A Tutorial on AI Music Composition

Xu Tan
Senior Researcher
Microsoft Research Asia
Beijing, China
xuta@microsoft.com

Xiaobing Li
Composer & Professor
Central Conservatory of Music
Beijing, China
lxbmusic@188.com

ABSTRACT

AI music composition is one of the most attractive and important topics in artificial intelligence, music, and multimedia. The typical tasks in AI music composition include melody generation, song writing, accompaniment generation, arrangement, performance generation, timbre rendering, sound generation, and singing voice synthesis, which cover different modalities (e.g., symbolic music score, sound) and well match to the theme of ACM Multimedia. As the rapid development of artificial intelligence techniques such as content creation and deep learning, AI based music composition has achieved rapid progress, but still encountered a lot of challenges. A thorough introduction and review on the basics, the research progress, as well as how to address the challenges in AI music composition are timely and necessary for a broad audience working on artificial intelligence, music, and multimedia. In this tutorial, we will first introduce the background of AI music composition, including music basics and deep learning techniques for music composition. Then we will introduce AI music composition from two perspectives: 1) key components, which include music score generation, music performance generation, and music sound generation; 2) advanced topics, which include music structure/form/style/emotion modeling, timbre synthesis/transfer/mixing, etc. At last, we will point out some research challenges and future directions in AI music composition. This tutorial can serve both academic researchers and industry practitioners working on AI music composition.

ACM Reference Format:
Xu Tan and Xiaobing Li. 2021. A Tutorial on AI Music Composition. In *Proceedings of the 29th ACM International Conference on Multimedia, Oct 20-24, 2021, Chengdu, China (ACM MM'21).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3474085.3478875

ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3478875

1 BIO

1.1 Xu Tan

Xu Tan is a Senior Researcher in Machine Learning Group, Microsoft Research Asia (MSRA)[1]. His research interests cover machine learning, deep learning, and their applications on natural language, speech and music processing, including music understanding and generation, singing voice synthesis, speech synthesis and recognition, neural machine translation, pre-training, etc. He has designed several popular language and speech models, such as MASS and FastSpeech, and has transferred many research works (e.g., HiFiSinger/XiaoiceSing, StructMelody, MASS, FastSpeech 1/2, LRSpeech, AdaSpeech) to the important products in Microsoft (e.g., Azure, Bing). The machine translation systems developed by him have achieved human parity on Chinese-English machine translation in 2018 and won several champions on WMT machine translation competition in 2019. He is the area chair of several top AI conferences (e.g., NeurIPS 2021, AAAI 2021).

Xu Tan has rich research experience in AI music, including song writing, accompaniment/arrangement generation, singing voice synthesis, and music understanding. He has designed several models for AI music composition, including: 1) SongMASS [29], a lyric-to-melody and melody-to-lyric generation system with pre-training and lyric-melody alignment modeling; 2) DeepRapper [32], a neural network based rap generation system with rhyme and rhythm modeling; 3) StructMelody, a melody generation system based on music structure information; 4) MusicBERT [34], a large-scale pre-trained model based on huge music data for music understanding; 5) PopMAG [26], a music accompaniment generation model with efficient encoding and long-term sequence modeling; and 6) HiFiSinger [4], DeepSinger [27], and XiaoiceSing [21], several neural singing voice synthesis systems.

1.2 Xiaobing Li

Xiaobing Li, composer, Professor of the Central Conservatory of Music (CCOM)[2] in China, and the head[3] of the Department of Music AI and Information Technology[4] in CCOM, Director of the Art and Artificial Intelligence Committee of the Chinese Association for Artificial Intelligence, and chief expert of major national social science projects. He is a researcher and advocator of "3D music", and an expert in electronic music, computer music, and music technology. He has integrated deep learning into artistic creation

[1] Some userful links of Xu Tan: 1) Homepage: https://www.microsoft.com/en-us/research/people/xuta/; 2) Google Scholar: https://scholar.google.com/citations?user=tob-U1oAAAAJ; 3) AI music project page: https://www.microsoft.com/en-us/research/project/ai-music/.
[2] http://en.ccom.edu.cn/2020/
[3] http://www.ccom.edu.cn/jxyx/ai/xyls/201909/t20190908_59340.html
[4] http://en.ccom.edu.cn/2020/endep/202001/t20200110_67073.html

and made outstanding contributions to music artificial intelligence and music information technology. His works cover almost all types of music, and some works are loved by general audience and have national-wide influence in China. He has won the Golden Bell Award, the Wenhua Award, the Wenhua Music Composition Award, the first prize of National Opera and Dance Drama, and the "Five One Project" of the Central Propaganda Department in China, and other domestic and foreign awards.

Xiaobing Li leads the Department of Music AI and Information Technology in CCOM, and guides the research on AI music. He has deep understanding in both music and AI technology, and has developed several high-quality AI music systems on song writing, accompaniment generation, and singing voice synthesis.

2 PREFERENCE FOR HALF- OR FULL-DAY EVENT

The tutorial will be a half-day event, with two breaks in the middle.

3 MOTIVATION OF THIS TUTORIAL

AI music composition is one of the most attractive and important topics in artificial intelligence, music, and multimedia. The typical tasks in AI music composition include melody generation, song writing, accompaniment generation, arrangement, performance generation, timbre rendering, sound generation, and singing voice synthesis, which cover different modalities (e.g., symbolic music score, sound) and well match to the theme of ACM Multimedia. As the rapid development of artificial intelligence techniques such as content generation and deep learning, AI based music composition has achieved rapid progress, but still encountered a lot of challenges. A thorough introduction and review on the basics, the research progress, as well as how to address the challenges in AI music composition are timely and necessary for a broad audience working on artificial intelligence, music, and multimedia.

4 COURSE DESCRIPTION

1. Background
   1.1 Music Basics [3, 23]
   1.2 Music Composition Pipeline [17]
   1.3 Deep Leaning Techniques for Music [3, 10]
2. Music Score Generation
   2.1 Song Writing
   2.1.1 Lyric Generation [11, 22, 32]
   2.1.2 Melody Generation [12, 14, 28]
   2.1.3 Melody-to-Lyric Generation [18, 30]
   2.1.4 Lyric-to-Melody Generation [2, 29, 33]
   2.2 Music Arrangement
   2.2.1 Multi-Track Music Generation [7, 20]
   2.2.2 Accompaniment Generation [26, 36]
3. Music Performance Generation [16, 24, 25]
4. Music Sound Generation
   4.1 Singing Voice Synthesis [4, 19, 21]
   4.2 Music Sound Generation [6, 9]
5. Advanced Topics in AI Music Composition
   5.1 Music Structure/Form Modeling [1, 31]
   5.2 Music Style/Emotion Modeling [34, 35]
   5.3 Transfer/Control in Music Generation [5, 15]
   5.4 Timber Synthesis/Sound Mixing [8, 13]
6. Challenges and Future Directions

5 ANTICIPATED TARGET AUDIENCE

This tutorial targets for those audiences who work on: 1) music composition, computer music, electronic music; 2) speech, audio, music, signal processing, and multimedia; 3) deep learning and artificial intelligence. The expected number of attendees is 1000+.

REFERENCES

[1] Richard Ashley. 2017. Musical Structure: Form. In *The Routledge Companion to Music Cognition*. Routledge, 179–190.

[2] Hangbo Bao, Shaohan Huang, Furu Wei, Lei Cui, Yu Wu, Chuanqi Tan, Songhao Piao, and Ming Zhou. 2019. Neural melody composition from lyrics. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 499–511.

[3] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. 2017. Deep learning techniques for music generation–a survey. *arXiv preprint arXiv:1709.01620* (2017).

[4] Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu. 2020. HiFiSinger: Towards High-Fidelity Neural Singing Voice Synthesis. *arXiv preprint arXiv:2009.01776* (2020).

[5] Shuqi Dai, Zheng Zhang, and Gus G Xia. 2018. Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841* (2018).

[6] Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis Bach. 2018. Sing: Symbol-to-instrument neural generator. *arXiv preprint arXiv:1810.09785* (2018).

[7] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[8] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. 2020. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1x1ma4tDr

[9] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*. PMLR, 1068–1077.

[10] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[11] Daniel Taesoo Lee Harrison Gill and Nick Marwell. [n.d.]. Deep Learning in Musical Lyric Generation: An LSTM-Based Approach. ([n. d.]).

[12] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer: Generating Music with Long-Term Structure. In *International Conference on Learning Representations*.

[13] Sicong Huang, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, and Roger B Grosse. 2018. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620* (2018).

[14] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop Music Transformer: Beat-based modeling and generation of expressive Pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1180–1188.

[15] Yun-Ning Hung, I Chiang, Yi-An Chen, Yi-Hsuan Yang, et al. 2019. Musical composition style transfer via disentangled timbre representations. *arXiv preprint arXiv:1905.13567* (2019).

[16] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, and Juhan Nam. 2019. Score and performance features for rendering expressive music performances. In *Proc. of Music Encoding Conf.*

[17] Shulei Ji, Jing Luo, and Xinyu Yang. 2020. A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions. *arXiv preprint arXiv:2011.06801* (2020).

[18] Hsin-Pei Lee, Jhih-Sheng Fang, and Wei-Yun Ma. 2019. icomposer: An automatic songwriting system for chinese popular music. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 84–88.

[19] Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, and Kyogu Lee. 2019. Adversarially Trained End-to-End Korean Singing Voice Synthesis System. *Proc. Interspeech 2019* (2019), 2588–2592.

[20] Xia Liang, Junmin Wu, and Jing Cao. 2019. MIDI-Sandwich2: RNN-based Hierarchical Multi-modal Fusion Generation VAE networks for multi-track symbolic music generation. *arXiv preprint arXiv:1909.03522* (2019).

[21] Peiling Lu, Jie Wu, Jian Luan, Xu Tan, and Li Zhou. 2020. XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System. *Proc. Interspeech 2020* (2020), 1306–1310.

[22] Xu Lu, Jie Wang, Bojin Zhuang, Shaojun Wang, and Jing Xiao. 2019. A syllable-structured, contextually-based conditionally generation of chinese lyrics. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 257–265.
[23] Peter Manning. 2013. *Electronic and computer music*. Oxford University Press.
[24] Eita Nakamura, Yasuyuki Saito, and Kazuyoshi Yoshii. 2020. Statistical learning and estimation of piano fingering. *Information Sciences* 517 (2020), 68–85.
[25] Sageev Oore, Ian Simon, Sander Dieleman, and Doug Eck. 2017. Learning to create piano performances. (2017).
[26] Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1198–1206.
[27] Yi Ren, Xu Tan, Tao Qin, Jian Luan, Zhou Zhao, and Tie-Yan Liu. 2020. Deepsinger: Singing voice synthesis with data mined from the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1979–1989.
[28] Adam Roberts, Jesse Engel, Colin Raffel, Ian Simon, and Curtis Hawthorne. 2018. MusicVAE: Creating a palette for musical scores with machine learning, March 2018.
[29] Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2020. SongMASS: Automatic Song Writing with Pre-training and Alignment Constraint. *arXiv preprint arXiv:2012.05168* (2020).
[30] Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. A melody-conditioned lyrics language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 163–172.
[31] Jian Wu, Xiaoguang Liu, Xiaolin Hu, and Jun Zhu. 2020. PopMNet: Generating structured pop music melodies using neural networks. *Artificial Intelligence* 286 (2020), 103303.
[32] Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L. Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. 2021. DeepRapper: Neural Rap Generation with Rhyme and Rhythm Modeling. In *ACL 2021*.
[33] Yi Yu, Abhishek Srivastava, and Simon Canales. 2021. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 1 (2021), 1–20.
[34] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training. In *ACL 2021*.
[35] Kun Zhao, Siqi Li, Juanjuan Cai, Hui Wang, and Jingling Wang. 2019. An emotional symbolic music generation system based on lstm networks. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE, 2039–2043.
[36] Hongyuan Zhu, Qi Liu, Nicholas Jing Yuan, Chuan Qin, Jiawei Li, Kun Zhang, Guang Zhou, Furu Wei, Yuanchun Xu, and Enhong Chen. 2018. Xiaoice band: A melody and arrangement generation framework for pop music. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2837–2846.