Offer #2023-05881

PhD Position F/M Audio-visual Speech Enhancement: Bridging the Gap between Supervised & Unsupervised Approaches

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Context

This is a fully-funded PhD position as part of the**REAVISE** project: *"Robust and Efficient Deep Learning-based Audio-visual Speech Enhancement"* (2023-2026). REAVISE aims to develop a unified audio-visual speech enhancement (AVSE) framework that robustly integrates acoustic data (noisy speech signal) with accompanying visual information (video of speaker's lip movements) in order to recover an intelligible, high-quality estimate of the clean speech signal with low computational power and independently of the acoustic and visual noise environments. These objectives will be achieved by leveraging the recent methodological breakthroughs in statistical signal processing, machine learning, computer vision, and deep neural networks.

The PhD candidate will join theMULTISPEECH team at Inria, Nancy - Grand Est., France, and will work under the co-supervision of Mostafa Sadeghi (researcher, Inria), and Romain Serizel (associate professor, University of Lorraine). The candidate will benefit from the research environment, expertise, and powerful computational resources of the team.

Assignment

Background: Audio-visual speech enhancement (AVSE) refers to the task of improving the intelligibility and quality of a noisy speech signal by incorporating visual information, i.e., video of speaker's lip movements [1]. Visual modality can help to differentiate the target speech signal from background noise, especially in highly noisy environments. Recently, AVSE has been extensively revisited due to the great success and progress of deep neural network (DNN) architectures. Existing DNN-based AVSE methods are categorized into *supervised* and *unsupervised* approaches. In the former category, a DNN is trained to map a noisy speech signal and the associated video frames of the speaker into a clean estimate of the target speech signal. On the other hand, unsupervised methods [2] utilize a statistical model-based approach that combines deep generative models such as variational autoencoders (VAEs) [3] with DNNs to learn the prior distribution of clean speech signals without training on noisy data. The estimated clean speech signal is obtained in a probabilistic manner by combining the learned prior distribution with a statistical observation model.

Supervised methods require very deep and complex neural networks, with millions of parameters, and a large audio-visual dataset with diverse enough noise instances to achieve robustness against acoustic noise. There is also no systematic way to efficiently handle visual noise, e.g., head movements, face occlusions, changing illumination conditions, or missing video frames. Unsupervised methods, on the other hand, have a higher potential for performance generalization and can achieve robustness to visual noise, thanks to their probabilistic modeling framework [2, 5]. Despite these potential advantages, unsupervised methods have been less explored and may have some limitations, such as a complex and iterative inference phase.

Main tasks: The principal objective of this PhD project is to combine the strengths of both supervised and unsupervised AVSE approaches to create a unified framework that bridges the gap between the two. To this end, we will target three main tasks: *1) Developing novel neural architectures that efficiently and robustly integrate the acoustic and visual modalities*, *2) Designing data-efficient AVSE models and frameworks that generalize well to different acoustic and visual environments*, and *3) Developing fast and computationally efficient inference algorithms*. Regarding Task 1, we will explore, design, and implement novel audio-visual fusion methodologies that incorporate the strengths of existing fusion approaches, especially attention-based mechanisms [4], while alleviating their shortcomings. To accomplish Task 2, we will devise efficient acoustic noise modeling frameworks, a robust reliability-aware visual processing module [5, 6], and a systematic noise-aware training procedure without increasing the overall model complexity. Finally, concerning Task 3, we will explore lightweight models, e.g., based on Transformers [7], and also formulate a unified optimization-based inference procedure, inspired by [8], involving all the latent variables and parameters. Efficient and parallelizable optimization algorithms exploiting the recent breakthroughs in the optimization domain will also be developed. Furthermore, we will explore deep unrolling techniques [9] to bridge the gap between the inference phases of supervised

and unsupervised AVSE approaches. Publicly available audiovisual speech datasets, including AVSpeech [10], LRS3 [11], and TCD-TIMIT [12] will be used for this project.

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] 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.

[3] D. P. Kingma and M. Welling. "An introduction to variational autoencoders." Foundations and Trends® in Machine Learning 12, no. 4, pp. 307-392, 2019.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[5] M. Sadeghi and X. Alameda-Pineda, "Switching variational autoencoders for noise-agnostic audio-visual speech enhancement," in ICASSP, 2021.

[6] Z. Kang, M. Sadeghi, R. Horaud, and X. Alameda-Pineda, "Expression-preserving face frontalization improves visually assisted speech processing," International Journal of Computer Vision (IJCV), January 2023.

[7] 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.

[8] M. Sadeghi and R. Serizel, "Fast and Efficient Speech Enhancement with Variational Autoencoders," in IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Rhodes island, June 2023.

[9] V. Monga, Y. Li, and Y. C Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," IEEE Signal Processing Magazine, vol. 38, no. 2, pp. 18–44, 2021.

[10] 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.

[11] T. Afouras, J. S. Chung, and A. Zisserman. "LRS3-TED: a large-scale dataset for visual speech recognition." *arXiv preprint arXiv:1809.00496,* 2018.

[12] 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.

Main activities

- Literature review and research on the subject
- Developing original algorithms and methodologies
- Implementing and evaluating the developed algorithms in Python/PyTorch
- Writing & presenting scientific articles on the obtained results
- Dissertation writing and thesis defense

Skills

- Master's degree, or equivalent, in the field of speech/audio processing, computer vision, machine learning, or a related field,
- Experience in deep learning and neural architectures,
- Proficiency in programming languages (Python and PyTorch),
- Ability to work independently as well as in a team,
- High written and spoken English skills.

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

Salary: 2051€ gross/month for 1st and 2nd year. 2158€ gross/month for 3rd year.

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Scientific computing (BAP E)
- **Town/city :** Villers lès Nancy
- **Inria Center :** Centre Inria de l'Université de Lorraine
- **Starting date :** 2023-10-02
- **Duration of contract :** 3 years
- **Deadline to apply :** 2023-06-18

Contacts

- **Inria Team :** MULTISPEECH
- **PhD Supervisor :**
  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) and Romain Serizel (romain.serizel@loria.fr), attaching their CV, motivation letter, and transcripts. They should also apply via the Inria job platform (https://jobs.inria.fr/).

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.