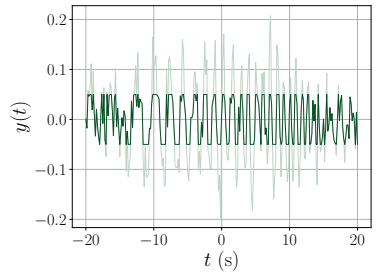
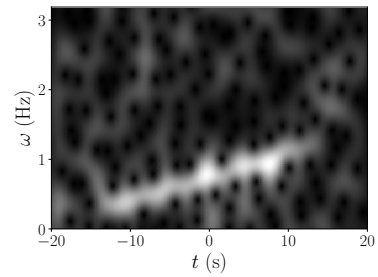
Master internship for Spring 2024 on Diffusion Models for Audio Inpainting

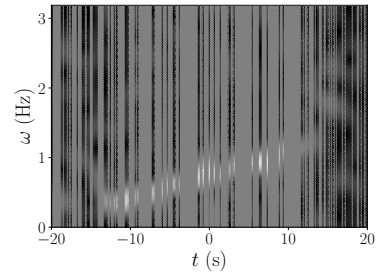
Title	Audio inpainting in the time-frequency domain through probabilistic diffusion models
Dates and duration	Spring 2024, from 4 to 6 months
Supervision	Barbara Pascal, Researcher CNRS, LS2N barbara.pascal@cnrs.fr Mathieu Lagrange, Researcher, CNRS, LS2N mathieu.lagrange@ls2n.fr
Hosting institution	Laboratoire des Sciences du Numérique de Nantes, École Centrale Nantes, Équipe SIMS
Salary	$\gtrsim$ 600 euros per month (standard internship compensation)
PhD funding	No

Context: Available audio signals often have undergone different types of degradation, either during their acquisition, e.g., the recording of a piece of music on a physical medium, and/or during their transmission, e.g., in phone conversations. A very common yet severe degradation consists in *missing samples*, encompassing the case when samples have been distorted so much that they no longer contain relevant information [1]. E.g. in top figure, the observed signal $y(t)$, plotted in dark green, is the typical output obtained from a saturating sensor: the values of the ground truth signal, plotted in light green, are cut when their amplitude is larger than 0.05, inducing locally a total loss of information. *Audio inpainting* consists in reconstructing missing samples with maximal possible accuracy.



Objective: The main goal of the project is to design probabilistic diffusion models for audio inpainting. To that aim, we propose to resort to the definition of audio inpainting stated in [4] which is formulated in terms of *spectrogram* masking. The *spectrogram* is a widely used tool in audio processing which provides the temporal evolution of the frequency content of a signal. The spectrogram of the ground truth signal, in light green in the top figure, is displayed in the second figure: it shows a line of energy maxima in white, which is the signature of a linearly increasing frequency. At times when saturation occurs, the signal and its spectrogram contain no information, resulting in the masked spectrogram shown in the third figure, where vertical shaded areas correspond to missing data to be reconstructed.





Methods: Audio inpainting, and particularly audio inpainting in the time-frequency domain, is directly inspired from *image inpainting* [2] consisting in recovering masked part of an image, as illustrated on the Cameraman photography in the fourth figure. Recently, image inpainting has been revisited with great success in the framework of deep learning. More specifically, generative models are today particularly promising to tackle inpainting problems in image processing and beyond. This project targets the use of *probabilistic diffusion models* which rely on a forward diffusion process $x_t \mid x_{t-1} = \sqrt{1-\beta_t}x_{t-1} + \zeta_t$, where $\zeta_t \sim \mathcal{N}(0, \beta_t \mathbf{I})$, starting from a clean image x_0 and ending with a pure white noise image x_T by successively adding noise of increasing variance β_t to the intermediates x_t, $t \in (0, T)$. Forward diffusion on training samples is then leveraged to learn an *inverse diffusion process* $y_t \mid y_{t+1} = \alpha_t y_{t+1} + \nu_t, \quad \nu_t \sim \mathcal{N}(\mu_t, \sigma_t^2 \mathbf{I})$ going from white noise y_T to a clean image y_0. Now widely used for image denosing [3], diffusion models have recently be applied successfully to image inpainting [5].

Research program: *i)* Reproduce the results of [5] on image inpaint with diffusion models.
ii) Run RePaint on masked spectrograms of synthetic and real world audio signals.
iii) Quantify performance in terms of standard audio metrics as a function of the mask coverage and geometry.
iv) Compare to state-of-the-art in resolution of audio inverse problems with diffusion models [6].

Prerequisite: The recruited intern is expected to be at ease with the basic concepts of deep learning and with Python programming, preferably with PyTorch. Some notions of audio processing will be a great plus and a mathematical background in stochastic processes and probabilities will be appreciated.

References

[1] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley. Audio Inpainting. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):922–932, 2011.

[2] C. Guillemot and O. Le Meur. Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine*, 31(1):127–144, 2013.

[3] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[4] A. M. Krémé, V. Emiya, C. Chaux, and B. Torrésani. Time-frequency fading algorithms based on gabor multipliers. *IEEE Journal of Selected Topics in Signal Processing*, 15(1):65–77, 2020.

[5] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[6] E. Moliner, J. Lehtinen, and V Välimäki. Solving audio inverse problems with a diffusion model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.