2022-05549 - Efficient Attention-based Audio-visual Fusion Mechanisms for Speech Enhancement

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 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

Audiovisual speech enhancement (AVSE) is defined as the task of improving the quality and intelligibility of a noisy speech signal by utilizing the complementary information provided by the visual modality, i.e., lip movements of the speaker [1]. Visual modality is especially important in high-noise situations, as it is less affected by acoustic noise. Because of that, AVSE could be exploited in several practical applications, including hearing assistive devices. Numerous works have already studied the integration of visual modality with audio modality to improve the performance of speech enhancement. While the majority of audiovisual speech enhancement algorithms rely on deep neural networks and supervised learning, they require very large audiovisual datasets with diverse noise instances to have good generalization performance.

A recently introduced AVSE approach is based on unsupervised learning [2,3], where during a training phase, the statistical distribution of clean speech is learned from a clean audiovisual dataset. This is done using a deep generative model, e.g. variational autoencoder (VAE) [4]. Then, at test (inference) time, the learned distribution is combined with a noise model to estimate the clean speech signal from the available noisy speech observations.

Main activities

An important element of AVSE is audio-visual feature fusion, which should robustly and efficiently combine the two modalities. Current fusion mechanisms used for unsupervised AVSE are based on simple feature concatenation, which is not effective, as it treats the two feature streams on an equal basis. In fact, the audio modality usually contributes more than the visual modality, but in general, their contributions should be robustly balanced and weighted. In this project, we are going to develop efficient feature fusion modules based on attention models [5], which have proven very successful in different applications. The designed fusion module is supposed to robustly and efficiently incorporate the potentially different uncertainty (reliability) levels of the two modalities. We will then evaluate its effectiveness for AVSE.

References:

[1] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learningbased audio-visual speech enhancement and separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1368–1396, 2021.

[2] M. Sadeghi and X. Alameda-Pineda, "Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[3] 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.

[4] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," Foundations and Trends in Machine Learning, 2019.

[5] 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, 2017.

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

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Statistics (Big data) (BAP E)
- **Town/city :** Villers lès Nancy
- **Inria Center :** CRI Nancy - Grand Est
- **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).

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.

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.

Remuneration

About 500 euros/month