Offer #2023-06843

Diffusion-based Unsupervised Audio-visual Speech Enhancement

**Contract type:** Fixed-term contract

**Level of qualifications required:** Master's or equivalent

**Fonction:** Internship Research

**Context**

- Afrikaans
- Albanian - shqipe
- Arabic - عربية
- Armenian - հայերեն
- Azerbaijani - azərbaycanca
- Basque - euskara
- Belarusian - беларуская
- Bengali - বাংলা
- Bulgarian - български
- Catalan - català
- Chinese - 中文
- Croatian - hrvatski
- Czech - čeština
- Danish - dansk
- Dutch - Nederlands
- English
- Esperanto - esperanto
- Estonian - eesti
- Filipino
- Finnish - suomi
- French - français
- Galician - galego
- Georgian - ქართული
- German - Deutsch
- Greek - Ελληνικά
- Gujarati - ગુજરાતી
- Haitian Creole - kreyòl ayisyen
- Hebrew - אידיש
- Hindi - हिन्दी
- Hungarian - magyar
- Icelandic - íslenska
- Indonesian - Bahasa Indonesia
- Irish - Gaeilge
- Italian - italiano
- Japanese - ญี่ปุ่น
- Kannada - ಕನ್ನಡ
- Korean - 한국
- Latin - Lingua Latina
- Latvian - latviešu
- Lithuanian - lietuvių
- Macedonian - македонски
- Malay - Bahasa Melayu
- Maltese - Malti
- Norwegian - norsk
- Persian - فارسی
- Polish - polski
- Portuguese - português
- Romanian - română
- Russian - روسی
e- Serbian - српски
- Slovak - slovenčina
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 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 intern will benefit from the research environment, expertise, and powerful computational resources (GPUs & CPUs) 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]. These models operate in two distinct phases. During a so-called forward diffusion process, training data are gradually perturbed by Gaussian noise. Then, during a reverse diffusion process, a denoising neural network is learned to revert the forward process by gradually removing 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 the task of speech enhancement [2], [3]. However, all these works operate in a supervised settings, where a paired training dataset of clean and noisy speech signals is used. This could greatly limit their generalization performance, as the test data might contain noise conditions not seen during training, e.g., a new noise type and/or noise level. In a recent work [4], we have developed an unsupervised approach based on diffusion models to overcome the generalization issue. In this method, the training is done over only clean speech data, and the noise is modelled separately at test time, thus adapting to new noise conditions.

**Main activities**
The primary goal of this project is to expand upon the diffusion-based audio-only speech enhancement framework presented in [4] to address the Audiovisual Speech Enhancement (AVSE) task [5]. AVSE involves enhancing the quality and intelligibility of a noisy speech signal by leveraging the complementary information provided by the visual modality, namely the speaker's lip movements. The visual modality proves especially valuable in high-noise environments, as it is less susceptible to acoustic interference. Consequently, AVSE holds the potential for various practical applications, including hearing assistive devices.

The fundamental idea here is to develop an audio-visual diffusion-based generative model for clean speech, with speech generation conditioned on video data (image frames) capturing the speaker's lip movements. This necessitates the implementation of efficient visual data encoders and fusion networks, enabling the robust integration of visual information into the speech generation framework. Ideally, the trained generative model would be capable of generating speech signals that align with the conditioned visual data, i.e., the speaker's lip movements (see [6]). Subsequently, this model will be employed for speech enhancement during testing by combining it with a parametric noise model, following the same methodology as in [4].

References


Skills

translator

• Afrikaans
• Albanian - shqipe
• Arabic - العربية
• Armenian - Հայերեն
• Azerbaijani - azərbaycanca
• Basque - euskara
• Belarusian - белarusкая
• Bengali - বাংলা
• Bulgarian - български
• Catalan - català
• Chinese - 中文
• Chinese - 中文
• Croatian - hrvatski
• Czech - čeština
• Danish - dansk
• Dutch - Nederlands
• English
• Esperanto - esperanto
• Estonian - eesti
• Filipino
• Finnish - suomi
• French - français
• Galician - galego
• Georgian - ქართული
• German - Deutsch
• Greek - ελληνικά
• Gujarati - ગુજરાતી
• Haitian Creole - kreyòl ayisyen
• Hebrew - עברית
• Hindi - हिन्दी
• Hungarian - magyar
• Icelandic - íslenska
• Indonesian - Bahasa Indonesia
• Irish - Gaeilge
• Italian - italiano
• Japanese - 日本語
• Kannada - ಕನ್ನಡ
• Korean - 한국어
• Latin - Lingua Latina
• Latvian - latviešu
• Lithuanian - lietuvių
• Macedonian - македонски
• Malay - Bahasa Melayu
• Maltese - Malti
• Norwegian - norsk
• Persian - پارسی
• Polish - polski
• Portuguese - português
• Romanian - română
• Russian - русский
• Serbian - српски
• Slovak - slovenčina
• Slovenian - slovenščina
• Spanish - español
• Swahili - Kiswahili
• Swedish - svenska
• Tamil - தமிழ்
• Telugu - తెలుగు
• Thai - ภาษาไทย
• Turkish - Türkçe
• Ukrainian - українська
• Urdu - اردو
Preferred qualifications for candidates include a strong foundation in statistical (speech) signal processing, 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**

Smic level

**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**: 2024-04-01
- **Duration of contract**: 5 months
- **Deadline to apply**: 2024-02-11

**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.