October 2024

EXPLORING AI APPLIED TO AUDIO AND MUSIC

Matthieu Cervera
matthieu.cervera@hotmail.com

Introduction

This document has a dual purpose. First, it aims to provide a concise overview of how Artificial Intelligence (AI) is recently evolving in the field of audio and music. Additionally, it intends to explore potential research directions based on the current state of the domain and my own perspectives and interests. My greatest interest in AI is generative AI as I find it theoretically really fascinating and it is obviously mainly used in creative domains such as music. That's why I will focus mainly on generative AI in this document.

Generative models for music took a great stride in the last years. Starting from generating short audio musical segments and symbolic music such as MIDI, models are now capable of generating longer, high-quality, realistic audio signals. However we are still in the early stages, especially when it comes to the use of these models by artists. There is much work ahead in fostering effective human-AI collaboration. I believe the most interesting, effective and ethical approach is to focus on how AI models could best be integrated into artists' workflows. The central question we should keep in mind is : **As an artist, composer, producer, performer, how can AI help me and improve my workflow and creativity ?**

Context

To achieve audio and music generation, the state-of-the-art (SOTA) models are mostly driven by autoregressive-based models, non-autoregressive models introducing masking strategies, and diffusion-based models. Autoregressive models allow coherent and high-fidelity sounds but have a high computational cost during training and inference [1]. Some models are thus operating on quantized latent space in order to address this issue. For non-autoregressive models, masked generative transformers are often used to diminish computational costs [2] [3]. Diffusion-based generative models show controllable and high quality outputs [4]. For computational efficiency, diffusion models often operate on latent spaces too [5] [6]. An approach to improve the output of these models is to introduce conditioning : it allows more control over the generated outputs. The recent progress on generated audio is largely due to the development of Large Language Models (LLMs) and the integration of LLMs in audio generation ("text-to-music" models). I think large text-conditioned generative models (such as MusicLM [7] or MusicGen [8]) are a great breakthrough in the domain and I find their contribution crucial. Text-to-music generation is a fascinating tool for creators as it becomes easier to control the generated output. We can also note relevant other methods using parametric or audio conditioning. The most effective models often allow to combine several conditioning methods [9] [10] [11] [12].

Ideas

Text-conditioned models face several limitations. First, music has specific rules : musicality and musical coherence are really difficult challenges to address. This leads to an overall quality of the output that can be improved (e.g.generic LLMs can have insufficient integration of music theory or knowledge). Then, I believe the major issue of these models is that, as it stands, they can't be integrated into artists' creation process. Indeed the limited explicit controls over the generated result, the inadequate features that fail to capture the artistic intent of musicians or the significant computational and data needs are among the challenges worth exploring.

Improving text conditioning: To face those challenges, we might initially consider improving the text conditioning. For example, we could finetune the LLM on musical theory as in [13] or [14] to allow artists to better describe what they want with high-level features (BPM, chords...). Finetuned and improved LLM captions can also enable the model to interpret complex audio signals more accurately for other Music Information Retrieval (MIR) related tasks.

DAW integration: To integrate such models more widely into artists' workflows, we could create Digital Audio Workstation (DAW) plug-ins that would focus on simple actions : generate a synthesizer sound, a sound FX, create a reverberation effect, etc... Composer's Assistant 2 [15] is a good example of a tool integrated in a DAW. It allows to fill up MIDI patterns via an interactive control.

Physical parameters and audio conditioning: As using only a textual input could be restricting, the models could also be conditioned with physical Digital Signal Processing parameters or concepts. These would be chosen by producers either as real physical concepts, or assigned to more abstract concepts (color, timber...) that the model could easily describe [16]. If the model can also be audio-conditioned it could of course allow style transfer of a sound or sample [17]. This could be an interesting tool for producers and sound designers. We can also think of having a control over the chords or the melody like [18]. Sticking with the same idea of multi-conditioned models, earlier this year, I discussed with *Gael Richard* at *Telecom Paris* about a model that could learn notions of acoustics and 3D rendering to guide a generative model in the context of reverberation style transfer. This idea can be generalized for different audio effects [19].

Music theory conditioning: Using real music theory (not only through a text model) to improve the quality and the musicality of the generated audio could also be a good idea. In [20], the notation software could help diminish computation needs and improve the learning of symbolic music structure. Furthermore, investigating the similarity in music [21] [22] could help guide the generative model into a person's tastes and musical influences. That could serve musicians (e.g. to influence plug-in and model outputs) or listeners (e.g. to improve music discovery and classification).

Image Conditioning: Finally, an approach could also be to condition a model with images (using efficient SOTA models) [23]. Given the high proportion of music for the image in the world, it could be a useful tool for composers. Even if the composition isn't intended to be used for other media, the musicians, performers and composers often have mental images that characterize the music they are playing or composing.

Further considerations

The ideas outlined above naturally involve re-using, merging and optimizing existing models and techniques. There are too many different techniques to list them all here. For instance, methods like quantization are crucial for reducing computational costs, making them highly important to look over. Obviously, it is also interesting to try tackling other signal processing and acoustic related subjects such as source separation. Besides, source separation investigation could also fit in well with the ideas mentioned above [24] since it can be useful to provide more control to users. Please note also that the references provided are my selection of worthy examples among many other really interesting SOTA papers and this document doesn't aim to contain an exhaustive literature review.

References

[1] 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, 2016.

[2] Marco Comunità, Zhi Zhong, Akira Takahashi, Shiqi Yang, Mengjie Zhao, Koichi Saito, Yukara Ikemiya, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. Specmaskgit: Masked generative modeling of audio spectrograms for efficient audio synthesis and beyond, 2024.

[3] Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer, 2024.

[4] Simon Rouard and Gaëtan Hadjeres. Crash: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis, 2021.

[5] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models, 2023.

[6] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion, 2024.

[7] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023.

[8] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024.

[9] Simon Rouard, Yossi Adi, Jade Copet, Axel Roebel, and Alexandre Défossez. Audio conditioning for music generation via discrete bottleneck features, 2024.

[10] Shahan Nercessian, Johannes Imort, Ninon Devis, and Frederik Blang. Generating sample-based musical instruments using neural audio codec language models, 2024.

[11] Javier Nistal, Marco Pasini, Cyran Aouameur, Maarten Grachten, and Stefan Lattner. Diff-a-riff: Musical accompaniment co-creation via latent diffusion models, 2024.

[12] Fang-Duo Tsai, Shih-Lun Wu, Haven Kim, Bo-Yu Chen, Hao-Chung Cheng, and Yi-Hsuan Yang. Audio prompt adapter: Unleashing music editing abilities for text-to-music with lightweight finetuning, 2024.

[13] Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation, 2024.

[14] Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, Yizhi Li, Yinghao Ma, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenwu Wang, Guangyu Xia, Wei Xue, and Yike Guo. Composerx: Multi-agent symbolic music composition with llms, 2024.

[15] Martin E. Malandro. Composer's assistant 2: Interactive multi-track midi infilling with fine-grained user control, 2024.

[16] Liwei Lin, Gus Xia, Junyan Jiang, and Yixiao Zhang. Content-based controls for music large language modeling, 2024.

[17] Nils Demerlé, Philippe Esling, Guillaume Doras, and David Genova. Combining audio control and style transfer using latent diffusion, 2024.

[18] Or Tal, Alon Ziv, Itai Gat, Felix Kreuk, and Yossi Adi. Joint audio and symbolic conditioning for temporally controlled text-to-music generation, 2024.

[19] C. J. Steinmetz, S. Singh, M. Comunità, I. Ibnyhahya, S. Yuan, E. Benetos, and J. D. Reiss. ST-ITO: Controlling audio effects for style transfer with inference-time optimization. Presented at the 25th International Society for Music Information Retrieval Conference (ISMIR), 2024.

[20] Stephen Ni-Hahn, Weihan Xu, Jerry Yin, Rico Zhu, Simon Mak, Yue Jiang, and Cynthia Rudin. A new dataset, notation software, and representation for computational schenkerian analysis, 2024.

[21] Julia Barnett, Hugo Flores Garcia, and Bryan Pardo. Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model, 2024.

[22] Roser Batlle-Roca, Wei-Hisang Liao, Xavier Serra, Yuki Mitsufuji, and Emilia Gómez. Towards assessing data replication in music generation with music similarity metrics on raw audio, 2024.

[23] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models, 2023.

[24] Karn N. Watcharasupat and Alexander Lerch. A stem-agnostic single-decoder system for music source separation beyond four stems, 2024.