2022-05548 - Diffusion-based Deep Generative Models for Audio-visual Speech Modeling

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 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

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].

Principales activités

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.

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 :** CRI Nancy - Grand Est
- **Date de prise de fonction souhaitée :** 2023-03-01
- **Durée de contrat :** 6 mois
- **Date limite pour postuler :** 2023-02-15

Contacts

- **Equipe 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 200 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3500 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 180 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).

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.