Generative AI for Music and Audio

Hao-Wen Dong

My research aims to **empower music and audio creation with machine learning**. My long-term goal is to **lower the barrier of entry for music composition and democratize audio content creation**. I am broadly interested in music generation, audio synthesis, generative AI, multimodal learning and music information retrieval.

Generative AI has been transforming the way we interact with technology and consume content. The recent success of large language model-based chatbots (e.g., OpenAI's ChatGPT and Google's Bard), AI assistants (e.g., GitHub and Microsoft Copilot) and text-to-image generation systems (e.g., Adobe Firefly, Midjourney and Stable Diffusion) showcases how AI-powered technology can be integrated into professional workflows and boost human productivity. In the next decade, generative AI technology will also reshape how we create audio content in the $2.3 trillion global entertainment industry, including the music, film, TV, podcast and gaming sectors. Take AI-powered music creation for example: On one hand, we have witnessed major progress in automatic music composition, which has long been considered as a grand challenge of AI. On the other hand, our expectations of *AI Music* today has expanded to cover the whole music creation process—from composition, arrangement, sound production, recording to mixing. With a growing momentum in both academia and industry, AI-powered audio creation has been gaining attention in the broader AI community.

My research springs from two fundamental questions: 1) *How can AI help professionals or amateurs create music and audio content?* 2) *Can AI learn to create music in a way similar to how humans learn music?* From a musical perspective, technology has always been a driving factor of music evolution. For example, the study of acoustics and musical instrument making fostered the development of classical music; the invention of synthesizers and drum machines helped popularize electronic music. I am thus interested in exploring how the latest AI technology can empower artists to create novel contents. From a technical perspective, music possesses a unique complexity in that music follows rules and patterns while being creative and expressive at the same time. I am thus fascinated about the idea of building intelligent systems that can learn, create and play music like humans do. I envision the future development of AI Music to be a two-way process—*new technology creates new music; new music inspires new technology*.

Motivated by this belief, I study a wide range of topics centered around **Generative AI for Music and Audio**, including multitrack music generation [1–6], automatic instrumentation [7], automatic arrangement [1, 6], automatic harmonization [8], music performance synthesis [9], text-queried sound separation [10], text-to-audio synthesis [11, 12] and symbolic music processing software [13, 14]. My research can be categorized into three main pillars: *multitrack music generation*, *assistive music creation tools* and *multimodal learning for audio and music*.

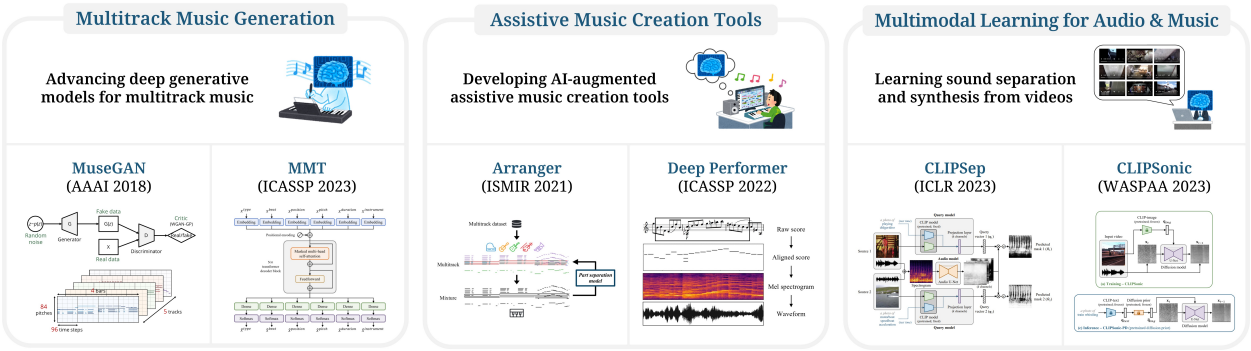

Figure 1: An overview of my research

Past Research

Advancing Deep Generative Models for Multitrack Music

Researchers have been working on automatic music composition for decades, and it has long been viewed as a grand challenge of AI. I started this thread of research on multitrack music generation in 2017. Back then, prior work on deep learning-based music generation had focused on generating melodies, lead sheets (i.e., melodies and chords) or four-part chorales. However, modern pop music often consists of multiple instruments or tracks. To make deep learning technology applicable in modern music production workflow, it is important to modernize deep learning models for multitrack music generation.

Figure 2: The AWS DeepComposer was debuted at AWS re:Invent 2019, featuring my work on multitrack music generation [1] as one of the backbone models.

I have done significant work on advancing deep generative models for multitrack music [1–6]. A representative example is my work on multitrack music generation, which was the *first deep neural network that can generate multitrack, multi-pitch music from scratch* [1, 2]. In this work, I investigated generating five-track pop music excerpts using generative adversarial networks (GANs). I showed that convolutional GANs are well-suited for modeling music using the piano roll representation. Moreover, I proposed a novel mechanism that allows the user to control the characteristics for each generated track on top of the overall music style for improved controllability. To train the proposed MuseGAN model, I compiled a dataset of more than twenty thousand pop songs, which has been used in many follow-up papers. Notably, **MuseGAN was featured as one of the backbone models implemented in the AWS DeepComposer, an AI-powered keyboard made and sold by Amazon**.[1,2]

Witnessing the lack of infrastructure when conducting these research projects, I have also made significant contributions to consolidate the infrastructure of music generation research [13, 14]. The Python toolkits I developed to process symbolic music for machine learning applications have been widely used in the field. With these libraries, researchers can easily download commonly used datasets programmatically and be liberated from reimplementing tedious data processing routines. The toolkits also allowed me to conduct *the first large-scale experiment that measures the cross-dataset generalizability of deep neural networks for music* [14].

Developing AI-augmented Assistive Music Creation Tools

Music creation today is still largely limited to professional musicians for it requires a certain level of knowledge in music theory, music notation and music production tools. Apart from generating new music content from scratch, another line of my research focuses on developing AI-augmented tools to assist amateurs to create and perform music. My long-term goal along this research direction is to lower the entry barrier of music composition and make music creation accessible to everyone.

In this direction, I study automatic instrumentation [7], automatic arrangement [1, 6], automatic harmonization [8] and music performance synthesis [9]. For example, in [7], I developed *the first deep learning model for automatic instrumentation*. Instrumentation refers to the process where a musician arranges a solo piece for a certain ensemble such as a string quartet or a rock band. This can be challenging for amateur composers as it requires domain knowledge of each target instrument. In

[1] https://www.amazon.com/dp/B07YGZ4V5B/
[2] https://aws.amazon.com/blogs/aws/aws-deepcomposer-now-generally-available-with-new-features/

this work, I proposed a new machine learning model that can produce convincing instrumentation for a solo piece by framing this problem as a sequential multi-class classification problem. Such an automatic instrumentation system can suggest potential instrumentation for amateur composers, especially useful when arranging for an unfamiliar ensemble. Further, the proposed model can empower a musician to play multiple instruments on a single keyboard at the same time.

Another example is my work on music performance synthesis [9]. While synthesizers play a critical role and are intensively used in modern music production, existing synthesizers either requires an input with expressive timing or allows only monophonic inputs. In light of the similarities between text-to-speech (TTS) and score-to-audio synthesis, I showed in this work that we can adapt a state-of-the-art TTS model for music performance synthesis. Moreover, I proposed a novel mechanism to enable polyphonic music synthesis. This work represents *the first deep learning based polyphonic synthesizer that can synthesize a score into a natural, expressive performance*.

Learning Sound Separation and Synthesis from Videos

The third line of my research focuses on multimodal learning for audio and music. Sound is an integral part of movies, dramas, documentaries, podcasts, games, short videos and audiobooks. In these media, audio and music production tools need to interact with inputs from other modalities such as text and images, and thus multimodal models are critical in enabling controllable creation tools for music and audio in these applications.

Along this direction, I have worked on text-queried sound separation [10] and text-to-audio synthesis [11, 12]. Unlike existing work that relies on a large amount of paired audio-text data, I explore a new direction of approaching bimodal learning for text and audio through leveraging the visual modality as a bridge. The key idea behind my study is to combine the naturally-occurring audio-visual correspondence in videos and the multimodal representation learned by contrastive language-vision pretraining (CLIP). Based on this idea, I developed *the first text-queried sound separation model that can be trained without any text-audio pairs* [10]. Text-queried sound separation aims to separate a specific sound out from a mixture of sounds given a text query, which has many downstream applications in audio post-production such as editing and remixing. I showed that the proposed model can successfully learn text-queried sound separation using only noisy unlabeled videos, and it even achieves competitive performance against a supervised model in some settings. Moreover, I built *the first text-to-audio synthesis model that requires no text-audio pairs during training* [11, 12]. The proposed model learns to synthesize audio given text queries, which can find applications in video and audio editing software. One of the key benefits of the approach studied in my work lies in its scalability to large video datasets in the wild as we only need unlabeled videos for training.

Future Directions

Multimodal generative AI with music and audio. Multimodal content generation has quickly become the next frontier of generative AI. Many recently-released large pretrained multimodal contrastive models lay the foundations for exciting creative applications to film, video and audiobook generation. I am particularly interested in working on multimodal generative models for background music and sound effect generation for videos, audiobooks and games. I would also like to explore fusing multiple controlling signals from different modalities (e.g., text, image, video, audio, emotion measurements, etc.) for controllable music and audio generation. My long-term goal along this direction is to *develop next-generation interfaces for music and audio editing equipped with intuitive multimodal controls*. I will seek collaborations with other faculty members in computer vision and natural language processing to pursue research along this direction.

Interactive AI tools for music and audio production. While recent deep learning-based music and audio generation systems can create short, plausible music excerpts, they offer limited usability and controllability for humans to step in. Instead of building a fully-automated generation system, I want to *develop interactable music and audio production tools equipped with intermediate controls that humans can interact with*. For example, recently we have seen preliminary results on leveraging large language models (LLMs) for building a compositional, human-usable audio generation system [15]. Moreover, subject-driven personalization [16] and instruction-based editing [17] has lately been attracting attentions in the image generation community, and I would like to explore opportunities in these directions with an eye to integrate these tools into professional creative workflows in music and audio production software. I will seek collaborations with other faculty members in human-computer interaction to explore new creative interfaces for music and audio production.

Human-like machine leaning algorithms for music. Richard Feynman once said: "What I cannot create, I do not understand." This echos my motivations of pursuing music generation research—generation represents the highest-level of understanding. A long-term direction of my research is to *develop human-like machine learning algorithms that can learn to create music in a way similar to how humans learn music*. For example, existing data-driven approaches for music generation usually rely on *reading* a large collection of musical scores. Unlike machines, however, humans learn music mostly through listening and practicing music rather than reading scores over and over again. I am thus interested in exploring novel machine learning models that can learn symbolic music composition through listening to a large collection of musical audio data. Some recent work [18] has shown preliminary results towards this direction. In my view, music possesses a unique complexity that might lead to new breakthroughs in AI and contribute towards the long-lasting pursuit of artificial general intelligence.

Funding and Industrial Collaborations

My research has been funded by the IEEE SPS Scholarship, Taiwan Government Scholarship to Study Abroad, J. Yang Scholarship and UCSD ECE Department Fellowship, awarded over $100K USD in total. Part of my work [9–12] is published in collaboration with researchers at Amazon, Dolby and Sony during my internships. To support my research group, I will actively apply for NSF funding and seek industrial collaborations with tech companies, including Amazon, NVIDIA, Adobe, Dolby, Sony and Yamaha, with whom I have had close collaborations in the past.

Broader Impacts

I envision my research to be integrated into the audio content creation workflow for professional artists and amateurs. Through providing new tools and interfaces to make music, my research could lower the barrier for music composition and empower novices to create their own music. Moreover, it could provide content creators (e.g., TikTokers, YouTubers and Twitch streamers) with royalty-free materials to avoid unintended copyright infringement. My research could also find applications in music education and therapy, where creating personalized courses can be costly. Finally, we could gain insights into the future of human-AI music co-creation though the interactions between human and automatic music composition systems. I envision this to foster the discussions in human-AI relationships in other fields.

References

[1] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[2] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Demonstration of a Convolutional GAN Based Model for Generating Multi-track Piano-rolls," *ISMIR Late-Breaking Demos*, 2017.

[3] Hao-Wen Dong and Yi-Hsuan Yang, "Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation," *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[4] Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick, "Multitrack Music Transformer," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[5] Weihan Xu, Julian McAuley, Shlomo Dubnov, and Hao-Wen Dong, "Equipping Pretrained Unconditional Music Transformers with Instrument and Genre Controls," *IEEE Big Data Workshop on AI Music Generation (AIMG)*, 2023.

[6] Hao-Min Liu, Hao-Wen Dong, Wen-Yi Hsiao, and Yi-Hsuan Yang, "Lead sheet and Multi-track Piano-roll generation using MuseGAN," *GPU Technology Conference (GTC) Taiwan*, 2018.

[7] Hao-Wen Dong, Chris Donahue, Taylor Berg-Kirkpatrick, and Julian McAuley, "Towards Automatic Instrumentation by Learning to Separate Parts in Symbolic Multitrack Music," *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[8] Yin-Cheng Yeh, Wen-Yi Hsiao, Satoru Fukayama, Tetsuro Kitahara, Benjamin Genchel, Hao-Min Liu, Hao-Wen Dong, Yian Chen, Terence Leong, and Yi-Hsuan Yang, "Automatic Melody Harmonization with Triad Chords: A Comparative Study," *Journal of New Music Research (JNMR)*, 50(1):37–51, 2021.

[9] Hao-Wen Dong, Cong Zhou, Taylor Berg-Kirkpatrick, and Julian McAuley, "Deep Performer: Score-to-Audio Music Performance Synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[10] Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick, "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *International Conference on Learning Representations (ICLR)*, 2023.

[11] Hao-Wen Dong, Gunnar A. Sigurdsson, Chenyang Tao, Jiun-Yu Kao, Yu-Hsiang Lin, Anjali Narayan-Chen, Arpit Gupta, Tagyoung Chung, Jing Huang, Nanyun Peng, and Wenbo Zhao, "CLIPSynth: Learning Text-to-audio Synthesis from Videos using CLIP and Diffusion Models," *CVPR Workshop on Sight and Sound*, 2023.

[12] Hao-Wen Dong, Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya, Santiago Pascual, Joan Serrà, Taylor Berg-Kirkpatrick, and Julian McAuley, "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

[13] Hao-Wen Dong, Wen-Yi Hsiao, and Yi-Hsuan Yang, "Pypianoroll: Open Source Python Package for Handling Multitrack Pianoroll," *ISMIR Late-Breaking Demos*, 2018.

[14] Hao-Wen Dong, Ke Chen, Julian McAuley, and Taylor Berg-Kirkpatrick, "MusPy: A Toolkit for Symbolic Music Generation," *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[15] Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D. Plumbley, and Wenwu Wang, "WavJourney: Compositional Audio Creation with Large Language Models," *arXiv preprint arXiv:2307.14335*, 2023.

[16] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023.

[17] Tim Brooks, Aleksander Holynski, and Alexei A. Efros, "InstructPix2Pix: Learning to Follow Image Editing Instructions," *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023.

[18] Rodrigo Castellon, Chris Donahue, and Percy Liang, "Codified audio language modeling learns useful representations for music information retrieval," *International Society for Music Information Retrieval Conference (ISMIR)*, 2021.