Open PhD Position – GIPSA-lab, Grenoble (France)

Title: *Grounding a Multimodal Speech Language Model Through Physical and Social Interaction*
Supervision: Dr. Thomas Hueber, GIPSA-lab (CNRS / Grenoble Alpes Univ.)
Co-supervision : Dr. Stéphane Lathuilière (INRIA-UGA), Pr. Laurent Girin (GIPSA-lab)
Funding: Fully funded by the Grenoble AI Research Institute MIAI Cluster (3 years)

Context

This PhD position is part of the **DevAI&Speech** project, funded by the **Grenoble AI Research Institute MIAI Cluster**. The project aims to advance conversational AI by drawing inspiration from **infant language acquisition**. It explores **textless Speech Language Models (SpeechLMs)** — generative speech models that learn without textual supervision, mimicking how children acquire language before literacy. DevAI&Speech focuses on how **sensorimotor experience, embodiment, and social interaction** can shape language learning. By combining AI and developmental science, it seeks not only to build better AI systems but also to deepen our understanding of human language acquisition.
The specific PhD project will investigate how **grounding SpeechLMs in multimodal interaction** — both physical and social — can improve their language learning capabilities.

Scientific Objectives

The PhD project will pursue the following goals:

- Investigate how multimodal inputs (e.g., visual, prosodic, contextual cues) support **audio stream segmentation** and **lexicon acquisition**. While LLMs perform well with textual input, segmentation and grounding remain unsolved for raw speech.
- Study the role of **multimodal communicative interaction** in the **grounded learning of word meanings**, with an emphasis on the lexical level.

Methodology & Roadmap

1. **Multimodal SpeechLM design & lexical metrics:** Extend a textless SpeechLM with multimodal inputs (e.g., audio & vision). Use metrics like *spot-the-word* (sWuggy) to evaluate lexical learning.
2. **Child–Parent Simulation Framework:** Simulate interactions between a “child” SpeechLM (textless) and a “parent” model (text-based, pretrained). After a passive self-supervised pretraining phase (analogous to early statistical learning), the child model is prompted with an image to describe. The parent model provides corrections. This setup will allow systematic

studies of the **impact of communicative cues**, especially **prosody**(e.g., contrastive focus), on learning outcomes.

3. **Interactive Learning with a Robot:** Embed a multimodal SpeechLM into the **humanoid robots** available at GIPSA-lab (iCub or Furhat). Design experiments where human participants interact with the robot, asking it to name objects (e.g., "Look, this is a blue …") and correct its responses through **naturalistic feedback** (e.g., "That's right!" / "No, it's a cow."). Inspired by recent work in **human-in-the-loop reinforcement learning** (e.g., Cohen & Billard, 2018; Deichler et al., 2023), we will explore how **multimodal feedback** can define a reward function for **online learning**. Over time, task complexity will increase (e.g., full sentence generation: "The white cat is watching the red ball").

   Publication and dissemination: The project aims to publish its results in top-tier venues such as ACL, Interspeech, ICASSP, CVPR, ICLR, NeurIPS and to release the code as open source

---

Expected Profile

We are seeking a motivated PhD candidate with a strong background in one or more of the following areas:

- Speech processing, NLP, computer vision, machine learning
- Solid programming skills (including the PyTorch library)
- Interest in connecting AI with human cognition

Prior experience with LLM, SpeechLMs, RL algorithms, or robotic platforms is a plus, but not mandatory.

Salary & Funding

- Fully funded 3-year position (standard French doctoral salary)
- Additional funding for conference travel and equipment

Application & Contact

Start Date: December 1st 2025 (flexible)
Application Deadline: Open until filled
To apply, please send the following documents to:
📧 **[Thomas Hueber]** – thomas.hueber@*grenoble-inp.fr*
📧 **[Laurent Girin]** – laurent.girin@*grenoble-inp.fr*
📧 **[Stéphane Lathuilière]** – stephane.lathuiliere@inria.fr

- CV
- Motivation letter

- Academic transcripts
- Reference contacts or letters (if available)

Related studies

- Cohen L, Billard A. (2018) Social babbling: The emergence of symbolic gestures and words. Neural Networks, vol. 106, pp.194-204.
- Cui W., Yu D., Jiao X., Meng Z., et al.. (2024). Recent Advances in Speech Language Models: A Survey, https://arxiv.org/pdf/2410.03751.
- Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. Cognition, 173, 43-59.
- Georges, M-A, Lavechin, M., Schwartz, J-L., Hueber., T., Decode, Move and Speak! Self-supervised Learning of Speech Units, Gestures, and Sound Relationships Using Vocal Imitation. Computational Linguistics 2024;
- Kuhl. P. (2004) Early language acquisition: cracking the speech code, Nature Reviews Neuroscience, vol. 5, pp. 831–843
- Nikolaus, M., & Fourtassi, A. (2021). Modeling the Interaction Between Perception-Based and Production-Based Learning in Children's Early Acquisition of Semantic Knowledge. In A. Bisazza & O. Abend (Eds.), Proc. of the 25th Conference on Computational Natural Language Learning, pp. 391–407
- Peng, P., Harwath, D. (2022) Word Discovery in Visually Grounded, Self-Supervised Speech Models. Proc. of Interspeech, pp. 2823-2827
- Tellex, S.; Gopalan, N.; Kress-Gazit, H.; and Matuszek, C. 2020. Robots That Use Language. Annual Review of Control, Robotics, and Autonomous Systems 3(1): 25–55.