



Offre n°2024-07443

PhD Position F/M End-to-end speech-to-sign language generation

Type de contrat : Fixed-term contract

Niveau de diplôme exigé : Graduate degree or equivalent

Fonction : PhD Position

Contexte et atouts du poste

The PhD project will be carried out in the [MULTISPEECH team](#) at [Loria](#). The PhD candidate will be supervised by [Slim Ouni](#) (professor, University of Lorraine) and [Mostafa Sadeghi](#) (researcher, Inria), and will benefit from the research environment, expertise, and powerful computational resources (GPUs & CPUs) of the team.

Mission confiée

Motivation and context

Sign language generation involves translating the spoken or written language into the visual-manual modality of sign language, effectively converting auditory or text information into corresponding sign language gestures and expressions. An automatic translation system for this task requires access to a sufficiently large parallel corpus of aligned speech and sign data. Moreover, previous work on sign language translation has shown that having an intermediate-level presentation of sign meta-symbols, known as gloss, is beneficial for translation performance. Gloss is essentially a morpheme-by-morpheme "translation" using English words. However, the field of sign language research does not have large-scale gloss-annotated corpora that would allow for the immediate use of a sign language generation system. Most existing corpora come from small discourse domains with a limited vocabulary, such as weather forecasts [1]. These corpora often present inherent problems with the acquisition itself, such as low resolution, motion blur, and interlacing artifacts.

Moreover, a main limitation of existing sign language generation systems is that the introduction of any intermediate representation removes some information from the source message. More precisely, the intermediation of text, obtained from input speech using automatic recognition systems, removes prosodic information carried by speech. The intermediation of glosses removes information about the inflection of the execution on signs with respect to their citation form.

Principales activités

Objectives

This project involves modeling the generation of sign gestures from speech. It aims to achieve direct translation from continuous speech, rather than text, to sign language through an end-to-end approach, bypassing the need for gloss annotations. Its main goal is to create a model that can produce high-quality, photorealistic animations of a 3D avatar straight from speech inputs. This will be accomplished by utilizing the latest developments in large-scale speech and vision-language modeling [2], self-supervised/unsupervised learning [3], and natural language processing techniques.

We will be building upon the work of [4], to develop a system based on a diffusion model [5]. We will build a conditional generative model capable of generating gesture data conditioned on input speech. In this process, we discard the intermediate conversion stage from text to gloss and directly perform a more efficient translation from spoken language to pose. For this project, we will use public corpora of parallel sign data, for fine-tuning and semi-supervised learning purposes, and a large corpus of unannotated sign language gestures and speech collected and partially preprocessed for German.

Addressing the challenge of limited labeled data, our project also explores the impact of applying a transfer learning strategy. This method aims to enhance the model's capacity for gesture representation and uncover deeper insights into the gesture production process. Transfer learning, a strategy where a model trained on one task is adapted for use on a related but different task, is particularly valuable in scenarios with scarce data. Through this investigation, we aim not only to improve gesture generation quality but also to achieve a more profound understanding of model behavior. This could lead to the

development of models that are not only more interpretable but also capable of generating more natural and expressive gestures.

References

- [1] H. Cooper and R. Bowden, Learning signs from subtitles: A weakly supervised approach to sign language recognition, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. [2568-2574](#), 2009.
- [2] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 20871–20881.
- [3] Guo, Z., He, Z., Jiao, W., Wang, X., Wang, R., Chen, K., Tu, Z., Xu, Y. and Zhang, M., 2024. Unsupervised Sign Language Translation and Generation. arXiv preprint [arXiv:2402.07726](#).
- [4] Fang, S., Sui, C., Zhang, X., Tian, Y. SignDiff: Learning Diffusion Models for American Sign Language Production. arXiv preprint [arXiv:2308.16082](#), 2023.
- [5] 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.

Compétences

Preferred qualifications for candidates include expertise in machine learning and proficiency with deep learning frameworks, particularly PyTorch. A background in statistical signal processing (especially speech) and/or computer vision is a plus.

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

2100€ gross/month the 1st year

Informations générales

- **Thème/Domaine** : Language, Speech and Audio Statistics (Big data) (BAP E)
- **Ville** : Villers lès Nancy
- **Centre Inria** : [Centre Inria de l'Université de Lorraine](#)
- **Date de prise de fonction souhaitée** : 2024-10-01
- **Durée de contrat** : 3 years
- **Date limite pour postuler** : 2024-04-30

Contacts

- **Équipe Inria** : [MULTISPEECH](#)
- **Directeur de thèse** :
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

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.

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.