasks such as visual question answering, image captioning, and referential ambiguity resolution.

Prerequisites: *This topic requires good Python programming skills as well as experience in NLP. Background in linguistics and cognitive science is highly appreciated.*

Junior supervision: *Jing Liu*

Senior supervision: *Abdellah Fourtassi*

Note*:* If you are interested in this project, please directly send your CV and cover letter to: *abdellah.fourtassi@gmail.com*.

References

[1] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

[2] Chung, J., Lim, S., Jeon, J., Lee, S., & Yu, Y. (2024). Can visual language models resolve textual ambiguity with visual cues? Let visual puns tell you!. *arXiv preprint arXiv:2410.01023*.

[3] MacWhinney, B. (2000). *The CHILDES project: The database*(Vol. 2). Psychology Press..

[4] Nikolaus, M., Agrawal, A., Kaklamanis, P., Warstadt, A., & Fourtassi, A. (2024). Automatic Annotation of Grammaticality in Child-Caregiver Conversations. *arXiv preprint arXiv:2403.14208*.

[5] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.

[6] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, *36*.

[7] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., ... & Tang, J. (2023). Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

[8] OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

S2: Using Speech-Based AI to Study Communicative Development

Large Language Models, such as ChatGPT, have shown impressive abilities in text-based tasks. Beyond practical applications, they have also sparked scientific discussions about the nature of human language and cognitive development, including debates around Chomsky's theories on the emergence of syntax.[1]

However, these models have limitations in advancing our understanding of how children acquire language. First, they rely on vast amounts of text data for training. Children do not acquire language through exposure to written text; their language learning is grounded in speech—an inherently multimodal signal that combines linguistic and paralinguistic information such as prosody. These features are understood to play a critical role in shaping children's communicative development.[2] Second, children are not passive learners, they actively engage in (proto-)conversational exchanges with caregivers. Through interactions, they influence their linguistic environment, creating a dynamic feedback loop that is vital for learning.[3]

Recent advances in speech language modeling provide a scientific infrastructure for the study of how multimodality and interaction shape early language development. Models like Moshi [4] represent a significant step forward by processing speech directly, without first converting it into text. This approach allows an effective integration of both linguistic and paralinguistic cues. Moshi also models *interactive* speech communication, enabling it to listen and respond simultaneously—just as humans do.

This project aims to use such speech-based models to study children's communicative development in unprecedented ways, addressing key questions about how early conversational dynamics, prosody, and meaning interact to support language acquisition and use. Beyond its scientific contributions, this work has significant societal implications. In education, it can guide the development of more engaging, low-latency e-tutoring systems. In health, it can improve the accuracy of tools for early detection of communicative disorders, such as autism, through analysis of markers like turn-taking dynamics and prosody.

---

[1] Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, 353-414.
[2] Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and speech*, *51*(1-2), 61-75.
[3] Murray, L., & Trevarthen, C. (1986). The infant's role in mother–infant communications. *Journal of child language*, *13*(1), 15-29.
[4] Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., ... & Zeghidour, N. (2024). Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

The internship will focus on the Generative Spoken Language Model (dGSLM),[5] a direct precursor to Moshi. dGSLM is well-suited for an M2 internship due to its relative simplicity, while still being capable of producing significant scientific results. The main components of dGSLM include (see Figure, extracted from the original paper):

- **Encoder:** HuBERT, a self-supervised speech model that encodes linguistic and paralinguistic features from raw audio
- **Decoder:** HiFi-GAN, a vocoder for generating realistic audio.
- **Model Architecture:** Duplex transformer, which supports bidirectional processing of conversational dynamics.

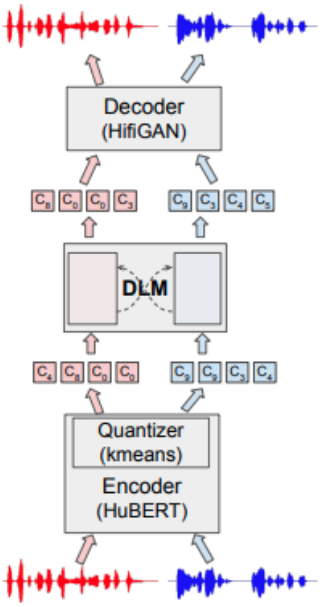

Figure 1: General Schema for dGSLM: A discrete encoder (HuBERT+kmeans) turns each channel of a dialogue into a string of discrete units $(c_1, ..c_N)$. A Dialogue Language Model (DLM) is trained to autoregressively produce units that are turned into waveforms using a decoder (HifiGAN).

We will fine-tune dGSLM on around 150 hours of child-adult conversations from the OCSC corpus, which includes data from 303 children aged 4 to 9 years. This fine-tuning will adapt the model to study child-directed communication. In particular, we will explore how prosody influences turn-taking dynamics, employing methods analogous to those we use to study children's behavior in the lab.[6]

Junior supervision: *Jing Liu*
Senior supervision: *Abdellah Fourtassi*

Note*:* If you are interested in this project, please directly send your CV and cover letter to: *abdellah.fourtassi@gmail.com*.

[5] Nguyen, T. A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W. N., Elkahky, A., ... & Dupoux, E. (2023). Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, *11*, 250-266.
[6] Ekstedt, E., & Skantze, G. (2022). How much does prosody help turn-taking? investigations using voice activity projection models. *arXiv preprint arXiv:2209.05161*.

S3: Diffusion-based forward model for articulatory trajectories

Self-supervised learning (SSL) representations of speech have shown remarkable performance in various downstream tasks, such as ASR, speaker diarisation or speech translation. Along this line of work, a study showed that the latent SSL representations in fact correlate very strongly with articulatory kinematic trajectories [1]. Recently, it was also shown that the articulatory-informed latent representations of a single speaker enable generalisation to unseen languages and speakers in speech synthesis [2]. *This offers the possibility of building a text-to-speech (TTS) model with an articulatory-control bottleneck.*

On the other hand, deep generative models, such as diffusion probabilistic models (DPMs), have lately gained traction in many machine learning tasks, across the text, image and speech modalities. In particular, DPMs have been used to produce high-quality audio by reconstructing the acoustic features [3] or the raw waveform [4]. *Given their generative nature, DPMs have the potential to produce controllable trajectories in articulatory space based on discrete representations.*

The goal of this internship project is to build a diffusion-based forward model to generate controllable, latent articulatory kinematic trajectories (*i.e.* the dynamics of the latent SSL representations) from discrete phonological configurations (*i.e.* the IPA-like, ternary-valued phonological features). The tasks will consist of (i) the preparation of a (forced-aligned) multilingual dataset, (ii) the implementation of the forward model and (iii) the synthesis-based evaluation.

This topic requires strong Python programming skills as well as a deep background in DL. Hence, ***end-of-study Master students*** *(M2 or equivalent) will be considered* ***only****. A background in phonology is appreciated but not required.*

Junior supervision: *Angelo Ortiz Tandazo*

Senior supervision: *Emmanuel Dupoux, Thomas Hueber*

References

[1] Cho, C. J., Wu, P., Mohamed, A., & Anumanchipalli, G. K. (2023). Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE. LINK

[2] Cho, C. J., Wu, P., Prabhune, T. S., Agarwal, D., & Anumanchipalli, G. K. (2024). Coding Speech through Vocal Tract Kinematics. *arXiv preprint arXiv:2406.12998*. LINK

[3] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., & Kudinov, M. (2021). Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In International Conference on Machine Learning (pp. 8599-8608). PMLR. LINK

[4] Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2021). DiffWave: A Versatile Diffusion Model for Audio Synthesis. In International Conference on Learning Representations. LINK

S4: Hierarchical under-specification of phonological features

The types of phonological representations are diverse and depend on the units of study of their underlying phonological theories. For instance, the Generative Phonology (GP) theory initiated by Chomksy and Halle [1] proposes a feature set based on the *acoustic* properties of and synchronised at speech segments, but Articulatory Phonology uses *articulatory* features at the segmental level and Autosegmental Phonology involves suprasegmental *tiers*.

A feature set for the GP theory describes segments according to manner, place and laryngeal binary features [2]. It notably includes the so-called “zero-valued” features: these are features that are irrelevant or that depend on the context of the segment. A recent study used these features to train a linear model to predict articulatory trajectories on top of simple phonological forward models (linear and cubic interpolations) [3]. One of the conclusions was that keeping under-specified features (*i.e.* zero-valued features), instead of arbitrarily binarising them – thus, deferring their value interpolation to the forward models – was beneficial in terms of the correlation coefficient.

In this internship project, we seek to establish the impact of the different levels of phonological features under-specification on the prediction of articulatory trajectories (for English). The under-specification approaches to test include, but are not limited to, (i) random features by segment class, (ii) top-down under-specification (starting from fully specified features) and (iii) bottom-up specification (starting from a zero matrix).

This topic is more suitable for a Master's student with a background in (computational) linguistics and/or phonology.

Junior supervision: *Angelo Ortiz Tandazo*

Senior supervision: *Emmanuel Dupoux, Thomas Hueber*

References

[1] Chomsky, N. & Halle, M. (1968). The sound pattern of English. Harper & Row.
[2] Hayes, B. (2011). Introductory phonology. John Wiley & Sons.
[3] Ortiz Tandazo, A., Schatz, T., Hueber, T., Dupoux, E. (2024). Simulating articulatory trajectories with phonological feature interpolation. Proc. Interspeech 2024, 3595-3599, doi: 10.21437/Interspeech.2024-2192. LINK

S5: Fine-tuning SSL models for speaker segmentation

Audio segmentation based on speaker identification presents unique challenges when working with child speech data. The distinctive characteristics of children's voices, including higher pitch, greater variability, and less stable speech patterns, make traditional speaker segmentation systems less effective. Additionally, most existing models are primarily trained on adult speech, creating a significant domain gap when applied to child speech data.

Self-Supervised Learning[^1][^2] models have emerged as the state-of-the-art approach for speech processing tasks. Models such as wav2vec, HuBERT, and WavLM have demonstrated exceptional performance by leveraging massive amounts of unlabeled speech data. These models learn robust speech representations that can be effectively transferred to downstream tasks.

The intern project will begin with a migration of our current speech segmentation framework to the SpeechBrain ecosystem[^3], allowing us to leverage the existing training and fine-tuning pipelines. After that, the intern will focus on
adapting state-of-the-art speech foundation models using our proprietary dataset comprised of 20k hours of unannotated audio. Finally, these fine-tuned models will be used as sophisticated feature extractors for the downstream speaker classification task. If the time allows it, multiple speech foundation models will be compared.

Junior supervision: Tarek Kunze
Senior supervision: Marvin Lavechin, Alejandrina Cristia

Prerequisites

- good coding skills, especially in python
- knowledge of pytorch is appreciated
- knowledge of transformer-like encoder-decoder architectures

References

[^1]: Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., … Goldblum, M. (2023). A Cookbook of Self-Supervised Learning. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2304.12210
[^2]: S. -w. Yang et al., "A Large-Scale Evaluation of Speech Foundation Models," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2884-2899, 2024, doi: 10.1109/TASLP.2024.3389631.
[^3]: Ravanelli, M. et al. (2024) Open-Source Conversational AI with SpeechBrain 1.0.

S6: Extending Whisper to perform speaker segmentation

Speaker segmentation based on speaker identification presents unique challenges when working with child speech data. The distinctive characteristics of children's voices, including higher pitch, greater variability, and less stable speech patterns, make traditional speaker segmentation systems less effective. Additionally, most existing models are primarily trained on adult speech, creating a significant domain gap when applied to child speech data.

Whisper[^1] has emerged as a powerful and versatile Speech Foundation Model[^2], demonstrating remarkable capabilities across transcription and translation tasks. Its encoder-decoder architecture, trained on a vast and diverse corpus of audio data, provides robust speech representations that can be adapted for various downstream applications. It has been shown to work well for speaker diarization on Child-Adult dyadic interactions data[^3]

The internship will focus on advancing speaker segmentation capabilities by extending the Whisper architecture. Inspired by nanodrz[^4] The project will be structured around designing and implementing a novel decoder architecture that extends Whisper's capabilities to perform precise speaker classification.
The intern will first develop a custom decoder that outputs three key elements: temporal boundary tokens (start/end positions) and the speaker class tokens.
The implementation will be followed by a systematic fine-tuning phase using our internal dataset. Finally, comprehensive evaluation protocols will be established to assess the model's performance across various metrics, including temporal precision, speaker class identification accuracy and more.

Junior supervision: Tarek Kunze
Senior supervision: Marvin Lavechin, Alejandrina Cristia

Prerequisites

- good coding skills, especially in python
- knowledge of Pytorch is appreciated
- knowledge of transformer-like encoder-decoder architectures

References

[^1]: Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv [Eess.AS]. Retrieved from http://arxiv.org/abs/2212.04356
[^2]: S. -w. Yang et al., "A Large-Scale Evaluation of Speech Foundation Models," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2884-2899, 2024, doi: 10.1109/TASLP.2024.3389631.
[^3]: Xu, A., Huang, K., Feng, T., Shen, L., Tager-Flusberg, H., Narayanan, S. (2024) Exploring Speech Foundation Models for Speaker Diarization in Child-Adult Dyadic Interactions. Proc. Interspeech 2024, 5193-5197, doi: 10.21437/Interspeech.2024-717
[^4]: Coultas Blum, H. (01 2024). nanodrz (Version 1.0.0). Retrieved from https://github.com/mogwai/nanodrz

S7: Pretraining on formal languages for sample efficiency

Children are incredibly data-efficient language learners, and language models are not. Children are exposed to less than 100 million words by age 13 [1], while modern language models typically need 3 or 4 orders of magnitude more data. But infants do not start from scratch: they have inductive biases to learn language inherited from evolution. This project aims to induce strong inductive biases to learn the structure of language by pretraining text-based language models on formal languages before transferring them to natural language.

This project situates itself within the literature of training Transformer-based language models on formal languages [2,3]. It aims to explore two main questions: how well these models transfer to natural language tasks [4,5,6,7] and what kinds of knowledge they can effectively acquire [8,9,10]. Following the approach in [11], the formal languages will be constructed by defining a set of formal primitives and developing a probabilistic model that combines these primitives to create definitions of languages.

The evaluation will be focused on the BabyLM challenge [12]: first pretraining on formal languages and then transfer on a small quantity of natural language, and evaluating the model by its zero-shot lexical and syntactic knowledge and its performance on downstream tasks. This can be followed by going from a pretrain-then-transfer paradigm to a meta-learning one as in [13,14] or by using approaches from the model merging literature [15,16].

Junior supervision: *Maxime Poli*

Senior supervision: *Emmanuel Chemla, Emmanuel Dupoux*

References

[1] Gilkerson and al. Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis American Journal of Speech-Language Pathology, 2017.
[2] Z. Allen-Zhu, Y. Li. Physics of Language Models: Part 1, Learning Hierarchical Language Structures
[3] Y. Wu, F. Li, and P. Liang. "Insights into Pre-training via Simpler Synthetic Tasks". NeurIPS 2022
[4] C.-H. Chiang and H. Lee. On the Transferability of Pre-trained Language Models: A Study from Artificial Datasets. AAAI 2022
[5] I. Papadimitriou and D. Jurafsky. "Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models". EMNLP 2020
[6] I. Papadimitriou and D. Jurafsky. Injecting structural hints: Using language models to study inductive biases in language learning Findings of EMNLP 2023
[7] R. Ri and Y. Tsuruoka. "Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models". In: ACL 2022
[8] F. Cagnetta, M. Wyart Towards a theory of how the structure of language is acquired by deep neural networks NeurIPS 2024
[9] J. Kallini et al. Mission: Impossible Language Models ACL 2024
[10] J. White and R. Cotterell. "Examining the Inductive Bias of Neural Language Models with Artificial Languages". ACL IJCNLP 2021
[11] Y. Yang, S.T. Piantadosi, One model for the learning of language, PNAS. 2022
[12] Alex Warstadt et al. "Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora". In: BabyLM at CoNLL 2023
[13] R. T. McCoy and T. L. Griffiths. Modeling Rapid Language Learning by Distilling Bayesian Priors into Artificial Neural Networks.
[14] R. T. McCoy et al. Universal Linguistic Inductive Biases via Meta-Learning CogSci 2020
[15] Alexandre Rame et al. "Model Ratatouille: Recycling Diverse Models for Out-of-Distribution Generalization". ICML 2023
[16] Mitchell Wortsman et al. "Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy without Increasing Inference Time". ICML 2022

S8: 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 Chemla, 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