Deep Generative Modeling of Multi-microphone Speech Signals

General information

Position: Master internship.
Duration: 5 to 6 months, starting in March or April 2024.
Location: CentraleSupélec campus of Rennes (France).
Affiliation: AIMAC team of the IETR laboratory (UMR CNRS 6164).
Supervisor: Simon LEGLAIVE.

Context

This master internship is part of the DEGREASE project (2024-2027) funded by the French National Research Agency (ANR).

DEGREASE aims to develop speech enhancement methods that can leverage real unlabeled noisy and reverberant speech recordings at training time and that can adapt to new unseen acoustic conditions at test time. At the crossroads of audio signal processing, probabilistic graphical modeling, and deep learning, the DEGREASE project proposes a methodology based on deep generative and inference models specifically designed for the processing of multi-microphone speech signals.

Description and objectives

High-dimensional data such as natural images, speech, or music signals exhibit some form of regularity, preventing their dimensions from varying independently from each other. From a generative perspective, this suggests that there exists a latent variable of much lower dimension from which the observed high-dimensional data were generated. Uncovering the hidden explanatory features involved in the generation of complex data is the goal of representation learning, and deep generative models with latent variables have emerged as promising unsupervised approaches. In particular, the variational autoencoder (VAE) (Rezende et al., 2014; Kingma & Welling, 2014), which is equipped with both a generative and inference model, allows for the analysis, transformation, and generation of various types of data. Over recent years, the VAE has been extended in many ways, including for processing multimodal data (Wu & Goodman, 2018), sequential data (Girin et al., 2021), or multimodal and sequential data (Sadok et al., 2023a).

Deep generative models of speech signals in the literature have focused on a single-microphone scenario. The objective of this internship is to develop a deep generative model of multi-microphone speech signals where the spatial and spectro-temporal characteristics of the signal are disentangled in the learned latent representation.

A key to learning a meaningful latent representation of multi-sensor temporal data lies in the structuring of the dependencies in the probabilistic graphical model (Sadok et al., 2023a). In this internship, we propose to build hybrid models combining (i) insights from traditional multi-microphone audio signal processing (Gannot et al., 2017); (ii) recent advances in disentangled representation learning of speech signals with (dynamical) VAEs (Girin et al., 2021; Sadok et al., 2023a, 2023b); (iii) temporal convolutional neural architectures for efficient and high-quality synthesis of audio waveforms (Caillon & Esling, 2021; Zeghidour et al., 2021; Défossez et al., 2022).

The model will be validated through quantitative and qualitative evaluations of multi-microphone speech signal manipulations in the learned latent space. It might also be validated as a clean speech model for a speech enhancement task (e.g., Bie et al., 2022).

This internship could lead to a Ph.D. position fully funded by the DEGREASE project, on the topic of deep generative models for weakly-supervised speech enhancement.

Candidate profile

The candidate will be pursuing his/her last year of Master's or engineer's degree. He/she should have good knowledge and practical skills in machine learning and/or audio signal processing. A good practice in Python is required, experience with PyTorch would be a plus. The candidate should also have good oral and written communication skills.

How to apply

Interested candidates should submit their transcripts, a detailed CV, and a cover letter to Simon LEGLAIVE (simon.leglaive@centralesupelec.fr).

Work environment

The intern will be supervised by Simon LEGLAIVE and will integrate the AIMAC team of the IETR laboratory, located in the CentraleSupélec's campus of Rennes (in Brittany, France). CentraleSupélec offers accommodation on the campus.

The intern will benefit from the research environment of CentraleSupélec, in particular the computational resources of the Mésocentre.

The intern will receive the legal internship gratification of about 600 €/month.

References

Bie, X., Leglaive, S., Alameda-Pineda, X., & Girin, L. (2022). Unsupervised speech enhancement using dynamical variational autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 2993-3007.

Caillon, A., & Esling, P. (2021). RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv preprint arXiv:2111.05011.

Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High fidelity neural audio compression. arXiv preprint arXiv:2210.13438.

Gannot, S., Vincent, E., Markovich-Golan, S., & Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4), 692-730.

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X. (2021). Dynamical variational autoencoders: A comprehensive review. Foundations and Trends in Machine Learning, 15(1-2), 1-175.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representation (ICLR).

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. International Conference on Machine Learning (ICML)

Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., & Séguier, R. (2023a). A multimodal dynamical variational autoencoder for audiovisual speech representation learning. arXiv preprint arXiv:2305.03582.

Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., & Séguier, R. (2023). Learning and controlling the source-filter representation of speech with a variational autoencoder. Speech Communication, 148, 53-65.

Wu, M., & Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. Advances in neural information processing systems (NeurIPS), 31.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., & Tagliasacchi, M. (2021). SoundStream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 495-507.