M2 Internship: Using Speech-Based AI to Study Communicative Development

Requirement: M1 in computer science

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 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.

We will fine-tune dGSLM on around 150 hours of child-adult conversations from a new 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]

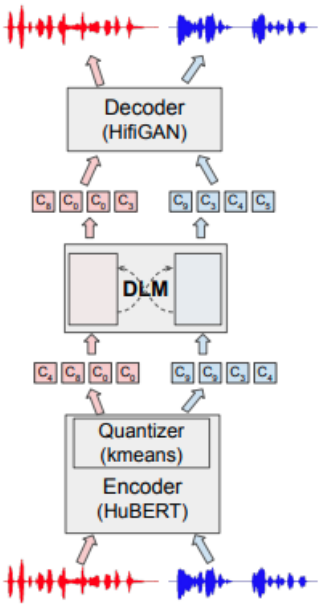

Figure 1: General Schema for dGSLM: A discrete encoder (HuBERT+kmeans) turns each channel of a dialogue into a string of discrete units $(c_1, ..c_N)$. A Dialogue Language Model (DLM) is trained to autoregressively produce units that are turned into waveforms using a decoder (HifiGAN).

Practicalities

The internship will be funded ~600 euros per month for a duration of 5 to 6 months. It will take place in Marseille within the TALEP research group at LIS/CNRS on the Luminy campus. The intern will collaborate with other interns from this project, as well as PhD students and researchers from the research group.

How to apply: send as soon as possible a short application letter, transcripts, and CV to abdellah.fourtassi@gmail.com

- Expected start: February or March 2025

[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*.