Offer #2024-07569

Post-Doctoral Research Visit F/M Deep generative models for robust and generalizable audio-visual speech enhancement

Contract type : Fixed-term contract

Level of qualifications required : PhD or equivalent

Fonction : Post-Doctoral Research Visit

Context

This postdoctoral research is part of the **REAVISE** project: "Robust and Efficient Deep Learning based Audiovisual Speech Enhancement" (2023-2026) funded by the French National Research Agency (ANR). The general objective of REAVISE is to develop a unified audio-visual speech enhancement (AVSE) framework. This will leverage recent breakthroughs in statistical signal processing, machine learning, and deep neural networks to create a robust and efficient AVSE system.

The postdoctoral researcher will be supervised by Mostafa Sadeghi (researcher, Inria), Romain Serizel (associate professor, University of Lorraine), as members of the Multispeech team, and Xavier Alameda-Pineda (Inria Grenoble), member of the RobotLearn team. The team has access to powerful computational resources, including efficient GPUs and CPUs, required for the experiments planned in this project.

Work environment: Multispeech team, Inria Nancy, France.

Starting date & duration: October 2024 (flexible), for two years.

Assignment

Background. Audio-visual speech enhancement (AVSE) aims to improve the intelligibility and quality of noisy speech signals by utilizing complementary visual information, such as the lip movements of the speaker [1]. This technique is especially useful in highly noisy environments. The advent of deep neural network (DNN) architectures has led to significant advancements in AVSE, prompting extensive research into the area [1]. Existing DNN-based AVSE methods are divided into supervised and unsupervised approaches. In supervised approaches, a DNN is trained on a large audiovisual corpus, like AVSpeech [2], which includes a wide range of noise conditions. This training enables the DNN to transform noisy speech signals and corresponding video frames into a clean speech estimate. These models are typically complex, containing millions of parameters.

On the other hand, unsupervised methods [3-5] employ statistical modeling combined with DNNs. These methods use deep generative models, such as variational autoencoders (VAEs) [6] and diffusion models [7], trained on clean datasets like TCD-TIMIT [8], to probabilistically estimate clean speech signals. Since these models do not train on noisy data, they are generally lighter than supervised models and may offer better generalization capabilities and robustness to visual noise, as indicated by their probabilistic nature [3-5]. Despite these advantages, unsupervised methods remain less explored compared to their supervised counterparts.

Main activities

Objectives. In this project, we aim to develop a robust and efficient AVSE framework by thoroughly exploring the integration of recent deep-learning architectures designed for speech enhancement, encompassing both supervised and unsupervised approaches. Our goal is to leverage the strengths of both strategies alongside cutting-edge generative modeling techniques to bridge their gap. This includes the implementation of computationally efficient multimodal (latent) diffusion models, dynamical VAEs [9], temporal convolutional networks (TCNs) [10], and attention-based methods [11]. The main objectives of the project are outlined as follows:

1. Develop a neural architecture that assesses the reliability of lip images—whether they are frontal, non-frontal, occluded, in extreme poses, or missing—by providing a normalized reliability score at the output [12];
2. Design deep generative models that efficiently exploit the sequential nature of data and effectively fuse audio-visual features;
3. Integrate the visual reliability analysis network within the deep generative model to selectively use visual data. This will enable a flexible and robust framework for audio-visual fusion and enhancement.

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," in *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," in *SIGGRAPH* 2018.

[3] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," in *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 1788 –1800, 2020.

[4] A. Golmakani, M. Sadeghi, and R. Serizel, "Audio-visual Speech Enhancement with a Deep Kalman Filter Generative Model," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Rhodes Island, June 2023.

[5] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised Speech Enhancement with Diffusion-based Generative Models," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Seoul, Korea, April 2024.

[6] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," in*Foundations and Trends in Machine Learning*, vol. 12, no. 4, 2019.

[7] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2021.

[8] N. Harte and E. Gillen, "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," in*IEEE Transactions on Multimedia*, vol.17, no.5, pp.603-615, May 2015.

[9] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," in *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 (ECCV)*, 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 (NeurIPS)*, 2017, pp. 5998–6008.

[12] Z. Kang, M. Sadeghi, R. Horaud, and X. Alameda-Pineda, "Expression-preserving Face Frontalization Improves Visually Assisted Speech Processing", in *International Journal of Computer Vision (IJCV)*, 2022.

Skills

The preferred profile is described below.

- Master's degree, or equivalent, in the field of 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.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration

2788€ gross/month

General Information

- **Theme/Domain :** Language, Speech and Audio

Statistics (Big data) (BAP E)

- **Town/city :** Villers lès Nancy
- **Inria Center :** Centre Inria de l'Université de Lorraine
- **Starting date :** 2024-10-01
- **Duration of contract :** 2 years
- **Deadline to apply :** 2024-08-18

Contacts

- **Inria Team :** MULTISPEECH
- **Recruiter :**
  Sadeghi Mostafa / mostafa.sadeghi@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

Interested candidates are encouraged to contact Mostafa Sadeghi (mostafa.sadeghi@inria.fr), Xavier Alameda-Pineda (xavier.alameda-pineda@inria.fr), and Romain Serizel (romain.serizel@loria.fr), and upload the required documents (CV, transcripts, motivation letter, and recommendation letters) to the dedicated Inria Job platform.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.