Offre n°2022-05549

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

Le descriptif de l'offre ci-dessous est en Anglais

Niveau de diplôme exigé : Bac + 4 ou équivalent

Fonction : Stagiaire de la recherche

Contexte et atouts du poste

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.

Mission confiée

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.

Principales activités

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.

Compétences

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

Avantages

- 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

Rémunération

About 500 euros/month

Informations générales

- **Thème/Domaine :** Optimisation, apprentissage et méthodes statistiques
  Statistiques (Big data) (BAP E)
- **Ville :** Villers lès Nancy
- **Centre Inria :** Centre Inria de l'Université de Lorraine
- **Date de prise de fonction souhaitée :** 2023-03-01
- **Durée de contrat :** 6 mois
- **Date limite pour postuler :** 2023-02-15

Contacts

- **Équipe Inria :** MULTISPEECH
- **Recruteur :**
  Sadeghi Mostafa / mostafa.sadeghi@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

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

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.