A Tool for Composing Music via Graphic Scores in the style of György Ligeti's Artikulation using Self-supervised Representation Learning

Berker Banar and Simon Colton

School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
{b.banar, s.colton}@qmul.ac.uk

Abstract

Graphic scores are powerful and expressive symbolic music notations which are promising for music generation in multi-modal settings. However, it is a challenging task to decipher the relationship between the graphic scores and their corresponding musical pieces to explicitly use the creative mapping between them in generative settings. In this work, we connect graphic score and audio domains using self-supervised representation learning to reveal the mapping between these modalities, and utilise this technique to compose music using graphic scores in the universe of György Ligeti's Artikulation, which is a well-known electronic piece. To experiment with and disseminate this approach, we have built an interactive web application in Hugging Face Spaces, designed using the Gradio SDK.

Introduction

In contemporary music, graphic scores are used as alternatives to traditional score-based music notations. Graphic scores are expressive and arguably easier to interact with for inexperienced music practitioners as they consist of graphical objects, colours and their visual compositions. Yet, mapping between the sonic material and the graphic scores is not universal and deciphering this mapping is not a trivial task given complicated sonic textures and high-level descriptions in the legends of graphic scores (Banar and Colton 2022). Rather than explicitly defining the rules of this mapping, having it learnt by a neural model can enable multi-modal generative settings, where in the case of graphic scores and music, the drawings in the style of a graphic score can be converted into music expressively.

Self-supervised representation learning has been proven to be powerful for deriving multi-modal connections given the success of models such as CLIP (Radford et al. 2021). Also, these models have been effectively utilised in generative settings as shown in examples where CLIP is combined with BigGAN and VQGAN (Brock, Donahue, and Simonyan 2019) (Esser, Rombach, and Ommer 2021).

In the work described here, we apply a self-supervised representation learning technique to connect graphic scores and music in audio form, and to demonstrate this approach, focus on a well-known electronic piece, namely György Ligeti's Artikulation. The neural model consists of two variational autoencoders (VAEs) to encode and decode both the graphic score and audio material and a contrastive learning framework is used to connect two different modalities. Arguably, Artikulation is a suitable piece in this context given its exquisite sonic textures and rich graphical world. Also, we present a web application for music composition via graphic scores in Hugging Face Spaces using the Gradio SDK, which can be accessed via this link[1].

Implementation

The original graphic score of Artikulation, which is designed by artist Rainer Wehinger, is available in pieces representing fragments of roughly 5 to 10 seconds duration. As per (Banar and Colton 2022), we extracted these fragments (an example can be found in Figure 1), and concatenated these fragments into a unified graphic score for the piece as part of the data processing. Then, we windowed the unified graphic score to have 2 seconds of segments with the stride amount of 1 second and restricted the colour palette to 10 colours to make the learning procedure easier. Similar to the windowing procedure of the graphic score, we extracted 2 seconds of audio samples from the audio recording of the piece, where samples match to their corresponding graphic score fragments.

We implemented the self-supervised representation learning framework using two pipelines for graphic scores and audio processing, which is based on (Tatar, Bisig, and Pasquier 2021) with the architecture schematic depicted in Figure 2. Both of these pipelines consist of variational autoencoder architectures (Kingma and Welling 2013) as per (Banar and Colton 2022). In the audio pipeline, audio waveforms are converted into CQT spectrograms (Schörkhuber and Klapuri 2010), which are then reconstructed in the output using fast Grifin-Lim phase reconstruction as in (Tatar, Bisig, and Pasquier 2021). Two encoder-decoder pipelines are connected to each other in a contrastive learning setting using a duplet loss, where the encoded latent vectors of the corresponding graphic score and audio fragments are aimed to be placed as close to each other as possible. This architecture is utilised in a generative setting, where a user provided

[1] https://bit.ly/3ELSq6K

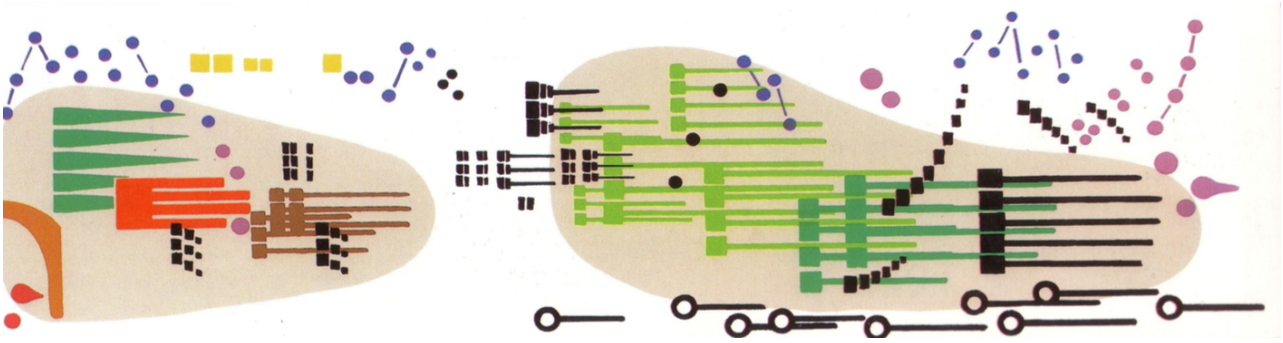

Figure 1: Graphic Score from Ligeti's Artikulation.

graphic score in the style of Artikulation is used to synthesize audio accordingly.

Web Application

Using the framework described, we developed an interactive web application on Hugging Face Spaces using the Gradio SDK (Abid et al. 2019). Using this, users can draw graphic score fragments in the style of Artikulation to compose musical segments. In Figure 3, we provide a screenshot of the interface, where a hand-drawn graphic score is given as an example (in the bottom left corner graphic of figure 3). In this interface, users can select the thickness of their brush and the color. We encourage users to first watch the YouTube video of Artikulation, whose link can be found here[2], to better understand the connections between the electronic piece and the graphical language.

Evaluation of the System

Based on our initial (subjective) evaluations, synthesized material reflects the nature of the textures in Artikulation and also the composition of the user drawn graphic scores. To test some fundamental capabilities of the system, we conducted controlled experiments with a formalism where we explore the effect of colour and repetitions on the generated audio material.

To test the effect of colour, following our 10-colour palette, we drew 10 circles individually in the colours of black, blue, brown, dark green, light green, orange, pink, red, white (the canvas) and yellow. These drawn graphic scores and the mel-spectrograms of their generated audio files are depicted in Figure 4. 128 mel bins are used in the spectrograms, and as per usual practice, horizontal and vertical axes correspond to time and frequency, respectively. Generated audio files for these graphic scores can be found in this SoundCloud page[3]. Based on these experiments, each circle is reflected in the audio file as an individual event and each colour represents a different sonic texture, which are also shown in the mel-spectrograms.

To experiment with the effect of repetitions, we drew an orange circle on the canvas and then added two more similar orange circles one by one. These graphic scores and the mel-spectrograms (similarly with 128 mel bins) of their generated audio files can be found in Figure 5. Similarly, generated audio files for these graphic scores are in the same SoundCloud page. The number of circles in the drawn graphic scores matches with the number of individual events

[2]https://bit.ly/3GvjDMu

[3]https://bit.ly/3XuqsCX

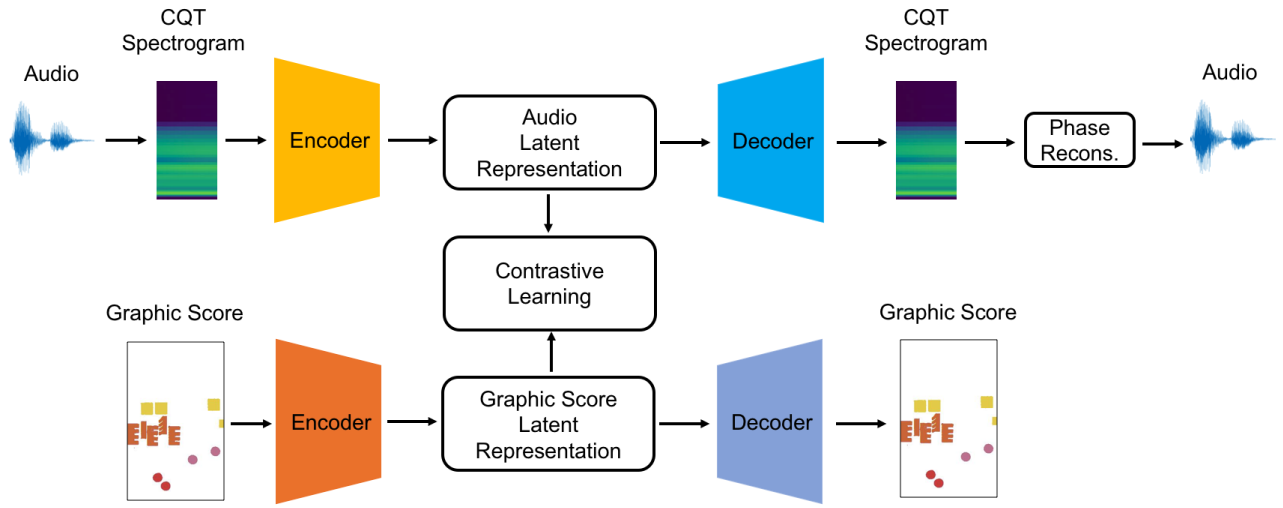

Figure 2: Architecture schematic.

Figure 3: Hugging Face Spaces interface.

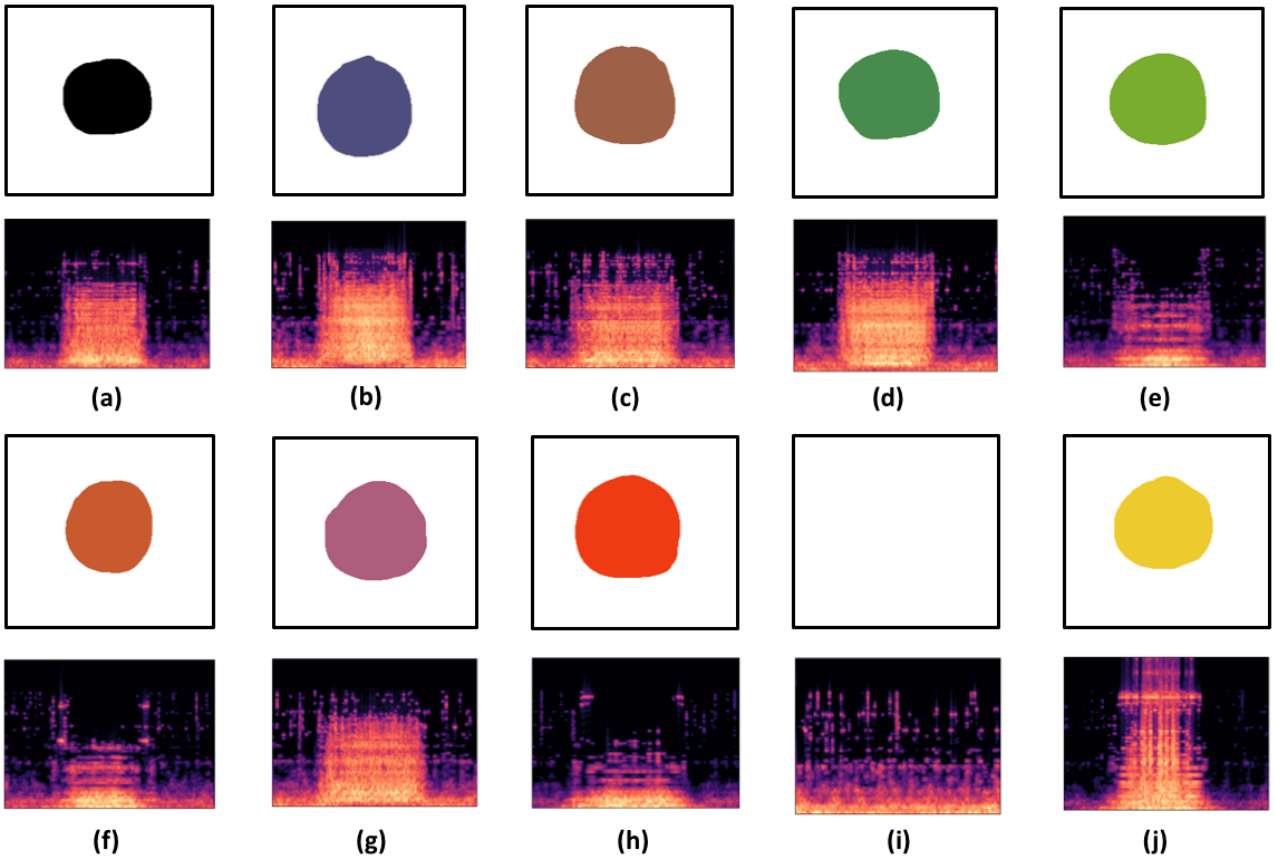

Figure 4: 10 drawn circles in different colours with the mel-spectrograms of their generated audio respectively: (a) black, (b) blue, (c) brown, (d) dark green, (e) light green (f) orange, (g) pink, (h) red, (i) white (the canvas) and (j) yellow.

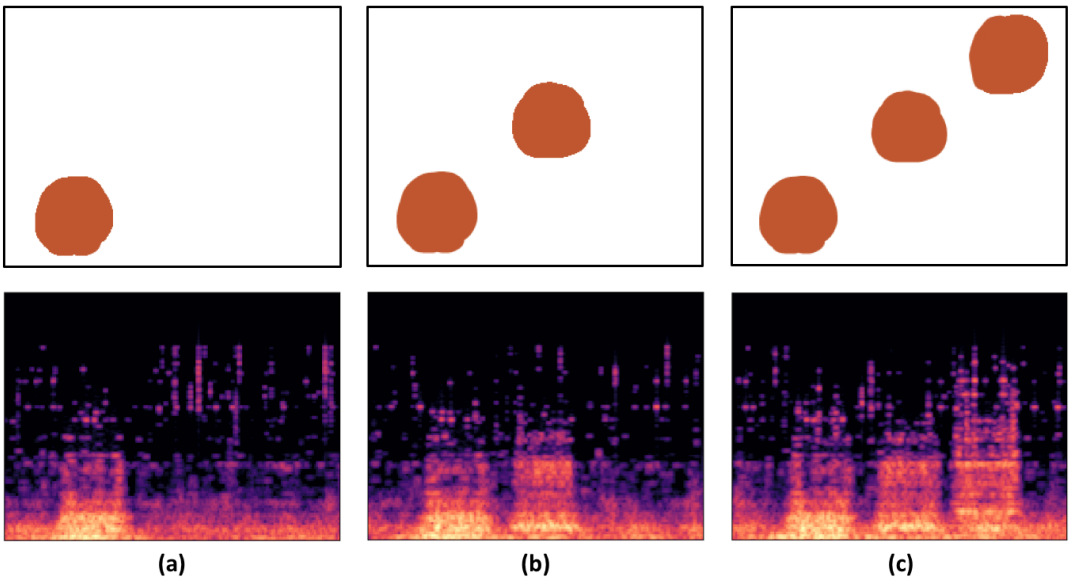

Figure 5: Graphic scores for the experiments repeating orange circles with the mel-spectrograms of their generated audio respectively.

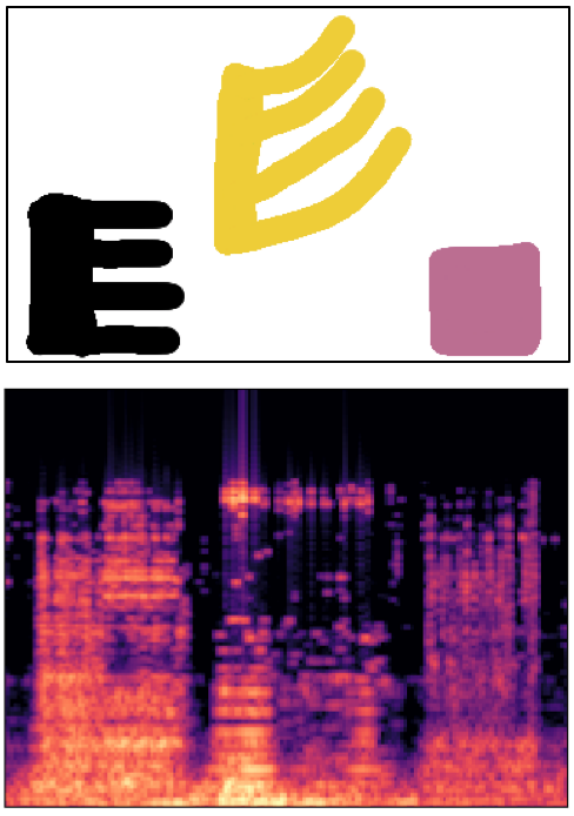

Figure 6: Graphic score for the composed example with the mel-spectrogram of its generated audio.

in their corresponding audio files and the vertical placement of the circles is reflected on the pitch of the events, where higher placement vertically corresponds to a higher pitch.

In addition to the colour and repetition experiments, we composed an example graphic score, which is depicted in Figure 6 with a mel-spectrogram corresponding to its generated audio. In this composed example, individual graphic objects are reflected in individual events with different sonic textures referencing the colours and their vertical placement monotonically corresponds to pitch as explained above.

In future work, we will further experiment with the tool to better understand and improve its sonic capabilities, and we encourage the community to explore the tool as well as it is publicly available on Hugging Face Spaces. In the workshop, we will present a demo of the system.

Conclusions

In this study, we present a tool that can be used to compose new music based on user-prompted graphical scores in the style of György Ligeti's Artikulation. Our tool consists of a Hugging Face Spaces front-end and graphic scores (symbolic music) and audio. Our hope is that the creative AI community will experiment with this tool, discover its potential, and also make suggestions for further improvement. With the recent advancements in multi-modal creative AI, we believe that graphical scores and defined visual spaces that can be mapped to a sonic world have much creative potential, especially in the context of contemporary electronic music, which we plan to further explore as part of future work. Also, we plan to add more features to our tool to enable users to create and host many pairs of graphic scores and audio and modularly curate longer pieces by combining these materials similar to Ligeti's Artikulation. Moreover, we are interested in exploring other pieces from the electroacoustic music repertoire which also have graphic score notations.

Ethical Statement

As this generative system relies on György Ligeti's Artikulation and Rainer Wehinger's graphic score, one ethical problem may be the ownership of the generated material, which is an ongoing discussion in the creative AI field.

Acknowledgements

Berker Banar is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London. We wish to thank the anonymous reviewers for their insightful comments.

References

Abid, A.; Abdalla, A.; Abid, A.; Khan, D.; Alfozan, A.; and Zou, J. 2019. Gradio: Hassle-free sharing and testing of ML models in the wild. *arXiv preprint arXiv:1906.02569*.

Banar, B.; and Colton, S. 2022. Connecting Audio and Graphic Score Using Self-supervised Representation Learning - A Case Study with György Ligeti's Artikulation. *In Proceedings of International Conference on Computational Creativity (ICCC)*.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale GAN training for high fidelity natural image synthesis. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *In Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763.

Schörkhuber, C.; and Klapuri, A. 2010. Constant-Q transform toolbox for music processing. *In Proceedings of the 7th Sound and Music Computing Conference, Barcelona, Spain*, 3–64.

Tatar, K.; Bisig, D.; and Pasquier, P. 2021. Latent Timbre Synthesis: Audio-Based Variational Auto-Encoders for Music Composition and Sound Design Applications. *Neural Computing and Applications*, 33(1): 67–84.