Offer #2024-08376

Internship - Contrastive Multimodal Pretraining for Noise-aware Diffusion-based Audio-visual Speech Enhancement

Contract type : Internship agreement

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 audio-visual speech enhancement (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 byMostafa 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 intern will benefit from the research environment, expertise, and powerful computational resources (GPUs & CPUs) of the team.

Assignment

Diffusion models represent a cutting-edge class of generative models highly effective in modeling natural data, such as images and audio [1]. These models function through a forward (noising) process that incrementally transforms training data into Gaussian noise, paired with a reverse (denoising) process that reconstructs the original data point from noise. Recently, diffusion models have demonstrated promising performance for unsupervised speech enhancement [2]. By leveraging these models as data-driven priors for clean speech, they enable the enhancement of noisy speech data by estimating clean speech through posterior sampling, effectively separating it from background noise. Additionally, the integration of video as conditioning information into the speech model further augments the enhancement capability, utilizing visual cues from the target speaker to improve the performance [3]. This approach underscores the potential of combining audio and visual data to improve speech quality, especially in highly noisy environments.

Contrastive learning further extends the functionality of multimodal integration, as evidenced by models like CLIP (Contrastive Language–Image Pre-training) [4] and CLAP (Contrastive Language–Audio Pre-training) [5], which bridge disparate modalities such as text with image and audio. These models create a shared multimodal embedding space that supports various applications, from text-to-image generation to sophisticated audio processing tasks. Although these models have been used in some audio tasks like source separation [6], generation [7], classification [8] or localization [9], their application in audio-visual speech enhancement is highly under-explored.

Main activities

The primary objective of this project is to refine and expand the capabilities of audio-visual speech enhancement through the strategic incorporation of additional modal information into the noise model. By utilizing either textual descriptions or visual representations of the noise environment, such as videos or images depicting the acoustic scene, we aim to enhance the model’s ability to identify and differentiate noise sources effectively. This would involve developing robust contrastive learning techniques to manage the discrepancies between training and testing conditions, such as training with textual noise descriptions and testing with visual data, thanks to the shared multimodal embedding space.

To address these challenges, we propose to:

- Develop a contrastive learning framework that can dynamically adapt to different modalities of noise information, ensuring that the system remains effective regardless of the variability in available data type at training and test times.
- Utilize the shared embedding space learned through contrastive methods as conditioning information for the noise model to improve the performance of speech enhancement systems, making them more adaptable and effective in diverse and noisy environments.

References

[1] 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.

[2] 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), 2024.

[3] J.-E. Ayilo, M. Sadeghi, R. Serizel, and X. Alameda-Pineda, Diffusion-based Unsupervised Audio-visual Speech Enhancement HAL preprint hal-04718254, 2024.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al., Learning transferable visual models from natural language supervision In International Conference on Machine Learning, pp. 8748-8763, PMLR, 2021.

[5] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, Clap learning audio concepts from natural language supervision In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, IEEE, 2023.

[6] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, and W. Wang Separate anything you describe arXiv preprint arXiv :2308.05037 2023.

[7] Y. Yuan, H. Liu, X. Liu, X. Kang, P. Wu, M. D. Plumbley, and W. Wang, Text-driven foley sound generation with latent diffusion model arXiv preprint arXiv :2306.10359 2023.

[8] A. Guzhov, F. Raue, J. Hees, A. Dengel Audioclip: Extending clip to image, text and audio. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022.

[9] T. Mahmud, and D. Marculescu Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization In IEEE/CVF Winter Conference on Applications of Computer Vision 2023.

Skills

Preferred qualifications for candidates include a strong foundation in statistical (speech) signal processing, and computer vision, as well as expertise in machine learning and proficiency with deep learning frameworks, particularly PyTorch.

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

€ 4.35/hour

General Information

- **Theme/Domain :** Language, Speech and Audio
  Scientific computing (BAP E)
- **Town/city :** Villers lès Nancy
- **Inria Center :** Centre Inria de l'Université de Lorraine
- **Starting date :** 2025-04-01
- **Duration of contract :** 6 months
- **Deadline to apply :** 2025-01-15

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

Prospective applicants are invited to submit their academic transcripts, a detailed curriculum vitae (CV), and, if they choose, a cover letter. The cover letter should highlight the reasons for their enthusiasm and interest in this specific project.

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.