Offer #2022-05548

Diffusion-based Deep Generative Models for Audio-visual Speech Modeling

Level of qualifications required : Master's or equivalent

Fonction : Internship Research

Context

This master internship 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 AVSE framework that leverages recent methodological breakthroughs in statistical signal processing, machine learning, and deep neural networks in order to design a robust and efficient AVSE framework.

The intern will be supervised by Mostafa Sadeghi (researcher, Inria) and Romain Serizel (associate professor, University of Lorraine), as members of the MULTISPEECH team, and will benefit from the research environment, expertise, and computational resources (GPU & CPU) of the team.

Assignment

Recently, diffusion models have gained much attention due to their powerful generative modeling performance, in terms of both the diversity and quality of the generated samples [1]. It consists of two phases, where during the so-called forward diffusion process, input data are mapped into Gaussian noise by gradually perturbing the data. Then, during a reverse process, a denoising neural network is learned that removes the added noise at each step, starting from pure Gaussian noise, to eventually recover the original clean data. Diffusion models have found numerous successful applications, particularly in computer vision, e.g., text-conditioned image synthesis, outperforming previous generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs), and normalizing flows (NFs). Diffusion models have also been successfully applied to audio and speech signals, e.g., for audio synthesis [2] and speech enhancement [3].

Main activities

Despite their rapid progress and application extension, diffusion models have not yet been applied to audiovisual speech modeling. This task involves joint modeling of audio and visual modalities, where the latter concerns the lip movements of the speaker, as there is a correlation between what is being said and the lip movements. This joint modeling effectively incorporates the complementary information of visual modality for speech generation. Such a framework has already been established based on VAEs [4]. Given the great potential and advantages of diffusion models, in this project, we would like to develop a diffusion-based audio-visual generative modeling framework, where the generation of audio modality, i.e., speech, is conditioned on the visual modality, i.e., lip images, similarly to text-conditioned image synthesis. This might then serve as an efficient representation learning framework for downstream tasks, e.g., audio-visual speech enhancement (AVSE) [4].

References

[1] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M. H. Yang, Diffusion models : A comprehensive survey of methods and applications arXiv preprint arXiv :2209.00796, 2022. 4

[2] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, Diffwave : A versatile diffusion model for audio synthesis arXiv preprint arXiv :2009.09761, 2020.

[3] Y. J. Lu, Z. Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, Conditional diffusion probabilistic model for speech enhancement IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[4] 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.

Skills

Background in statistical (speech) signal processing, computer vision, machine learning, and deep learning frameworks (Python, PyTorch) is preferred.

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

About 500 euros/month

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Statistics (Big data) (BAP E)
- **Town/city :** Villers lès Nancy
- **Inria Center :** Centre Inria de l'Université de Lorraine
- **Starting date :** 2023-03-01
- **Duration of contract :** 6 months
- **Deadline to apply :** 2023-02-15

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 should submit their transcripts, a detailed CV, and a cover letter (optional).

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.