HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230097

Co-Performing Music with AI: Real-Time Performance Control Using Speech and Gestures

Ilya BOROVIK [a,1] and Vladimir VIRO [b,2]
[a] *Skolkovo Institute of Science and Technology, Moscow, Russia*
[b] *Peachnote GmbH, Munich, Germany*

Abstract. We present an interactive music AI system that enables users to co-create expressive performances of notated music using speech and gestures. The system provides multi-modal interactive dialog-based control of performance rendering via smartphones and is accessible to people regardless of their musical background. We train a deep learning music performance rendering model on sheet music and associated performances with notated performance directions and user-system interaction data. Users have the opportunity to actively participate in the performance process. A speech- and gesture-based feedback loop with interactive learning improve the accuracy of performance rendering control. We believe that many people can express aspects of music performance using natural human expressions such as speech, voice, and gestures, and that by hearing the music follow their communicated intent, they can achieve deeper immersion and enjoyment of music than otherwise possible. With this work we pursue the goal of developing novel, fulfilling, and accessible music making experiences for large numbers of people who are not currently musically active.

Keywords. expressive music performance, human-computer interaction, mobile interface, deep learning, interactive learning

1. Introduction

The development of artificial intelligence and deep learning models has led to the creation of a new paradigm of human-centered machine learning, which aims to improve user experience and enhance human capabilities in various domains [1,2,3]. Music performance is an art form that requires expertise, practice, and physical ability [4,5]. The traditional paradigm of music interpretation and performance, where the musician interprets a score and translates the intended expression into the control of the musical instrument, can be difficult for those without musical training. Artificial intelligence and machine learning offer new approaches to music creation and performance [6,7], and interactive music creation systems have become increasingly popular [8,9].

In this work, we present a system for interactive co-creation of expressive performances of notated music using speech and gestures, which provides multi-modal inter-

[1]Corresponding Author: ilya.borovik@skoltech.ru
[2]Corresponding Author: vladimir@peachnote.de

active dialog-based control of performance rendering via smartphones and is accessible to people regardless of their musical background. Our system is designed to allow people to actively participate in the performance process by using their natural human expressions such as speech, voice, and gestures to control the performance of existing music. By hearing the music follow their communicated intent, they can achieve a deeper immersion and enjoyment of music than otherwise possible [10,11].

To accomplish this, we draw inspiration from the practice of music conducting, in which musicians translate the score and the conductor's gestures, facial expressions, and speech direction into music performance [12,13]. We focus on performance rendering for notated music, which provides written performance direction markings that composers use to communicate aspects of intended articulation to the musicians. We use existing music performances to ground these labels in music performance practices and learn their interpretations through a deep learning model to intuitively control performances. Our system uses state-of-the-art transformers for modeling sequential musical patterns [14] and variational autoencoders [15,16] that operate at different levels of the music hierarchy to advance research in expressive music performance rendering. Building on recent advances in multi-modal representation learning [17,18,19], we link user expression data in multiple modalities with music performance features and provide real-time or near real-time interactive music co-performance.

This paper presents the ongoing development of a deep learning based system for the interactive co-creation of expressive music performances for written music with:

1. real-time interactive music performance rendering;
2. human expression as performance control modalities;
3. accessibility to people without professional musical training and background;
4. maximum accessibility through inexpensive smartphone devices.

Our goal is to provide a new kind of fulfilling, engaging, and accessible music making experience, allowing people to perform great musical works using natural human expression. Our main contributions:

1. We develop a method to interactively participate in and control music performance rendering in real-time using speech and facial expressions;
2. We employ transformer models for controllable expressive music performance rendering and build interpretable connections between learned performance embeddings and notated music performance directions;
3. We implement a mobile web interface with low hardware requirements for interactive co-creation of music performances.

The paper is organized as follows. In Section 2, we present related work on interactive and deep learning-based music performance creation systems and how our work relates to them. Then, in Section 3, we present the design of our system with a description of its main parts: the music performance model, the interaction backend, and the mobile web application. Section 4 describes the examples of interactivity implemented in the system. Finally, we outline future work in Section 5 and make a conclusion in Section 6.

2. Related Work

2.1. Human-Centered Machine Learning

Human-centered machine learning is an actively emerging research field that explores the methods of aligning machine learning systems with human needs to make humans more effective and efficient [1,2,3]. The applications include medicine [20], education [21], music [9], art [22], software development [23], interactive and assistive technologies [24]. Recent work on tuning large-scale language models to build human-friendly assistants shows the great potential of using deep learning as a human companion in everyday tasks [24]. Our primary goal is to make the music-making experience accessible to more people through computational models. We develop a deep learning model with the focus on interactive inference and fine-tuning based on user feedback.

2.2. Interactive Music Performance

Interactive music performance systems contribute to the emerging field of human-centered machine learning [2,9]. They introduce new instruments for musical expression [25,26,27] and interfaces for controlling a generative model [28,29,30]. Wekinator [25] is a user-friendly computer application that learns to map camera-scanned sample input, such as gestures or facial expressions, to specific music performance actions. Piano Genie [27] is a machine learning controller that allows non-musicians to improvise on the piano. CoCoCo [29] provides multi-example sampling with revision and AI steering tools to control the diversity and high-level directions of a generative model. COSMIC [30] provides a novel way to create music through a textual dialog system. Our work follows the Wekinator approach and focuses on expressive performance rendering for a fixed score and accessibility through a multi-modal mobile web interface.

2.3. Music Generation with Deep Learning

Deep learning is commonly used for music generation [8]. The most popular architectures are transformers for learning long-term sequential musical patterns [31,32,33] and variational autoencoders for unsupervised style encoding and control [34,35,36,37]. The models provide offline control over performance style [35,32] or performance parameters [36,37]. Recent works aim to generate music from descriptions [37] and text [38], as an intuitive way for humans to express themselves musically. Our work builds on these advances, focusing on real-time and interactive music performance generation that allows humans to participate directly in the creative process.

2.4. Expressive Music Performance Rendering

Expressive music performance models render performances for written musical scores [7,39], either through rule-based [40,41] or machine learning approaches [42,43,44]. The latter mainly consist of variational autoencoders [15] for performance style encoding and control, and recurrent neural networks [45] for expressive performance representation. Our music performance model uses a transformer architecture [14] to improve long-term music dependency modeling and maps learned style representations to human expressive control inputs. In addition, in contrast to offline control of performance generation, we aim for interactive controllable generation and model fine-tuning.

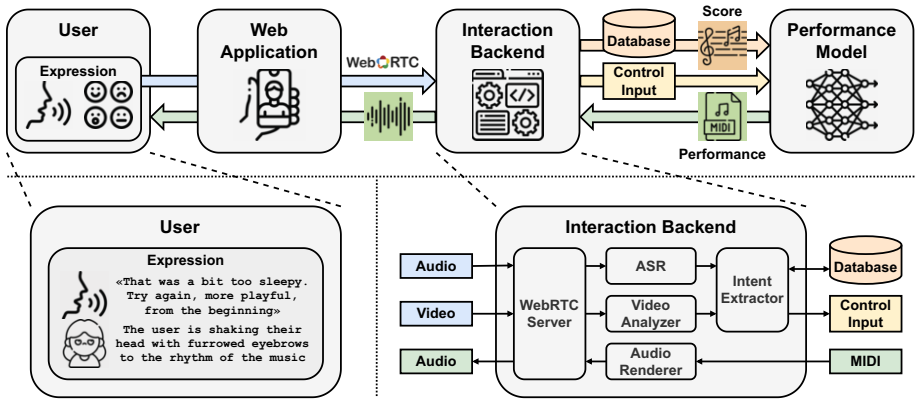

Figure 1. Interactive system for real-time co-creation of expressive music performances using speech and facial expressions. The user interacts with the music performance model through a mobile web application and interaction backend. The rendered music performance is played back to the user in the mobile application.

3. Interactive Music Performance Rendering System

We develop an interactive learning system for real-time co-creation of expressive music performances, shown in Figure 1. Its main components are:

1. Music Performance Model;
2. Interaction Backend;
3. Mobile Web Application.

The Music Performance Model enables the controllable creation of expressive performances for written music. The model learns from examples of human performance and is fine-tuned with user feedback to provide human-like musical expressiveness satisfying user requests. By automatically performing the written notes, the model removes the need for a user to play a musical instrument in order to perform a piece of music.

The Interaction Backend and Mobile Application connect users to the computational music performance model and allow interactive manipulation of the performance. By offering real-time performance rendering, we provide users with online interaction and immediate response. The speech and gestures allow users to express creative ideas using intuitive concepts such as text and emotion.

The following subsections describe the technical details behind the performance rendering model, interaction backend and mobile application.

3.1. Music Performance Model

Music Performance Model is a deep learning model for controllable expressive music performance rendering illustrated in Figure 2. It combines transformers [14] for sequential data modeling and variational autoencoders [15] for encoding performance style at different levels of the musical hierarchy. The building blocks are: performance encoder, performance decoder, and performance direction classifier.

Performance Encoder computes performance style representations at the note, beat, and bar levels. The transformer encoder [14] takes a sequence of score and performance features as input and outputs an embedding for each note. The embeddings are

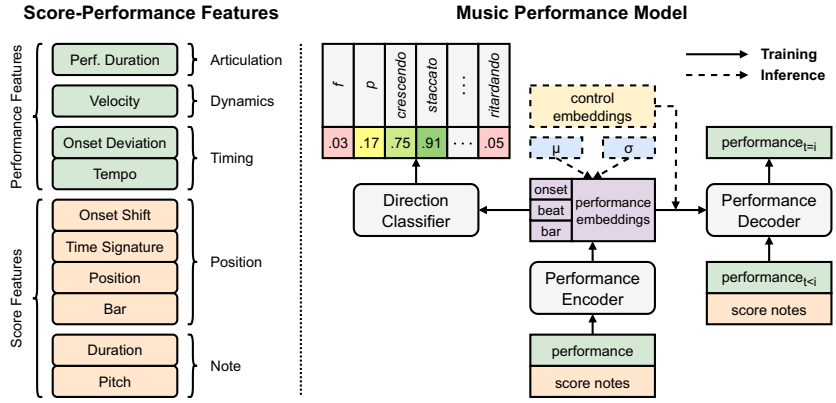

Figure 2. Music Performance Model. The performance encoder computes performance style embeddings on bar, beat, and onset levels. The performance decoder outputs performance features given score note features, past performance, and encoded style representations. The direction classifier associates the performance context with a set of performance direction labels.

averaged over bars, beats, and onsets (chords, or notes at the same position) and passed through a linear layer to compute latent bar-, beat-, and onset-level performance style embeddings, optimized using maximum mean discrepancy objective [16,46].

Performance Decoder operates with the score features (notes to play), the previous performance context (performance history), and the combined multi-level performance style embeddings computed by the Performance Encoder (style input). The decoder is a decoder transformer model autoregressively predicting the expressive performance features of the currently played note. The model is trained by maximizing the likelihood of the performance features.

Performance Direction Classifier learns an association between the performance embeddings and performance directions written in musical scores. It aims to provide an intuitive interpretation of the learned control space to be used to control the performance generation. The Direction Classifier classifies a local context of bar, beat and onset embeddings into performance direction classes:

- dynamic: degrees of *piano* and *forte*;
- dynamic changes: *crescendo* and *diminuendo*;
- tempo: *adagio*, *largo*, *presto*, etc.;
- tempo changes: *accelerando*, *ritardando*, *a tempo*, etc.;
- articulations: *legato*, *staccato*, *fermata*, etc.

The classifier predicts the likelihood of a direction being performed in a given performance context. Differences between embeddings with high and low likelihoods provide a direction for moving the generation toward a specified performance marking. We map these quantified per-direction embedding differences to natural language commands such as *"play more piano here"* or *"switch to largo"* to control performance rendering. This interaction is implemented in the interaction backend and application interface.

For model training, we preprocess the MIDI files from the ASAP dataset of aligned scores and piano performances [47]. The score features are note pitch, duration, bar, position in bar, time signature, and inter-onset position shift. The performance features are local performance tempo, note onset deviation, performed duration, and velocity.

During inference, the randomly sampled or modified existing performance embeddings can be used to generate and control music performances. Since the embedding space is optimized with the decoder performance generation objective, the latent space encodes features relevant for performance reconstruction. By connecting the classifier interpretation of the performance embedding with natural language instructions, we enable an intuitive controllable music performance rendering.

The initial model is trained isolated from the target user and generated music performances might be far from the user needs. To improve the quality of the model, we fine-tune it on the user-model interaction data. Currently, the active learning framework involves the offline periodic fine-tuning of the model on feedback scores. Given a set of performance-feedback score pairs, the performance decoder is optimized with an additional loss function maximizing the positive feedback per input performance sequence. In the future, we plan to implement fully interactive and online model updates.

3.2. Interaction Backend

Interaction Backend comprises multiple micro-services responsible for different tasks, such as handling the client connection, audio transcription, video analysis, audio rendering, etc. Services are implemented in different languages (Python, C++ and Go) and may be restarted, updated and rolled back independently. They all communicate via a messaging bus. A database stores scores, past performances, user feedback and performance directions which we use to optimize the performance model.

The JavaScript client connects to a multi-user WebRTC backend and establishes a bi-directional data channel and audio stream as well as a video stream from the client. Audio and video streams are analyzed in real-time. Audio is transcribed to text using Whisper [19], which is forwarded to GPT-3 [17] for intent extraction. The intent extractions works with input in multiple languages.

Currently, the system is sequencing and rendering MIDI performances to audio. In the future, we plan to generate audio directly through a deep learning model. The MIDI sequencer gets its cues from the gesture and intent recognition services and renders MIDI performances live. The MIDI stream is sent to the audio rendering node. Its audio output is sent back to the WebRTC server process that handles the client connection, and from there the music audio is sent back to the web interface and the user.

3.3. Mobile Web Application

Mobile Web Application[3] connects users to the interaction backend and the deep learning based music performance rendering system. The application requires a smartphone with a stable internet connection, camera and voice recorder. The web interface welcomes users and prompts them to turn on their camera and microphone to start interactive communication with the music performance model. Once the permissions are granted and the WebRTC connection is established, the user sees the camera image in the top half of the screen and the interaction button in the bottom half. The button allows the user to indicate that the system should pay attention to their input: audio, or video. The next section provides an overview of the mechanics of interacting with the system.

[3]Demo: https://d3dbzxyywswxzm.cloudfront.net.

4. Interaction with the System

Our application uses two primary intuitive interaction modalities which can be combined to create a highly expressive and dynamic musical performance:

1. **Speech**: the system analyzes the input audio stream and recognizes speech-specific phrases. The text embeddings of speech transcription are mapped to performance direction control embeddings as described in Section 3.1.
2. **Gestures**: the system processes the video stream and extracts facial expressions. The expression embeddings are mapped to performance direction classes and pre-defined user-system actions.

The web interface greets the user and prompts them to turn on their camera and microphone to begin interacting with the music performance model. Once permissions are granted, the user sees the camera image in the top half of the screen and the interaction button in the middle of the bottom half. Initially, the system selects a random musical composition from the database and begins rendering an arbitrary performance for this written music. The user can press the button and ask the system to do any of the following within a single phrase:

1. select a composition
   - *"Let's play Chopin's Mazurka in D major"*
2. pause or stop the performance
   - *"Please stop"*
3. navigate to a different place in the score
   - *"Let's play again from the beginning"*
4. provide feedback on the current performance
   - *"That was still a bit too slow and too much staccato"*
5. ask the system to play in a particular way
   - *"Could you play this like a mother singing a lullaby to her child?"*
6. show the system non-verbally how to play using facial gestures, for example making a blissful expression.

The interaction button relieves us of the need to continuously evaluate user input and judge whether it is intentional or accidental (the user does not intend to direct the performance, but moves or says something). While the button is pressed, the system continuously evaluates the video input and applies the analysis results to the performance. Audio input is not evaluated until the button is released. However, if a voice activity detector (VAD) detects speech, we immediately lower the volume of the performance to let the user know that we are listening and to make it easier to understand the speech.

The information we are looking for, such as navigation directions, feedback on past performance, and directions for future performance, is extracted from the transcribed speech using GPT-3. This allows us to successfully process free-form speech in multiple languages and provides great flexibility during development, at a cost in reliability and latency that we are currently willing to accept.

The interaction data and feedback are stored in the database for tuning the music performance model in subsequent iterations. Specifically, the backend stores the compressed video frame representations, the verbal commands and their embeddings, and the rendered performances. These features are then used to fine-tune the performance rendering model according to the desired input control as discussed in Section 3.1.

5. Future Work

As discussed throughout the paper, our main goal is to make musical expression accessible to people with no prior musical experience by designing simple interfaces and machine learning models. Currently, we support limited modalities for controlling music performance, namely speech and gesture. Our long-term goal is to integrate all forms of human expression used in musical contexts, such as conducting, teaching, and playing together, into our system. For example, using vocalization and full-body gestures to enhance interaction, allowing the user to modulate tempo, dynamics, and articulations.

Another important aspect of the human-computer interaction paradigm is personalization, the alignment of generated results with human requests and wishes. Understanding user intent, whether or not they provide control input, whether they are engaged in following the music rather than trying to direct its performance, is critical. We will study how different people describe and demonstrate music in relation to their expectations of the system's behavior. Our goal is to provide a personalized experience for each user, while collecting music descriptors that will help improve the system over time. Extensive human evaluation of the system is an important part of future research.

Regarding the technical solution, there are several points to consider. Currently, there are rare glitches in the rendered music performances coming from the trained deep learning model. We plan to solve them by collecting more data and using feedback to tune the models. Other shortcomings include support for piano music only and no control over the acoustic sound properties of the music, as the backend synthesizes rendered MIDI performances. We plan to perform music performance rendering in the audio domain using a deep learning model. Finally, we intend to improve the user interface to make it more user-friendly, robust, and inviting while keeping it simple. Various visualization options will complement and frame the musical functionality of the application.

6. Conclusion

In this work, we have presented a user-friendly interactive system for the co-creation of expressive music performances using speech and gestures. Through a mobile web application, it allows users to interact with an autonomous deep learning-based performance rendering model in real time. Our approach incorporates both verbal and non-verbal human expressive capabilities, allowing individuals to project emotions and affects through music using natural expressive language. This makes our system accessible to people without musical training or the ability to play musical instruments, and makes complex musical works more widely available for performance and interpretation.

We believe that this work will contribute to the fields of human-computer interaction, human-centered machine learning, and interactive music creation and performance, and will allow a greater number of people to experience the joy of musical expression. We hope that the system can be used in educational contexts to make musical tradition and practice more accessible, tangible, and engaging for young people. Since the web application does not require any setup on the part of the user, our system is easy to try out. If it produces interesting results right away, it has a chance of being used by many people. In future work, we will continue to incorporate different ways of interacting with the system to provide a complete, intuitive, and accessible musical experiences.

References

[1] Gillies M, Fiebrink R, Tanaka A, Garcia J, Bevilacqua F, Heloir A, et al. Human-Centred Machine Learning. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems; 2016. p. 3558-65. Available from: https://doi.org/10.1145/2851581.2856492.

[2] Kaluarachchi T, Reis A, Nanayakkara S. A Review of Recent Deep Learning Approaches in Human-Centered Machine Learning. Sensors. 2021;21(7):2514. Available from: https://www.mdpi.com/1424-8220/21/7/2514.

[3] Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review. 2022:1-50.

[4] Rink J. Musical Performance: A Guide to Understanding. Cambridge University Press; 2002.

[5] Palmer C. Music performance. Annual review of psychology. 1997;48(1):115-38.

[6] de Mantaras RL, Arcos JL. AI and Music: From Composition to Expressive Performance. AI Magazine. 2002 Sep;23(3):43. Available from: https://ojs.aaai.org/index.php/aimagazine/article/view/1656.

[7] Kirke A, Miranda ER. Guide to Computing for Expressive Music Performance. Springer; 2013. Available from: https://doi.org/10.1007/978-1-4471-4123-5.

[8] Hernandez-Olivan C, Hernandez-Olivan J, Beltran JR. A Survey on Artificial Intelligence for Music Generation: Agents, Domains and Perspectives. arXiv preprint arXiv:2210.13944. 2022.

[9] Dadman S, Bremdal BA, Bang B, Dalmo R. Toward Interactive Music Generation: A Position Paper. IEEE Access. 2022;10:125679-95.

[10] Cochrane T, Fantini B, Scherer KR. The Emotional Power of Music: Multidisciplinary perspectives on musical arousal, expression, and social control. Oxford University Press; 2013. Available from: https://doi.org/10.1093/acprof:oso/9780199654888.001.0001.

[11] Pipe L. The role of gesture and non-verbal communication in popular music performance, and its application to curriculum and pedagogy [Doctoral thesis]. University of West London; 2018. Available from: http://repository.uwl.ac.uk/id/eprint/6751/.

[12] Kelly SN. Using Conducting Gestures to Teach Music Concepts A Review of Research. Update: Applications of Research in Music Education. 1999;18(1):3-6. Available from: https://doi.org/10.1177/875512339901800101.

[13] Dannenberg R, Siewiorek D, Zahler N. Exploring Meaning And Intention In Music Conducting. In: Proceedings of the International Computer Music Conference, ICMC 2010; 2010. p. 327-30.

[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. p. 5998-6008. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[15] Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114. 2013.

[16] Zhao S, Song J, Ermon S. InfoVAE: Information Maximizing Variational Autoencoders. arXiv preprint arXiv:1706.02262. 2017.

[17] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc.; 2020. p. 1877-901. Available from: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[18] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International conference on machine learning. PMLR; 2021. p. 8748-63.

[19] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356. 2022.

[20] Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human–AI Collaborative Decision-Making. Proceedings of the ACM on Human-computer Interaction. 2019;3(CSCW):1-24.

[21] Jensen E, Dale M, Donnelly PJ, Stone C, Kelly S, Godley A, et al. Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020. p. 1-13. Available from: https://doi.org/10.1145/3313831.3376418.

[22] Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125. 2022.

[23] Chen M, Tworek J, Jun H, Yuan Q, Pinto HPdO, Kaplan J, et al. Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374. 2021.
[24] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155. 2022.
[25] Fiebrink R, Trueman D, Cook PR. A Meta-Instrument for Interactive, On-the-Fly Machine Learning. In: Proceedings of the International Conference on New Interfaces for Musical Expression; 2009. p. 280-5.
[26] Næss TR, Martin CP. A Physical Intelligent Instrument using Recurrent Neural Networks. In: Proceedings of the International Conference on New Interfaces for Musical Expression. Porto Alegre, Brazil: UFRGS; 2019. p. 79-82. Available from: http://www.nime.org/proceedings/2019/nime2019_paper016.pdf.
[27] Donahue, Chris and Simon, Ian and Dieleman, Sander. Piano Genie. In: Proceedings of the 24th International Conference on Intelligent User Interfaces; 2019. p. 160-4.
[28] Huang CZA, Hawthorne C, Roberts A, Dinculescu M, Wexler J, Hong L, et al. The Bach Doodle: Approachable music composition with machine learning at scale. arXiv preprint arXiv:1907.06637. 2019.
[29] Louie R, Coenen A, Huang CZ, Terry M, Cai CJ. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2020. p. 1–13. Available from: https://doi.org/10.1145/3313831.3376739.
[30] Zhang Y, Xia G, Levy M, Dixon S. COSMIC: A Conversational Interface for Human-AI Music Co-Creation. In: Proceedings of the International Conference on New Interfaces for Musical Expression. Shanghai, China; 2021. Available from: https://nime.pubpub.org/pub/in6wsc9t.
[31] Huang CZA, Vaswani A, Uszkoreit J, Shazeer N, Hawthorne C, Dai AM, et al. Music Transformer: Generating Music with Long-Term Structure. arXiv preprint arXiv:1809.04281. 2018.
[32] Choi K, Hawthorne C, Simon I, Dinculescu M, Engel J. Encoding Musical Style with Transformer Autoencoders. arXiv preprint arXiv:1912.05537. 2019.
[33] Yu B, Lu P, Wang R, Hu W, Tan X, Ye W, et al. Museformer: Transformer with Fine-and Coarse-Grained Attention for Music Generation. arXiv preprint arXiv:2210.10349. 2022.
[34] Roberts A, Engel J, Raffel C, Hawthorne C, Eck D. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In: Proceedings of the International Conference on Machine Learning. PMLR; 2018. p. 4364-73.
[35] Brunner G, Konrad A, Wang Y, Wattenhofer R. MIDI-VAE: Modeling Dynamics and Instrumenta- tion of Music with Applications to Style Transfer. arXiv preprint arXiv:1809.07600. 2018.
[36] Wu SL, Yang YH. MuseMorphose: Full-Song and Fine-Grained Piano Music Style Transfer with One Transformer VAE. arXiv preprint arXiv:2105.04090. 2021.
[37] von Rütte D, Biggio L, Kilcher Y, Hoffman T. FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control. arXiv preprint arXiv:2201.10936. 2022.
[38] Agostinelli A, Denk TI, Borsos Z, Engel J, Verzetti M, Caillon A, et al. MusicLM: Generating Music From Text. arXiv preprint arXiv:2205.05448. 2023.
[39] Cancino-Chacón CE, Grachten M, Goebl W, Widmer G. Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. Frontiers in Digital Humanities. 2018;5:25.
[40] Widmer G, Goebl W. Computational Models of Expressive Music Performance: The State of the Art. Journal of New Music Research. 2004;33(3):203-16.
[41] Friberg A, Bresin R, Sundberg J. Overview of the KTH rule system for musical performance. Advances in cognitive psychology. 2006;2(2):145.
[42] Jeong D, Kwon T, Kim Y, Lee K, Nam J. VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance. In: Proceedings of the 20th International Society for Music Information Retrieval Conference; 2019. p. 908-15. Available from: http://archives.ismir.net/ismir2019/paper/000112.pdf.
[43] Maezawa A, Yamamoto K, Fujishima T. Rendering Music Performance With Interpretation Variations Using Conditional Variational RNN. In: Proceedings of the 20th International Society for Music Information Retrieval Conference; 2019. p. 855-61. Available from: https://archives.ismir.net/ismir2019/paper/000105.pdf.
[44] Rhyu S, Kim S, Lee K. Sketching the Expression: Flexible Rendering of Expressive Piano Performance with Self-Supervised Learning. arXiv preprint arXiv:2208.14867. 2022.
[45] Salehinejad H, Baarbe J, Sankar S, Barfett J, Colak E, Valaee S. Recent Advances in Recurrent Neural

Networks. arXiv preprint arXiv:1801.01078. 2017.

[46] Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A. A Kernel Method for the Two-Sample-Problem. In: Schölkopf B, Platt J, Hoffman T, editors. Advances in Neural Information Processing Systems. vol. 19. MIT Press; 2006. p. 513—520. Available from: https://proceedings.neurips.cc/paper_files/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf.

[47] Foscarin F, Mcleod A, Rigaux P, Jacquemard F, Sakai M. ASAP: a Dataset of Aligned Scores and Performances for Piano Transcription. In: Proceedings of the 21st International Society for Music Information Retrieval Conference; 2020. p. 534-41. Available from: https://archives.ismir.net/ismir2020/paper/000127.pdf.