Multimodal approaches to predict turn-taking in natural conversations

Supervisors: Leonor Becerra, Philippe Blache and Eliot Maës
Aix-Marseille Université

Description of the project

Interaction between humans is *multimodal* by nature. When people interact with each other in a conversation, information from different modalities comes into play: verbal (use of language), visual (gestures, gaze, etc.), physiological (heart rate, skin temperature, etc.) and even cerebral (neural correlates between the different signals). Therefore, in order to have a complete vision of human interaction, it is crucial not to focus only on one of these modalities.

Turn-taking occurs in a conversation when one person listens while the other person speaks. This is a fundamental aspect of dialogue, since the participants need to coordinate who is speaking and when the next person can start to speak [5]. The cues used to coordinate their turn-taking have been widely studied, and have been found across different modalities, including verbal and non-verbal (such as gestures, breathing, etc.) cues. Although humans are good at this coordination, this is not the case for conversational systems [6]. Hence, the development of predictive models of turn-taking can help us to better understand social interactions, and can be applied to chatbots, for example, in order to identify the right moment to start speaking again, rather than relying on pause duration heuristics that affect the natural aspect of the conversation.

We propose in this project to explore multimodal approaches [1,2,3] to predict turn-taking in natural conversations. This study will be focused on a multimodal corpus named BrainKT [4] containing audio, video and neuro-physiological recordings of 28 interactions between dyads of participants. Conversations are in French and lasting 30 minutes, with different tasks (game and free conversation). This corpus is enriched with transcriptions automatically aligned to the audio signal. The goal of the project is to explore how different modalities are combined and complement each other to predict turn-taking.

Bibliography

[1] T. Baltrušaitis, C. Ahuja and L-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 41, Issue 2, 423-443, 2019.

[2] W. Guo, J. Wang and S. Wang, “Deep Multimodal Representation Learning: A Survey”, in IEEE Access, vol. 7, 63373-63394, 2019.

[3] P.P. Liang, A. Zadeh and L-P. Morency, “Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions”, 10.48550/arXiv.2209.03430. 2022.

[4] Maës, Eliot, et al. "Studying common ground instantiation using audio, video and brain behaviours: the BrainKT corpus." *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. 2023.

[5] M. Pickering and S. Garrod, “Understanding Dialogue”. Cambridge University Press, 2021

[6] G. Skantze, “Turn-taking in Conversational Systems and Human-Robot Interaction: A Review”, in Computer Speech and Language, vol. 67, 1-26, 2021.

Expected profile of the candidate

- Knowledge in Machine Learning and Natural Language Processing
- Programming skills in Python
- Ability to communicate effectively in English, both orally and in writing
- Curious, autonomous, rigorous mind