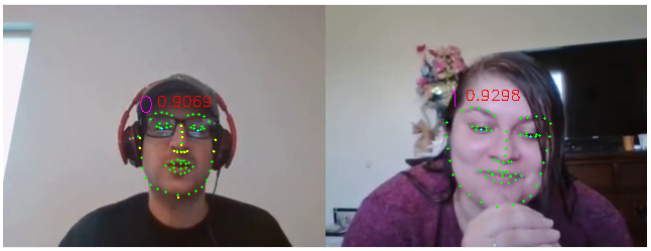
Internship: predictive multimodal model of dyadic conversations

Advisors: Benoit Favre, Magalie Ochs, Alice Delbosc, Stéphane Ayache

October 2023

1 Description

Context Speech is the primary and most natural communication medium. In face-to-face interactions, participants exploit a range of non-verbal signals beyond articulating an acoustic message: facial expressions, gaze, head movements, body pose and gestures. These actions support and augment the spoken message, allowing more engaging and robust communication. Automatic analysis and synthesis of multimodal conversation features has progressed rapidly, leading to increased levels of accuracy for detecting and sometimes synthesising individual phenomena: speech transcription and synthesis, head pose and facial action units recognition and generation, gaze and hand tracking, etc. In addition to allowing novel computer interactions targeting, for example, robotics, low-level and high-level multimodal descriptors support a range of research results in psychology, sociology and linguistics on the dynamics of conversational behavior. Yet, because of the effort required for manually annotating datasets, the varying degree of accuracy associated to those descriptors in uncontrolled settings hampers further research. Recent progress in representation learning could allow to build holistic models of dyadic multimodal conversations without the need for extensive supervision.



Problem statement The goal of this internship is to develop, train and evaluate a predictive model of dyadic conversations. The model will input the face and speech features of each participant at every time frame, and predict those features at future time frames.

Recent work [7] has shown that in the speech modality, realistic turn taking behavior, prosodic features and emotion can be synthesized in dialogic context, thanks to a twin transformer that inputs quantized speech (k-means over HuBERT units [4], or VQ-VAE [9]), generates future units with a next-token predictor, and finally generates back speech from units with a HifiGAN vocoder [5]. Using similar techniques, visual features such as facial action units or head movements, can be synthesized from cross-modal information such as transcripts [6] or speech [2].

We wish to extend these approaches to jointly model the vision and speech modalities. The work will be conducted on the CANDOR corpus [8] which contains about 800h of remote conversations with two participants and separate audio/video channels. Speech and visual features will be extracted with standard tools such as OpenFace [1] and OpenSmile [3]. In addition to the model, the intern will propose appropriate evaluation methodology for devising the quality of the model: automatic and manual evaluation of synthesized samples, probing of representations...

Objective and steps The goal of the internship is to extend the speech-only dialog model by [7] with the visual modality. For the sake of simplicity, the visual modality will be integrated as a vector of speaker-related features such as face landmarks, head orientation, facial action units. The expected contributions of the intern are the following:

1. Replicate results from [7] on the CANDOR corpus
2. Extract target visual features from videos
3. Build discrete representations for the visual modality with VQ-VAE

4. Integrate units from the visual modality to the transformer dialog model
5. Re-synthesize visual features from the discrete units with VQ-VAE
6. Evaluate samples from the model using both human and automatic evaluation

2 Profile

The intern will survey existing work, and leverage it to propose, implement and analyse multimodal models of dyadic conversations. The work will be implemented using Pytorch and relevant libraries. The candidate should have the following qualities:

- Excellent knowledge of deep learning methods (transformers, VQ-VAE...)
- Extensive experience with implementing Pytorch models, and handling research code bases
- Great scientific writing skills
- A hunch for the challenges of doing exciting research

The 6-month internship will take place at LIS/CNRS in Marseille during spring 2024. GPUs from the Jean-Zay super-computer will be available for training larger models.

3 Contact

Please send a CV, transcripts and letter of application to benoit.favre@lis-lab.fr and magalie.ochs@lis-lab.fr. Do not hesitate to contact us if you have any question.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[2] Alice Delbosc, Magalie Ochs, Nicolas Sabouret, Brian Ravenet, and Stéphane Ayache. Towards the generation of synchronized and believable non-verbal facial behaviors of a talking virtual agent. In *International Cconference on Multimodal Interaction*, pages 228–237. 2023.

[3] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[4] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[5] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[6] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10083–10093, 2023.

[7] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023.

[8] Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197, 2023.

[9] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.