Offer #2021-03399

PhD Position F/M Robust and Generalizable Deep Learning-based Audio-visual Speech Enhancement

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Context

Team: MULTISPEECH, Inria Nancy - Grand Est.

Contacts: Mostafa Sadeghi (mostafa.sadeghi@inria.fr) and Romain Serizel (romain.serizel@loria.fr)

The PhD thesis will be jointly supervised by**Mostafa Sadeghi** (Inria Starting Faculty Position) and **Romain Serizel** (Associate Professor, Université de Lorraine).

Assignment

Context: Audio-visual speech enhancement (AVSE) refers to the task of improving the intelligibility and quality of a noisy speech utilizing the complementary information of visual modality (lips movements of the speaker) [1]. Visual modality can help distinguish target speech from background sounds especially in highly noisy environments. Recently, and due to the great success and progress of deep neural network (DNN) architectures, AVSE has been extensively revisited. Existing DNN-based AVSE methods are categorized into supervised and unsupervised approaches. In the former category, a DNN is trained to map noisy speech and the associated video frames of the speaker into a clean estimate of the target speech. The unsupervised methods [2] follow a traditional maximum likelihood-based approach combined with the expressive power of DNNs. Specifically, the prior distribution of clean speech is learned using deep generative models such as variational autoencoders (VAEs) and combined with a likelihood function based on, e.g., non-negative matrix factorization (NMF), to estimate the clean speech in a probabilistic way. As there is no training on noisy speech, this approach is unsupervised.

Supervised methods require deep networks, with millions of parameters, as well as a large audio-visual dataset with diverse enough noise instances to be robust against acoustic noise. There is also no systematic way to achieve robustness to visual noise, e.g., head movements, face occlusions, changing illumination conditions, etc. Unsupervised methods, on the other hand, show a better generalization performance and can achieve robustness to visual noise thanks to their probabilistic nature [3]. Nevertheless, their test phase involves a computationally demanding iterative process, hindering their practical use.

Main activities

Project description: In this PhD project, we are going to bridge the gap between supervised and unsupervised approaches, benefiting from both worlds. The central task of this project is to design and implement a unified AVSE framework having the following features: 1- Robustness to visual noise, 2- Good generalization to unseen noise environments, and 3- Computational efficiency at test time. To achieve the first objective, various techniques will be investigated, including probabilistic switching (gating) mechanisms [3], face frontalization [4], and data augmentation [5]. The main idea is to adaptively lower bound the performance by that of audio-only speech enhancement when the visual modality is not reliable. To accomplish the second objective, we will explore techniques such as acoustic scene classification combined with noise modeling inspired by unsupervised AVSE, in order to adaptively switch to different noise models during speech enhancement. Finally, concerning the third objective, lightweight inference methods, as well as efficient generative models, will be developed. We will work with the AVSpeech [6] and TCD-TIMIT [7] audio-visual speech corpora.

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,” arXiv:2008.09586, 2020.

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

[4] Z. Kang, M. Sadeghi, R. Horaud, “Face Frontalization Based on Robustly Fitting a Deformable Shape Model to 3D Landmarks,” arXiv:2010.13676, 2020.

[5] S. Cheng, P. Ma, G. Tzimiropoulos, S. Petridis, A. Bulat, J. Shen, M. Pantic, “Towards Pose-invariant Lip Reading,” in ICASSP, 2020.

[6] 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.

[7] 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.

Skills

- 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),
- A decent 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

Salary: 1982€ gross/month for 1st and 2nd year. 2085€ gross/month for 3rd year.

Monthly salary after taxes : around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year. (medical insurance included).

General Information

- **Theme/Domain :** Language, Speech and Audio
- **Town/city :** Villers lès Nancy
- **Inria Center :** Centre Inria de l'Université de Lorraine
- **Starting date :** 2021-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2021-04-25

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

Application deadline: April 25th, 2021 (Midnight Paris time)

How to apply: Upload your file on jobs.inria.fr in a single pdf or zip file, and send it as well by email to

mostafa.sadeghi@inria.fr & romain.serizel@loria.fr. Your file should contain the following documents:

· Your CV.

· A cover/motivation letter describing your interest in this topic.

· A short (max one page) description of your Master thesis (or equivalent) or of the work in progress if not yet completed.

· Your degree certificates and transcripts for Bachelor and Master (or the last 5 years).

· Master thesis (or equivalent) if it is already completed and publications if any (it is not expected that you have any). Only the web links to these documents are preferable, if possible.

In addition, one recommendation letter from the person who supervises(d) your Master thesis (or research project or internship) should be sent directly by his/her author to mostafa.sadeghi@inria.fr & romain.serizel@loria.fr.

Applications are to be sent as soon as possible.

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.