Multimodal and Social Grounding of Speech Language Models for Studying Human Language Acquisition

Open PhD Position – GIPSA-lab & INRIA Grenoble (France)
Start Date: October 2026

Scientific Context and Motivation

Recent progress in self-supervised and generative speech modeling has led to the emergence of textless Speech Language Models (SpeechLMs), capable of learning directly from raw speech without textual supervision [1]. These models constitute a promising computational framework for studying language acquisition in a manner closer to how human infants learn language before literacy [2], while also offering an opportunity to draw inspiration from human learning mechanisms to design more adaptive and grounded conversational AI systems. Despite recent progress, current textless SpeechLMs still struggle to acquire higher-level linguistic competencies such as robust lexical representations, syntax, or semantics when trained on amounts of speech data comparable to those available to human infants during development, or when learning from ecologically realistic data [3, 4]. This discrepancy suggests that scale alone may not be sufficient for language acquisition. One possible explanation is that current models largely lack the multimodal and social grounding mechanisms that characterize early human learning [5]. In natural development, language emerges through rich perceptual and communicative interactions combining speech, vision, action, attention, and adaptive social feedback. Understanding the role of such grounding mechanisms therefore constitutes an important challenge for both developmental science and the next generation of conversational AI systems.

Recent computational models of visually grounded speech learning have progressively explored how lexical structure may emerge from the joint statistics of speech and vision. Early work on cross-situational word learning showed that learners can associate spoken forms with visual referents by accumulating evidence across individually ambiguous situations [6]. Subsequent computational models extended this idea to continuous speech, showing that visual grounding can also support speech segmentation and the discovery of word-like units [7, 9]. More recent visually grounded speech models based on self-supervised and contrastive learning further demonstrated that aligning raw speech with concurrent visual scenes can lead to the emergence of phonological and lexical representations without textual supervision [8, 10, 11].

Beyond perceptual grounding, a growing line of computational research emphasizes the role of social interaction in language acquisition. In natural development, children are not merely exposed to multimodal sensory input, but actively engage in communicative exchanges where they produce linguistic behaviors, receive feedback, and progressively adapt their internal representations through interaction. Recent computational models have therefore started to investigate how learning may emerge from the interplay between perception, production, and social feedback. In particular, Nikolaus and Fourtassi [12] proposed a neural model integrating both perception-based and production-based learning, showing that active language production and interaction-driven feedback improve semantic acquisition beyond passive perceptual learning alone. Their results highlight the importance of modeling language development not only as multimodal statistical learning from sensory input, but also as an interactive and socially guided

process in which learners actively participate in the construction of their linguistic knowledge.

Nevertheless, current computational models of multimodal and social grounding still suffer from several important limitations. Most visually grounded models are trained on highly simplified image-captioning datasets that only weakly reflect the richness and ambiguity of real infant experience. Conversely, models attempting to incorporate social interaction often rely on text-based representations as an intermediate backbone, thereby ignoring many crucial communicative signals conveyed directly through speech itself, including the prosodic and interactive cues characteristic of child-directed speech (CDS), such as prosodic emphasis, repetition, corrective feedback, and interactive clarification strategies. More fundamentally, existing approaches rarely model language acquisition as a dynamic co-adaptation process between the child and the caregiver. Modeling how such multimodal and social interactions shape language learning therefore remains a major challenge, and constitutes a central motivation of the present PhD project.

Methodology and Research Roadmap

Multimodal Grounding for Lexical Acquisition

The first objective will consist in developing a multimodal textless SpeechLM trained on realistic audiovisual environments going beyond standard caption-based datasets. In particular, the project will investigate controlled multimodal environments inspired by early infant experience, including realistic naming events, varying visual ambiguity, and different communicative speaking styles such as child-directed speech. The goal will be to study how multimodal perceptual information contributes to speech segmentation, lexical discovery, referential grounding, and the emergence of robust speech representations. Special attention will be paid to the interaction between acoustic variability, prosodic structure, and visual context during lexical acquisition.

Interactive and Socially-Grounded Language Learning

The second objective will investigate the role of communicative interaction in language acquisition through an interactive "child–caregiver" learning framework. The PhD researcher will develop an agentic interaction setup in which a textless "child" SpeechLM progressively acquires lexical knowledge, while a pretrained "caregiver" model dynamically adapts its communicative behavior according to the learner's current state. Inspired by developmental studies of caregiver–infant interaction, the caregiver model will be able to emphasize target words, provide corrective or reinforcing feedback, adapt lexical complexity and use contrastive naming strategies, and simulate multimodal communicative cues such as gaze or pointing to guide attention. This framework will enable controlled studies of how interactive feedback and communicative adaptation influence lexical learning.

Toward Real-World Human–Robot Interaction (Exploratory Direction)

As a more exploratory perspective, the developed multimodal SpeechLMs will be embedded into the humanoid robotic platforms available at GIPSA-lab and INRIA in order to study grounded communication in situated interaction settings. This direction is notably inspired by work on multimodal communicative behaviors in embodied conversational agents [13] as well as developmental robotic approaches investigating the emergence of language and gestures through social interaction and sensorimotor exploration [14]. Such an extension could enable preliminary human–robot interaction experiments in which participants naturally interact with the agent and provide multimodal communicative feedback during situated language learning tasks.

Research Environment and Project Context

This PhD position is **fully funded for 3 years** (standard French doctoral salary) by the DevAI&Speech Chair of the Grenoble AI Research Institute (MIAI Cluster), a research program that aims to advance conversational AI by drawing inspiration from language acquisition in children. The recruited PhD researcher will be jointly hosted at GIPSA-lab (CNRS / Université Grenoble Alpes), within the ComLearn research team (formerly the CRISSP team), and at the Inria Centre at the University Grenoble Alpes, within the RobotLearn team. The PhD candidate will be supervised by Thomas Hueber, Stéphane Lathuilière, and Laurent Girin. The PhD researcher will also interact with Mathilde Fort, a researcher in developmental psychology and a specialist in speech and language acquisition at the Grenoble BabyLab/LPNC. The position further includes opportunities for research stays at Tampere University (Finland) in collaboration with Okko Räsänen. The student will benefit from access to state-of-the-art computational infrastructures, including GPU servers available at GIPSA-lab and GRICAD, and the large-scale national computing facility Jean-Zay.

Responsibilities and Expected Activities of the PhD Candidate

The PhD candidate will conduct bibliographic research, design and implement computational models as well as experimental and simulation frameworks, run large-scale experiments, and analyze the obtained results. The PhD researcher will also actively contribute to the dissemination of the research through the writing of scientific articles and participation in international conferences and workshops. The project lies at the intersection of speech and natural language processing, computer vision, deep learning, robotics, and cognitive science. Target venues include top conferences and journals in these fields (e.g. NeurIPS, ICLR, ACL, EMNLP, Interspeech) as well as interdisciplinary venues related to cognitive and developmental modeling. The candidate will participate in regular PhD progress meetings, team meetings, and the broader scientific life of the laboratory. In addition, the doctoral researcher will be required to follow the mandatory courses and training program of the affiliated doctoral school.

Profile and Skills Required

Applicants should hold a Master's degree (or equivalent) in one or several of the following fields: natural language and speech processing, computer vision, computational linguistics, computer science, data science, machine learning, or related areas. Good programming skills in Python and experience with deep learning frameworks such as PyTorch are expected. The candidate should also demonstrate a strong interest in interdisciplinary research at the intersection of artificial intelligence, speech technologies, and cognitive science (an interest in bridging AI and human cognition is highly desirable). Strong communication and organizational skills are important, as the PhD student will be expected to work collaboratively within an interdisciplinary research environment and actively participate in scientific dissemination activities. A good level of spoken and written English is required, including the ability to present research results clearly at conferences and to write scientific publications.

Application & Contact

To apply, please send CV, academic transcripts and reference contacts (or letters) to:

`thomas.hueber@grenoble-inp.fr`

References

[1] Arora, S., Chang, K. W., Chien, C. M., Peng, Y., Wu, H., Adi, Y., & Watanabe, S. (2025). On the Landscape of Spoken Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2504.08528*.

[2] Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language learner. *Cognition*, 173, 43–59.

[3] Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., & Dupoux, E. (2021). On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9, 1336–1354.

[4] Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., & Cristia, A. (2024). Modeling early phonetic acquisition from child-centered audio data. *Cognition*, 245, 105734.

[5] Dupoux, E., LeCun, Y., & Malik, J. (2026). Why AI systems don't learn and what to do about it: Lessons on autonomous learning from cognitive science. *arXiv preprint arXiv:2603.15381*.

[6] Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.

[7] Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, 122(4), 792.

[8] Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., & Glass, J. (2018). Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of ECCV*, 649–665.

[9] Havard, W. N., Chevrot, J.-P., & Besacier, L. (2019). Word recognition, competition, and activation in a model of visually grounded speech. In *Proceedings of CoNLL*, 339–348.

[10] Chrupała, G. (2022). Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, 73, 673–707.

[11] Khorrami, K., & Räsänen, O. (2025). A model of early word acquisition based on realistic-scale audiovisual naming events. *Speech Communication*, 167.

[12] Nikolaus, M., & Fourtassi, A. (2021). Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, 391–407.

[13] Deichler, A., Wang, S., Alexanderson, S., & Beskow, J. (2023). Learning to generate pointing gestures in situated embodied conversational agents. *Frontiers in Robotics and AI*, 10:1110534.

[14] Cohen, L., & Billard, A. (2018). Social babbling: The emergence of symbolic gestures and words. *Neural Networks*, 106, 194–204.