Master thesis Internship / Stage de Master 2 - ingénieur

Sound and music transformation by hybrid deep neural synthesis

We have several openings for **Master students internships** (*Master 2 or Engineer*) at Télécom Paris, Institut polytechnique de Paris, in the Audio group (ADASP) of the "Signal, Statistics and Learning (S2A) team. *One of the topics is further described below.*

Place: Telecom Paris, 19 place Marguerite Perey, 91120 Palaiseau, France.
Start of the positions: March/April 2024
Duration: 4 to 6 months

Possible extension to PhD (Funding on the HI-Audio project)

Context: The positions will be a part of the **ERC Advanced (2022) – HI-Audio** (*Hybrid and Interpretable Deep neural audio machines*) project, which aims at building hybrid deep approaches combining parameter-efficient and interpretable models with modern resource-efficient deep neural architectures with applications in speech/audio scene analysis, music information retrieval and sound transformation and synthesis.

Subject: Controllable music generation with neural discrete representations and a multitrack MIDI-to-audio dataset

Hybrid deep learning is a rather recent concept where only a few attempts have been made in audio, for example, in using signal processing modules with deep learning for representation learning [1], audio synthesis [2] [3] or phase recovery [4]. Neural discrete representations bridge the gap between signals and tokens and allow for controllable generation of high-quality audio via the manipulation of tokens. One well-known method for learning audio codecs is EnCodec [5], which produces a quantized latent space and minimizes the perceptual distortions introduced by audio compression, although better methods such as RVQGAN [9] are also emerging. In this project, we will use audio codecs in novel conditional music generation settings. There are a number of works which have explored music generation with audio codecs. For example, VampNet [6] uses audio codecs for inpainting and presents various possibilities of designing prompts to elicit output. Recently, controllable music generation, with conditions for both coarse as well as frame-level, has also been enabled and improved by such codecs, for example, Coco-Mulla [7], based on MusicGen [8], uses chord progressions and rhythmic patterns as control parameters.
In this internship it is proposed :

- to conduct a survey of recent work in neural discrete representations for music and how they are used in various music generation settings.
- to create a dataset of acoustic piano recordings driven by existing MIDI datasets using the in-house recording facilities at our sound studio.
- to adapt existing models for codec-based music generation for controllability via frame-level structural controls using the created dataset.

Candidate Profile:
- Students involved in a master or engineering program in *Data Science*, *Machine learning*, *Signal Processing*, *or Speech/Audio/Music processing.*

Application:
- Contact: Gaël Richard, firstname.lastname@telecom-paris.fr.

Bibliography

[1] M. Ravanelli et Y. Bengio, «Interpretable Convolutional Filters with SincNet,» *arXiv, 1811.09725,,* 2019.

[2] X. Wang, S. Takaki et J. Yamagishi, «Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis,» *in IEEE/ACM Trans. on Audio, Speech, and Language Proc.,* vol. 28, 2020.

[3] J. Engel, L. Hantrakul, C. Gu et A. Roberts, «DDSP: Differentiable Digital Signal Processing,» chez *Int. Conf. on Learning Representations (ICLR)*, 2020.

[4] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa et N. Harada, «Deep Griffin–Lim Iteration: Trainable Iterative Phase Reconstruction Using Neural Network,» *IEEE Journal of Selected Topics in Signal Proc.,* vol. 15, n° %11, 2021.

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

[6] Garcia, H. F., Seetharaman, P., Kumar, R., & Pardo, B. (2023). Vampnet: Music generation via masked acoustic token modeling. ISMIR 2023.

[7] Lin, L., Xia, G., Jiang, J., & Zhang, Y. (2023). Content-based Controls For Music Large Language Modeling. arXiv preprint arXiv:2310.17162.

[8] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., ... & Défossez, A. (2023). Simple and Controllable Music Generation. arXiv preprint arXiv:2306.05284.

[9] Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., & Kumar, K. (2023). High-Fidelity Audio Compression with Improved RVQGAN. NeurIPS 2023.