

## [Research engineer/post-doc] Robust and generalizable deep learning-based audiovisual speech enhancement

**Context:** The [Multispeech team](#), at Inria Nancy, France, seeks a qualified candidate to work on *signal processing and machine learning techniques for robust audiovisual speech enhancement*. The candidate will be working under the co-supervision of [Mostafa Sadeghi](#) (researcher, [Multispeech team](#)), [Xavier Alameda-Pineda](#) (researcher and team leader of [RobotLearn team](#)), and [Radu Horaud](#) (senior researcher, [RobotLearn team](#)).

**Starting date & duration:** October 2022 (flexible), for a duration of one year (renewable depending on funding availability and performance).

**Background:** Audio-visual speech enhancement (AVSE) refers to the task of improving the intelligibility and quality of a noisy speech signal utilizing the complementary information of visual modality (lip movements of the speaker) [1], which could be very helpful in highly noisy environments. Recently, and due to the great success and progress of deep neural network (DNN) architectures, AVSE has been extensively revisited [1]. Existing DNN-based AVSE methods are categorized into *supervised* and *unsupervised* approaches. In the former category, a DNN is trained on a large audiovisual corpus, e.g., AVSpeech [2], with diverse enough noise instances, to directly map the noisy speech signal and the associated video frames of the speaker into a clean estimate of the target speech signal. The trained models are usually very complex and contain millions of parameters. The unsupervised methods [3] follow a statistical modeling-based approach combined with the expressive power of DNNs, which involves learning the prior distribution of clean speech using deep generative models, e.g., variational autoencoders (VAEs) [4], on clean corpora such as TCD-TIMIT [5], and estimating clean speech signal in a probabilistic way. As there is no training on noise, the models are much lighter than those of supervised methods. Furthermore, the unsupervised methods have potentially better generalization performance and robustness to visual noise thanks to their probabilistic nature [6-8]. Nevertheless, these methods are very recent and significantly less explored compared to the supervised approaches.

**Project description:** In this project, we plan to devise a robust and efficient AVSE framework by thoroughly investigating the coupling between the recently proposed deep learning architectures for speech enhancement, both supervised and unsupervised, benefiting from the best of both worlds, along with the state-of-the-art generative modeling approaches. This will include, e.g., the use of dynamical VAEs [9], temporal convolutional networks (TCNs) [10], and attention-based strategies [11,12]. The main objectives of this project are summarized as follows:

1. Developing a neural architecture that identifies reliable (either frontal or non-frontal) and unreliable (occluded, extreme poses, missing) lip images by providing a normalized score at the output;
2. Developing deep generative models that efficiently exploit the sequential nature of data;
3. Integrating the developed visual reliability analysis network within the deep generative model that accordingly decides whether to utilize the visual data or not. This will provide a flexible and robust audiovisual fusion and enhancement framework.

**Requirements & skills:** The preferred profile is described below.

- M.Sc. or Ph.D. degree in speech/audio processing, computer vision, machine learning, or in a related field,
- Ability to work independently as well as in a team,
- Solid programming skills (Python, PyTorch), and deep learning knowledge,
- Good level of written and spoken English.

**How to apply:** Interested candidates are encouraged to contact Mostafa Sadeghi, Xavier Alameda-Pineda, and Radu Horaud ([first name].[last name]@inria.fr), with the required documents (CV, transcripts, motivation letter).

**References:**

- [1] D. Michelsanti, Z. H. Tan, S. X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, “An overview of deep learning-based audio-visual speech enhancement and separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.
- [2] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.T. Freeman, M. Rubinstein, “Looking-to-Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation,” *SIGGRAPH* 2018.
- [3] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “Audio-visual speech enhancement using conditional variational auto-encoders,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 1788–1800, 2020.
- [4] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *International Conference on Learning Representations (ICLR)*, 2014.
- [5] N. Harte and E. Gillen, “TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech,” *IEEE Transactions on Multimedia*, vol.17, no.5, pp.603-615, May 2015.
- [6] M. Sadeghi and X. Alameda-Pineda, “Switching variational autoencoders for noise-agnostic audio-visual speech enhancement,” in *ICASSP*, 2021.
- [7] Z. Kang, M. Sadeghi, R. Horaud, “Face Frontalization Based on Robustly Fitting a Deformable Shape Model to 3D Landmarks,” in *International Conference on Computer Vision (ICCV) Workshops*, Montreal – Virtual, Canada, Oct. 2021, pp. 1–16.
- [8] Z. Kang, M. Sadeghi, R. Horaud, X. Alameda-Pineda, J. Donley, and A. Kumar, “The impact of removing head movements on audio-visual speech enhancement,” in *ICASSP*, 2022, pp. 1–5.
- [9] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, “Dynamical variational autoencoders: A comprehensive review,” *Foundations and Trends in Machine Learning*, vol. 15, no. 1-2, 2021.
- [10] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. “Temporal convolutional networks: A unified approach to action segmentation.” In *European Conference on Computer Vision*, pp. 47-54. Springer, Cham, 2016.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [12] J. Jiang, G. G Xia, D. B Carlton, C. N Anderson, and R. H Miyakawa, “Transformer VAE: A hierarchical model for structure-aware and interpretable music representation learning,” in *ICASSP*, 2020, pp. 516–520.