Ph.D. student position on deep generative models for weakly-supervised speech enhancement

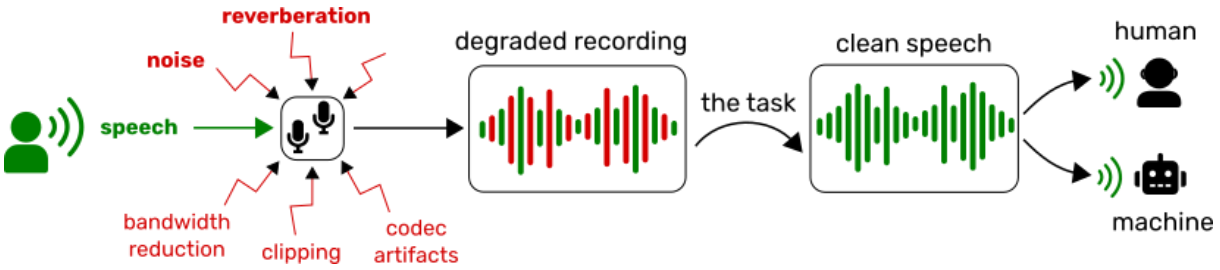
General information

Position: Fully funded Ph.D. student position.
Duration: 3 years, starting in September/October 2024 (flexible).
Location: CentraleSupélec campus of Rennes (France).
Affiliation: AIMAC team of the IETR laboratory, CNRS joint research unit (UMR) 6164.
Supervisor: Simon LEGLAIVE.
Application deadline: April 30th, 2024.

This Ph.D. student position is part of the DEGREASE project (2024-2027) funded by the French National Research Agency (ANR).

Description and objectives

Speech enhancement



Speech enhancement consists of improving the quality and intelligibility of speech in a degraded recording, for instance due to interfering sound sources and reverberation (Vincent et al., 2018). The task is to estimate a clean speech signal from the degraded recording, as illustrated above. Speech enhancement finds applications in various speech technologies for human and machine listening (hearing aids, assistive listening devices, vocal assistants, etc.)

The conventional fully-supervised approach

In recent years, there has been great progress in speech enhancement thanks to deep learning models trained in a supervised manner (Wang and Chen, 2018). Supervised speech enhancement involves three main ingredients:

1. A model, which provides a prediction of the clean speech signal given the noisy recording. State-of-the-art methods today rely on deep neural networks.
2. A metric, which measures the discrepancy between the clean speech estimate and the ground-truth signal. During training, the metric corresponds to a differentiable loss function, which is minimized to estimate the model parameters. At test time, the metric (which can differ from the training loss function) is used to evaluate the performance of

the trained model.

3. A labeled dataset, which consists of noisy speech signals and their corresponding clean reference signals; we say that the noisy speech signals are labeled with the clean speech signals. During training, the clean speech reference signals are used as the targets for the model.

Unfortunately, it is very difficult, if not impossible, to acquire labeled noisy speech signals in real-world conditions due to cross-talk between microphones. Therefore, datasets for supervised learning have to be generated artificially, by creating synthetic mixtures of isolated speech and noise signals, e.g., (Cosentino et al., 2020; Maciejewski et al., 2020). Artificially-generated training data are however inevitably mismatched with real-world noisy speech recordings, which can result in poor speech enhancement performance in case of severe mismatch (Pandey & Wang, 2020; Bie et al., 2022; Richter et al., 2023; Gonzalez et al., 2023). Moreover, when the test domain differs from the synthetic training domain, supervised learning necessitates rebuilding the synthetic training dataset and retraining the model, which is time-consuming and computationally intensive. These limitations of supervised deep learning methods for speech enhancement contrast with the impressive adaptability of the human auditory system when it comes to perceiving speech in unknown adversary acoustic conditions.

The Ph.D. project

This Ph.D. project aims to build a more effective approach to speech enhancement, where models can learn from and adapt to real, unlabeled noisy speech recordings. To reach this objective, we propose a methodology at the crossroads of audio signal processing, probabilistic graphical modeling, and deep learning, based on deep generative and inference models specifically designed for the processing of multi-microphone speech signals. The probabilistic generative modeling approach will allow us to consider the clean speech signal as partially-observed, hence enabling semi-supervised learning at training time and unsupervised adaptation at test time. Speech enhancement will be achieved by inverting the learned generative model, a.k.a performing inference.

The proposed methodology will leverage insights from multi-microphone audio signal processing (Gannot et al., 2017); advances in deep generative modeling with dynamical variational autoencoders (Girin et al., 2021), normalizing flow (Kingma et al., 2016), and diffusion models (Kingma et al., 2021; Lemercier et al., 2024); disentangled representation learning techniques for speech (Sadok et al., 2023, 2024); and temporal convolutional neural architectures for high-quality audio synthesis (Caillon & Esling, 2021; Zeghidour et al., 2021; Défossez et al., 2022).

Candidate profile

The candidate should hold a Master's or Engineer's degree, with a strong mathematical background (probability, statistics, linear algebra, optimization), good programming skills in Python, and particular interests in audio signal processing and machine/deep learning. The candidate should also have excellent oral and written communication skills.

Work environment

The Ph.D student 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 Ph.D. student will benefit from the research environment of CentraleSupélec, in particular the computational resources of Ruche, the HPC cluster of the "Mésocentre" computing center of Université Paris-Saclay, CentraleSupélec and École Normale Supérieure Paris-Saclay.

How to apply

Interested candidates should apply by filling out the following Google form before April 30th, 2024: https://forms.gle/WW7gZVtyzDpt7Cxh7

Candidates will be invited to upload (as PDF files) a resume, the official transcripts for each year of higher education, and between 1 and 3 recommendation letters.

In case of difficulty or question, please contact simon.leglaive@centralesupelec.fr.

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.

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

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., & Vincent, E. (2020). LibriMix: An open-source dataset for generalizable speech separation. arXiv preprint arXiv:2005.11262.

Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High fidelity neural audio compression. Transactions on Machine Learning Research.

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

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.

Gonzalez, P., Alstrøm, T. S., & May, T. (2023). Assessing the generalization gap of learning-based speech enhancement systems in noisy and reverberant environments. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

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

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. Advances in neural information processing systems (NeurIPS).

Kingma, D., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. Advances in neural information processing systems (NeurIPS).

Lemercier, J. M., Richter, J., Welker, S., Moliner, E., Välimäki, V., & Gerkmann, T. (2024). Diffusion models for audio restoration. arXiv preprint arXiv:2402.09821.

Maciejewski, M., Wichern, G., McQuinn, E., & Le Roux, J. (2020). WHAMR!: Noisy and reverberant single-channel speech separation. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Pandey, A., & Wang, D. (2020). On cross-corpus generalization of deep learning based speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

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

Richter, J., Welker, S., Lemercier, J.-M., Lay, B., & Gerkmann, T. (2023). Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

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.

Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., & Séguier, R. (2024). A multimodal dynamical variational autoencoder for audiovisual speech representation learning. Neural Networks.

Vincent, E., Virtanen, T., & Gannot, S. (Eds.). (2018). Audio source separation and speech enhancement. John Wiley & Sons.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

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.