Adaptive representation learning for the gestural control of deep audio generative models

Philippe Esling[1], Frédéric Bevilacqua[1], Geoffroy Peeters[2]
[1] ACIDS / ISMM / IRCAM – UMR STMS 9912, Sorbonne – CNRS, [2] LTCI / Telecom Paris – Institut Polytechnique

Abstract

In recent years, significant advances have been made in deep learning models for audio generation. Deep generative models have achieved impressive results in generating high-quality audio samples that reproduce the properties of a given training set distribution [4]. However, controlling such models remains a arduous and daunting task, which might hamper their creative use by both expert and non-expert users. Existing control methods mainly rely on massive sets of paired examples with labelled attributes. However, in real-life control problems, these labelled data are often scarce, task-oriented and subject to human biases. Musical applications especially rely on very limited datasets that match closely a given artistic intention. Hence, deep generative models lack intuitive personalized control methods that could leverage the non-linear topology of latent representations, while promoting co-creative human-machine interaction.

This PhD aims to overcome these limitations and propose novel methods to provide dynamic user-adapted control spaces on deep audio generative models. To do so, we propose to rely on a joint approach based on implicit regularization of the latent space, and explicit transformation of a subset of dimensions to provide user-centered control spaces. We aim to target adaptive representation learning by relying on self-supervised learning between both the audio and control latent spaces, by learning joint priors between these spaces. The project will be co-supervised by three experts covering the fields of deep audio generative models, human-machine interaction and self-supervised few-shot learning. We expect high impact in the fields of creative industries and music practitioners (artists, educators and amateurs).

1 Context

Deep generative models have achieved impressive results in diverse application fields, notably audio synthesis [21, 7, 4]. These advances transformed deep models into promising tools for musical creation. However, controlling such models remains a daunting task. Indeed, despite the significant advances in generative modelling, only few attempts have been made to address the need for *control* and *interactivity*, a lack that hampers their creative use by both expert and non-expert users. *Latent-based generative models* [13] may have the potential to address these limitations as they provide high-level features representations that can be organized in a low-dimensional embedded space, called a *latent space* [25]. To do so, they learn to model the underlying data distribution of a given set of training examples to generate new samples. These methods mostly rely on two different approaches. First, the *likelihood-based models* such as *Variational AutoEncoders* (VAE) [13] and *Normalizing Flows* (NF) [25] learn the probability density function via maximum likelihood. Second, *implicit probabilistic modelling*, such as *Generative Adversarial Networks* (GAN) [10], directly sample a latent space to model the targeted data distribution. Although likelihood-based models are able to learn meaningful representations, they either require strong restrictions on the model architecture (NFs), or must rely on surrogate objectives to approximate maximum likelihood which degrades the samples quality (VAEs). In contrast, GANs have shown impressive high-quality synthesis abilities, but the adversarial training often suffers from instabilities and unreliable latent representations. To alleviate these issues, state-of-the-art models usually combine several approaches. For instance, the recent RAVE model leverages both VAEs representation abilities and adversarial training to achieve high-quality synthesis [4]. However, controlling such models remains an open challenge: latent spaces still have a prohibitively large dimensionality, which precludes straightforward and intuitive control to navigate their non-linear topology. Hence, the creative use of these models is hampered by the lack of techniques to promote an intuitive and co-creative human-machine interaction [18]. Most of these models typically work as *black-boxes* [21, 7, 6], where the intended end-user has little to no control over the generation process. Additionally, they do not allow specific modes for interaction, so the user cannot selectively modify the generated content or some of its parts based on desired characteristics.

Existing approaches for controlling deep generative models either consist in *conditioning* [19, 8], *interpolation* [28], *transfer* [11, 3] or *disentanglement* techniques to identify underlying *factors of variations* [12, 22]. Most prior works rely on supervised learning based on paired labelled attributes to *explicitly* condition the generation. However, these techniques require massive sets of paired data and gathering these high-quality annotations necessitates extensive manual efforts and expert knowledge. To alleviate this issue, some methods rely on external pre-trained features extractor models to create artificial annotations [1]. However, in real-life control problems, labelled data are scarce, task-oriented and subject to human biases. Hence, existing *explicit* control methods bypass the fundamental need for personalised control, which can only be expressed by a small set of subjective paired examples that matches closely a given user's intent. Moreover, obtaining finer control often means having multiple labelled attributes over the same example. Therefore, it becomes even more tedious to preserve independence between each factor of variation, which results in undesired changes over other features [1]. Additionally, training datasets usually do not contain all attributes combinations, leading to poorly approximated control parameters and degraded synthesis quality. Finally, adding or updating control attributes usually requires the full re-training of the model [14], which is inefficient.

Alternatively, *implicit* control approaches rely on unpaired data to guide the generation process by providing a reference sample as input [29]. While these methods bypass the need for many paired *example-attributes* samples, they cannot provide fine-grained control to the end-users and, therefore, also lack customized control features. Yet, *implicit* control methods have emerged as a promising tool to promote co-creative human-machine interaction through multi-modal approaches providing more intuitive ways of control. Indeed, recent language generation models [24] leverage text-to-image models [23] to generate and edit realistic artistic images from a text description, which benefits both expert and non-expert users. This type of binding is lacking between sound generation and motion, which could allow to explore and guide the audio generation process in a dynamic and user-centered way. While in recent years, gestural control methods have been developed to explore and control sound/music spaces, to the best of our knowledge, no attempts has been made to bind it with the latent space of deep generative models [5]. As it is essential for users to personalize the movement-sound relationship to their context of use and individual need [9], a multi-modal motion-to-sound generation approach could provide relevant insights and intuitive controls for artists and performers in their creative exploration process.

2 Research goal and methodology

The goal of this research is to propose innovative methods to dynamically adapt latent representations of generative models according to specific users who would provide a small number of examples (few-shots). This will allow for providing customized control features for deep audio synthesis, which could be used for gestural control of sound environments. Specifically, our goal is to dynamically organize the musicians' latent representations with few data to provide smooth personalized control over musical attributes even in contexts of real-time constraints. We believe that such a customized control space will enable both creative and gestural control of the generated audio, leading to democratize the use of deep models in the music fields, especially in performances. To the best of our knowledge, this is the first attempt to provide dynamic user-adapted control spaces for deep generative models.

The research methodology is structured around three major and complementary axes. First, we will focus on regularization techniques to shape the latent representation of deep audio generative models according to the user's intent in order to provide *dynamic user-adapted control spaces* (Axis 1). Then, to ensure real-life applicability and bypass the need of massive sets of paired data with labelled attributes, we will explore few-shot and self-supervised learning (SSL) to dynamically adapt the latent representations in *low-data regimes* (Axis 2). Finally, we will leverage a multi-modal approach relying on the previously-introduced techniques to provide an unsupervised *custom mapping between motion and sound* in order to explore these dynamic user-adapted control spaces through gestures (Axis 3).

Axis 1. Dynamic user-adapted control spaces through latent space regularization

Latent-based generative models can provide high-level features representations embedded in an organized latent space [13]. However, this latent representation is not unique [16]. Indeed, it highly depends on inductive biases defined by both the model and the data. Hence, we aim to propose novel methods to organize the latent space by incorporating the inductive bias of the user's choices. Our goal is to obtain a user-adapted disentangled representation with explicit control features for deep audio synthesis. We aim to embed the temporal aspects of user inputs, by allowing to dynamically change the *personalized* control dimensions. This would allow to provide dynamic custom control spaces, which evolve by simply transforming the latent spaces of the same pre-trained model.

To do so, we will rely on the RAVE model [4] recently developed by one of the supervising team (Philippe Esling, ACIDS), which provides real-time high-quality audio synthesis. We will focus on latent regularization techniques relying on both *implicit* (data-driven) and *explicit* (user-driven) knowledge. Our idea is to first infer most of the data-driven disentangled representation by regularizing on *known audio descriptors*, providing the principal latent dimensions. Then, we will rely on the remaining latent dimensions to encode the perspective of the user input. To define the *implicit* latent transformations, we will rely on Normalizing Flows (NFs) [25], which enable to transform highly complex distributions, while retaining efficient sampling and density estimation. Thus, we can define an invertible mapping to transform the *audio features latent* space into a *disentangled user-adapted control* space. Here, the idea would be to define *partial target distributions* to transform the original latent space on a subset of dimensions. This would provide users with time-evolving control parameters adapted to one's intent following smooth distributions on pre-trained latent-based models [27]. Inspired by the work of [17], we will also explore Riemannian geometry in order to better reflect the *explicit* aspects of the user input. Considering the custom control space as a Riemannian manifold [2] will allow us to learn specific metrics, which describe the proximity of latent data points depending on the user's perspective. Indeed, manifold-informed methods can be understood as a geometric prior that encodes user assumptions about the data and imposes an inductive bias [17]. Therefore, we could obtain smooth latent trajectories adapted to the user's exploration strategies. Additionally, we can rely on external knowledge to impose the orthogonalization of the *custom* latent dimensions through user relevance-feedback. Combining our previous latent-based transformation model with *deep reinforcement learning* paradigm would enable to shape the latent space through direct user interaction. Our idea would be to directly learn a user-centered reward function in order to dynamically control and explore parameters spaces for digital sounds creation [26].

Axis 2. Adaptive representation learning in low-data regimes

Most of the existing methods for learning control over generative models require massive sets of paired data with labelled attributes. However, gathering high-quality annotated samples necessitates manual efforts and expert knowledge. Additionally, existing data for real-life control problems are usually scarce and subject to human biases. Musical applications especially need to rely on very limited datasets as they have to closely match a given artistic intent. Hence, the aim of this axis is to endow generative models with adaptive latent representations that can match subjective controls with only a few examples.

To do so, we will rely on *few-shot* and *self-supervised learning* (SSL), where the data itself provides supervision. In recent years, these techniques have received increasing attention as an alternative to supervised learning, especially in the field of music information retrieval. Hence, we will benefit from the research conducted in one of the supervising team (Geoffroy Peeters, LTCI), in order to leverage SSL in low-data regimes. Here, we aim to introduce controllable priors over time-evolving latent representations. We will consider small sets of examples from different usage to help the model understand and distinguish custom latent dimensions based on examples interactively annotated by the user. We aim to use these small sets to learn the user's perspective as a similarity metric. Our major hypothesis is that we can bypass the need of massive *paired* data sets by relying on the concept *matching priors*, where we learn an invertible transform between the generative (latent) prior over the audio feature, and the *control* prior defined by the user behavior [20]. This would also entail the possibility for *prior-guided inference*, where the user actions are complemented by inference from the prior in the latent space, leading to a true co-creative behavior. Eventually, as the learning signal will be sparse and real-time, we need to define efficient inference processes aware of the data characteristics and the evolutionary dynamics of the latent space. To do so, we could rely on *continual learning* with latent-based inference process methods to learn latent trajectories the user is likely to perform and update these predictions while performing. This would further improve the adaptive process to match closely the user's intent, even with real-time constraints.

Axis 3. Co-Creative Music Applications: leveraging a gestural approach to explore the user-adapted control spaces

In order to promote intuitive control, we aim to leverage the inherent dynamics of gestures for enabling both expert and non-expert users to explore the previously-defined control spaces for musical creation. As described in [9], it is essential for users to personalize the movement-sound relationships depending on both the context and individual needs. Hence, we aim to study the temporal dynamics of gestures to explore our proposed adaptive latent representations. We will analyze how the proposed methods provide a custom unsupervised mapping between user movements and time-evolving audio control attributes.

To do so, we will pursue the user-centered methodologies previously developed in one of the supervising team (IRCAM - ISMM team). To bind motion to sound, we will take inspiration from the CLIP model [23], and adopt a multimodal approach relying on SSL techniques to embed sound and motion features into a joint latent representation [15] and implicitly map their temporal dynamics. As in [9], we will collect small sets of multimodal sound-motion recordings that match closely a given artistic intention and leverage the methods described in Axis 1 and 2 to dynamically adapt the latent representations to the user's perspective in low-data regimes. This thesis will benefit from strong collaborations with composers and musicians that already exist at IRCAM around gesture control of interactive sound environments. This thesis will lead to original musical pieces and performances that will be created using our prototypes, enhancing the visibility of this work through general public workshops and concerts.

3 Relevance to the call themes

This PhD research is proposed under the program *Ile-de-France DIM AI4IDF*, as it perfectly fits the sub-call *Axe 4.3: L'IA dans la vie de l'humain: exemple de la création*. Indeed, this project aims to address the lack of control over deep audio generative models and overcome the limitations of existing approaches to promote co-creative human-machine interaction in musical applications. Therefore, it will both rely and develop on recent advanced deep machine learning questions (deep generative models, self-supervised few-shot learning), while promoting the creative use of the proposed methods.

Although the development of deep models appears to have reached an unprecedented pace, their use still seems to be restricted to expert (researcher) users. Especially in the field of creative applications, the use of deep models is still out of reach for the general public, and even most of the musicians and composers. Therefore, it is required to propose simplified interaction and control techniques that could allow to facilitate the inclusion of these methods inside existing creative workflows. Hence, this project will have a strong societal impact, as it aims to help democratize the use of deep generative models, especially for musical creation. This thesis will also lead to strong collaborations with composers and musicians at IRCAM, in order to provide musical pieces and performances that could enhance the visibility of this work through general public concerts. We believe that this work will also find applications in other domains in needs of a personalized control approach such as the medical field with physical rehabilitation systems or augmented virtual reality requiring custom environment interactions. Finally, this thesis also highly promotes international mobility, as it is supported by a signed hosting agreement with the University of Tokyo (Tatsuya Harada, MIL), which would allow to strengthen the collaboration between the AI institutes (SCAI, Hi!Paris) and our partners in Japan in the field of theoretical machine learning.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021.

[2] Georgios Arvanitidis, Søren Hauberg, and Bernhard Schölkopf. Geometrically enriched latent spaces. *arXiv preprint arXiv:2008.00565*, 2020.

[3] Adrien Bitton, Philippe Esling, and Axel Chemla-Romeu-Santos. Modulated variational auto-encoders for many-to-many musical timbre transfer. *arXiv preprint arXiv:1810.00222*, 2018.

[4] Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021.

[5] Baptiste Caramiaux, Jules Françoise, Wanyu Liu, Téo Sanchez, and Frédéric Bevilacqua. Machine learning approaches for motor learning: A short review. *Frontiers in Computer Science*, 2, 2020.

[6] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[7] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

[8] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

[9] Jules Françoise and Frederic Bevilacqua. Motion-sound mapping through interaction: An approach to user-centered design of auditory feedback using machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–30, 2018.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, and Marc' Aurelio Ranzato. Fader networks:manipulating images by sliding attributes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[15] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabeleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. Audio self-supervised learning: A survey. *arXiv preprint arXiv:2203.01205*, 2022.

[16] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019.

[17] Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *Advances in Neural Information Processing Systems*, 33:2503–2515, 2020.

[18] Jon McCormack, Patrick Hutchings, Toby Gifford, Matthew Yee-King, Maria Teresa Llano, and Mark D'inverno. Design considerations for real-time collaboration with creative artificial intelligence. *Organised Sound*, 25(1):41–52, 2020.

[19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[20] Trung Ngo, Ville Hautamäki, and Merja Heinäniemi. The transitive information theory and its application to deep generative models. *arXiv preprint arXiv:2203.05074*, 2022.

[21] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[22] Ashis Pati and Alexander Lerch. Is disentanglement enough? on latent representations for controllable music generation. *arXiv preprint arXiv:2108.01450*, 2021.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[25] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[26] Hugo Scurto, Bavo Van Kerrebroeck, Baptiste Caramiaux, and Frédéric Bevilacqua. Designing Deep Reinforcement Learning for Human Parameter Exploration. *ACM Transactions on Computer-Human Interaction*, 28(1):1–35, January 2021.

[27] Maciej Wołczyk, Magdalena Proszewska, Łukasz Maziarka, Maciej Zięba, Patryk Wielopolski, Rafał Kurczab, and Marek Śmieja. Plugen: Multi-label conditional generation from pre-trained models. *arXiv preprint arXiv:2109.09011*, 2021.

[28] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Supervision context

The PhD will be co-hosted by IRCAM (Sorbonne, CNRS UMR 9912 - member of SCAI) and Telecom ParisTech (Institut Polytechnique de Paris - member of Hi!Paris). Hence, the proposed PhD is deeply linked with current research objectives conducted within the different teams, as it seeks to develop new interactive and user-based tools for musical creation. Thus, as this PhD is focused around developing deep learning models that provide musicians with creative control over the generation, it is directly in continuity and will fully leverage previous researches conducted by the teams at both IRCAM and Telecom ParisTech, but also their extensive experience and network of users. Indeed, this thesis will be co-supervised by the following team of supervisors, with their respective field of specialty

PHILIPE ESLING (SCAI) - *Associate professor at IRCAM, Sorbonne Université*
Head of the Artificial Intelligence and Data Science (ACIDS) team at IRCAM, which specializes in deep generative models applied to audio synthesis. Since 2013, he has been an associate professor at IRCAM and Sorbonne University. During this period, he has written more than 20 articles in prestigious journals (ACM CSUR, PNAS, NAR, IEEE TSALP) and more than 40 conference papers and has received the AFIM young researcher award (2011), the CG94 1st prize doctoral award (2013) and several best paper awards. He also developed the first computer-assisted orchestration software, Orchids, used in musical pieces by renowned composers performed at international concerts. He is the principal researcher in machine learning applied to music generation at IRCAM. In this project, he will provide expertise in deep generative models applied to audio synthesis.

FREDERIC BEVILLACQUA (SCAI) - *Research director at IRCAM, CNRS*
Head of the Sound Music Movement Interaction team at IRCAM in Paris. His research concerns the modeling and the design of interaction between movement and sound, and the development of gesture-based interactive systems. He co-authored more than 150 scientific publications and co-authored 5 patents. He was keynote or invited speaker at several international conferences such as the ACM TEI'13. He was the co-chair of the International conference New Interfaces for Musical Expression, and he co-created the International Conference on Movement and Computing (since 2014). As the coordinator of the "Interlude project", he was awarded in 2011 the 1st Prize of the Guthman Musical Instrument Competition (Georgia Tech) and received the award "prix ANR du Numérique" from the National Research Agency (category Societal Impact, 2013). In this project, he will provide expertise in human-computer interaction and gestural control applied to musical creation.

GEOFFROY PEETERS (HI!PARIS) - *Professor at Telecom ParisTech, Institut Polytechnique de Paris*
Full professor in the Image-Data-Signal (IDS) department of Télécom Paris, Institut Polytechnique de Paris. Before that (from 2001 to 2018), he was Senior Researcher (Directeur de recherche) at IRCAM leading research related to Music Information Retrieval. His research topics concern signal processing and machine learning (including deep learning) for the processing of audio with a strong focus on music and environnmental sounds. He has participated in (or coordinated) many national or European projects, published numerous articles (h-index of 37) and several patents in these areas and co-author of the ISO MPEG-7 audio standard. He has been co-general-chair of the DAFx-2011 and ISMIR-2018 conferences and co-program-chair of the ISMIR-2019 conference. He is the current president-elected of the Internatonal Society of Music Information Retrieval. In this project, he will provide expertise in self-supervised deep learning applied to musical information and few-shot learning.

This supervision team is supported by a long-standing research relationship between all members (Geoffroy Peeters being a former researcher of IRCAM), but this PhD is also the opportunity to strengthen the collaboration between our respective institutes. Finally, this PhD also leverages international mobility, with an already agreed opportunity of hosting by Tatsuya Harada, which is head of the MIL (Machine Intelligence Laboratory - University of Tokyo), providing an expertise in theoretical aspects of machine learning, through an existing collaboration. Therefore, the PhD ideally sits at the crossroads of diverse experiences of the supervisors and will fully leverage both the current momentum generated by the thesis directors through various funded projects and the international opening offered by the Japanese MIL.

Doctoral school

This project will be hosted by the EDITE (Ecole Doctorale Informatique, Télécommunications et Electronique) doctoral school (ED130) at Sorbonne Université.