Connecting Audio and Graphic Score Using Self-supervised Representation Learning - A Case Study with György Ligeti's Artikulation

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

Music is a phenomenon that can be represented in various data modalities, such as MIDI, musical score, graphic score and audio. Connecting these modalities in an informative and intelligent way is important, especially for multi-modal music generation systems. In this study, we present a novel self-supervised representation learning approach that can be applied to finding a mapping between audio and graphic scores in a generative context. Our approach consists of two variational autoencoder-based generators and a contrastive learning mechanism. We demonstrate this technique using György Ligeti's Artikulation, which is an electronic music composition with a graphic score. In initial experiments, given manually designed graphic score excerpts in the style of Artikulation, we generated good quality audio correspondents with our model. We further suggest some ways of improving our approach and discuss some future directions for our work.

Introduction

Music can be represented in audio and symbolic (e.g. musical score, MIDI, or graphic score) domains. Commonly in generative music studies, just one of these data modalities is targeted and the generative system is specifically designed for the selected domain (Oord et al. 2016) (Payne 2019). While there are studies that focus on connecting some of these modalities (Wang and Yang 2019), the full potential of multi-modal representations has not been fully explored in generative contexts yet. This is especially true for a wide range of sonic and timbral options and various music representations. Connecting these different music representations is beneficial in an end-to-end multi-modal music generation pipeline, where the generation starts in the symbolic music domain, and then symbolic material is converted into audio via a mapping between symbolic and audio representations. In this multi-modal setting, we benefit from the advantages of both worlds, where the symbolic representation enables us to control the generation process in terms of some high-level musical attributes such as tonality, harmony and rhythmic complexity, and provides us with a confined format; while the audio representation allows us to introduce expressive, textural and complex elements in a sonic domain, where we appreciate music as people.

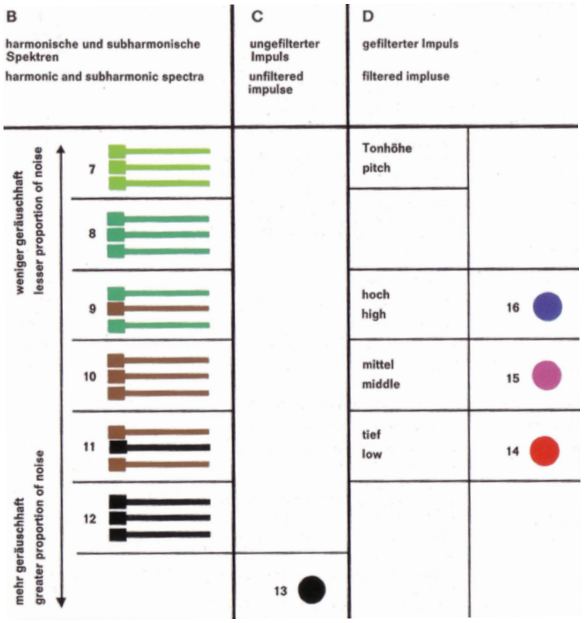

Figure 1: Legend for sonic objects in the graphic score of György Ligeti's Artikulation.

It has become clear recently that self-supervised representation learning can be highly effective, as highlighted by the success of the CLIP model (Radford et al. 2021) for mapping both images and text into the same latent space. Such contrastive learning can then be used in generative methods, for instance with CLIP being used to guide GAN image generation, such as with BigGAN (Brock, Donahue, and Simonyan 2018) or VQGAN (Esser, Rombach, and Ommer 2021). Inspired by these successes, we believe self-supervised representation learning approaches for connecting symbolic music and audio domains could enhance the creative potential of generative music models.

In the matter of symbolic music representations, traditional musical scores might be limited in terms of expressing the actual music itself, specifically in scenarios such as electroacoustic and acousmatic music. In contemporary classical music (Spencer 2015), graphic scores act as alternative music notations, and allow more expressive performance details to be represented, particularly for subtle and continuous

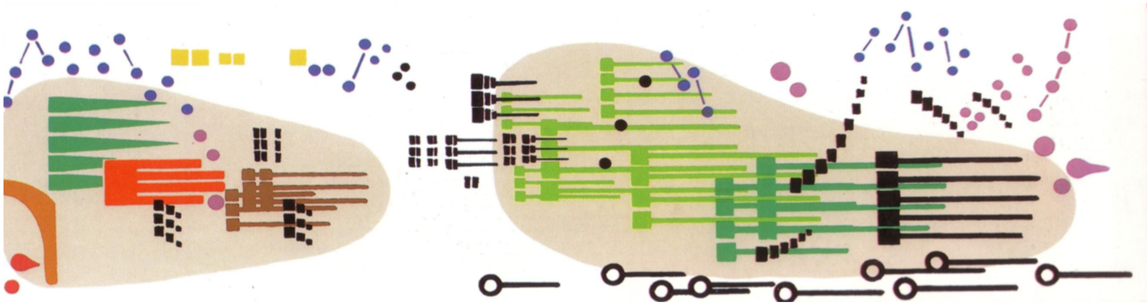

Figure 2: Graphic score fragment from Ligeti's Artikulation.

changes. Such scores engage performers to follow abstract visual mappings, which can be attractive to manipulate for inexperienced practitioners, e.g., for those without formal training on traditional Western scores. Graphic scores are not universal, however, and their organisation depends a lot on the unique mappings given in a legend, as in Figure 1; deciphering graphic scores is a challenging task. One well-known graphic score is for Artikulation by György Ligeti, designed by Rainer Wehinger, who first listened to the piece and then constructed coherent abstractions to illustrate the musical entities presented. An example of this score is depicted in Figure 2. In the organisation of this graphic score, the horizontal axis represents time, the vertical axis represents pitch and coloured shapes represent unique sonic entities that are used in the piece as themes and musical ideas.

In this study, we present a self-supervised representation learning framework to connect audio and graphic score domains, and demonstrate a creative composition use case that allows practitioners to compose in the style of Artikulation utilising its visual and sonic universe. Without such an approach, this task might not be possible, as it is challenging to separate and re-synthesize the complex textures in Artikulation by listening to the piece and looking at the graphic score and its abstract legend. Practically, in our use case, our system allows us to generate new audio segments, which are conditioned on manually created graphic score excerpts that are not part of the original graphic score, but in its graphical style. Demonstrating this feature, we exhibit some manually created graphic score fragments and their synthesised correspondents within the aesthetics of the piece. To conclude, we address potential ways of improving this system and some future directions for this study.

Data Processing

The original graphic score of Artikulation is presented in fragments of 5 to 10 seconds duration. First, we cropped these fragments and manually processed them to get rid of the time axis lines and canvas contours, then merged these processed fragments into a single long image file constituting the whole graphic score for the piece. Then, we extracted graphic score excerpts using 2 seconds of windows, where the stride amount is 1 second. As the piece is 227 seconds long, this excerpt extraction process gave us 226 windows in total. Then, we further processed these extracted images to restrict their palettes to 10 discrete colours to make the learning procedure easier. One caveat is that this procedure gets rid of the grey shaded regions in the original score, which represent the effect of reverb. In our future work, we will further experiment with graphic score excerpts that have such reverb regions.

We recorded the audio file of Artikulation while streaming the piece online from YouTube at 44.1kHz sampling rate and applied a similar data processing where 2 seconds of audio fragments were extracted, again with the stride amount of 1 second. These audio fragments were paired with their corresponding graphic score excerpts. Then, we used constant-q transform (CQT) (Schörkhuber and Klapuri 2010), which is a wavelet-based time-frequency transform, to generate spectrograms for each audio file, to be used in the learning process.

Model Architecture

Our architecture consists of three main sub-parts, which are an audio pipeline, a graphic score pipeline, and a contrastive learning block for self-supervised representation learning, as illustrated in Figure 3. Both the audio and graphic score pipelines utilise a variational autoencoder (VAE) architecture (Kingma and Welling 2013) and our contrastive learning mechanism is based on the cosine similarity between audio/graphic score latent representations using a duplet loss.

In the audio pipeline, we have an encoder-decoder architecture, which is taken from (Tatar, Bisig, and Pasquier 2021) and the audio data is presented to the network in CQT spectrogram format (Schörkhuber and Klapuri 2010). The encoder part has two consecutive 4096-dimensional dense layers that are followed by two parallel 4096-dimensional dense layers embedding in two 512-dimensional spaces, which are for the mean and the variance of variational sampling to a 512-dimensional space. The decoder part has three dense layers with 4096 dimensions. During the training procedure, we use the Adam optimiser (Kingma and Ba 2014) where the learning rate is 0.0001, β_1 is 0.9 and β_2 is 0.999. As per the typical configuration of VAEs, the loss function of this encoder-decoder architecture has two parts, namely the reconstruction loss and regularisation loss, and a mean squared error loss function is used for the reconstruction part, where KL-divergence (Kullback and Leibler 1951) is used for the regularisation. In this pipeline, our decoder generates CQT spectrograms, which are then converted into audio files using fast Grifin-Lim phase reconstruction as in (Tatar, Bisig, and Pasquier 2021).

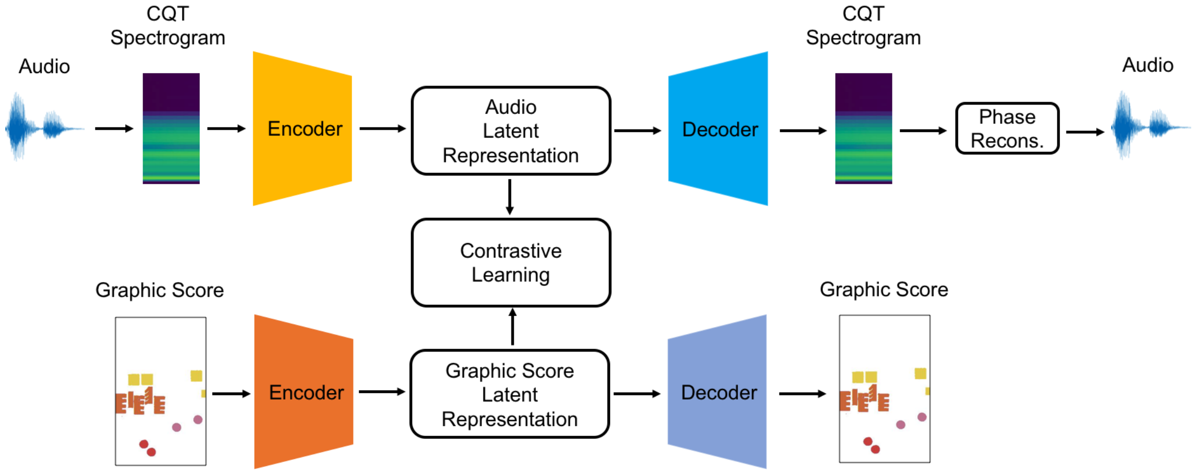

Figure 3: Architecture schematic for the two VAEs and contrastive learning block.

The graphic score pipeline also uses an encoder-decoder architecture, which directly takes and generates RGB images on both sides. The encoder here first flattens the RGB image, then passes it through a 2048-dimensional dense layer, which is followed by two parallel 2048-dimensional dense layers. The embedding space has 512 dimensions similar to the audio pipeline. The decoder part consists of two 2048-dimensional dense layers, which are followed by a deflattening procedure, which converts single stream decoder outputs into three channel RGB images. An adam optimiser is used as in the audio pipeline with the same parameters. Similar to the case above, we use mean squared error and KL-divergence for reconstruction and regularisation losses, respectively.

The contrastive learning block aims to make the corresponding embeddings of the graphic score and audio pairs as close to each other as possible, using cosine similarity between their mean and variance latent vectors in variational autoencoders. The training procedure utilises a multi-task optimisation process, where we train the VAE architectures for reconstruction and the contrastive learning block for self-supervised representation learning using a unified loss, simultaneously. Since we have two main objectives, which are the reconstruction quality of VAEs and creating a structured embedding space, we weight our VAE and contrastive losses. Based on our initial experiments, which suggest that audio reconstruction might be a more challenging task in this setting and require more attention, VAE losses for the graphic scores and audio are weighted as 10% and 90% with respect to each other. We also downscale the cosine similarity loss between 512-dimensional latent vectors by a factor of 50, in order to make it aligned with the VAE losses and have 0,0x decimal numbers for each of our losses at the beginning of our training procedure. We trained our complete model for 200 epochs with a batch size of 32.

We use this architecture in a multi-modal generative setting, where a user-designed graphic score is encoded into the embedding space and its latent vector is decoded using the audio decoder. Graphic score and audio embeddings share the same latent space due to the contrastive learning strategy, thus, the latent vector of a given graphic score can be interpreted keeping the semantic connections between two data modalities. This shared embedding space approach has been successfully demonstrated in the CLIP model (Radford et al. 2021), which uses text and image data, but CLIP requires a separate generator to create artefacts. In our approach, we combine the self-supervised representation learning and generation tasks in the same model and training procedure, and utilise this technique with graphic scores and audio, which allow us to create a generative universe in the style of a piece or a composer.

Experiments

To demonstrate the reconstruction capability of our model, we use four audio and four graphic score excerpts that are all from Artikulation, originally. We reconstruct these excerpts using our audio and graphic score pipelines that are trained separately without the contrastive learning block. We also reconstruct these original excerpts using our trained complete architecture. All of the reconstructed graphic scores including the originals are exhibited online (Figure 5, 6 and 7)[1] and all the reconstructed and original audio files (Original 1-4) are presented on a SoundCloud page[2]. For the separately trained pipelines, as demonstrated with the figures and audio files, reconstruction quality is high. For our trained complete architecture, although the reconstruction quality slightly decreases compared to the separately trained pipelines, which is expected due to introducing the contrastive learning block, reconstructed graphic scores and audio files exhibit intelligible graphical objects and good quality sonic entities, respectively.

In order to test our trained architecture in a multi-modal generation scenario, we manually designed four different graphic score fragments in the style of Artikulation, which are not exactly the same as any of the original graphic score fragments. This approach demonstrates the creative potential of the system, where creators can compose their own

[1] https://bit.ly/37A1CgV

[2] https://soundcloud.com/user-330551093/sets/audio-sym-ssrl

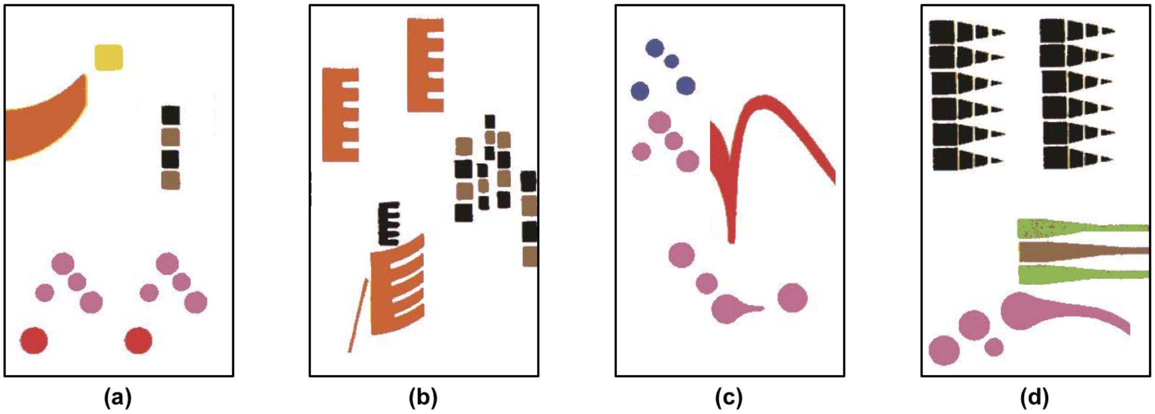

Figure 4: Four manually designed graphic score excerpts.

musical pieces by designing graphical scores in the universe of Artikulation. Our expectation is that combined models are able to generate an audio excerpt that sonically reflects the material presented in the given graphic score fragment in alignment with the characteristics of Artikulation. Our designed graphic scores are presented in Figure 4 and we display their reconstructed versions generated using our complete model via the same link[1] (Figure 8). Even though the reconstructed versions are lower in quality compared to the originals, they are successful in terms of representing the graphical content, shapes and colours. We exhibit the multi-modally generated audio files of four manually designed graphic score excerpts on the same SoundCloud page[2]. When we analyse these audio files, even though they are not considered to be good quality regarding clear sonic textures compared to Artikulation, the generated audio files still exhibit the textures of the piece and are reflective in terms of the visual composition.

When we evaluate these audio files in more detail, in the beginning of generated audio for graphic score (a), we have a sonic element with rising pitch similar to the curved orange shape on its graphic score. Generated audio for excerpt (b) demonstrates a similar rising pitch object, but lower in pitch compared to excerpt (a), which can be associated with the comb-shaped curvy figure on the lower side of the vertical axis, which corresponds to musical pitch. In the audio excerpt for graphic score (d), we have a unique and strong sonic statement, which might be reflected in the horizontal cone-like black figures. A similar sonic entity is repeated on the second half of this audio excerpt, but differently, which might correspond to the second set of black figures. The difference can be due to having horizontal green and brown shapes happening at the same time. In our future work, we plan to quantitatively analyse generated audio files using audio similarity metrics, to better evaluate the reflectiveness of their given graphic scores and the style of Artikulation in general.

Conclusions

In this study, we present a novel framework that connects audio and graphic score domains using self-supervised representation learning, which can be extended to other music data modalities. We demonstrate its use case in a scenario where we utilise Ligeti's Artikulation represented in both a graphic score and audio forms, and also exhibit the results of our initial experiments with the system, which generates music in audio format in the style of Artikulation based on unseen but stylistically similar graphic score excerpts presenting a creative use case of this generative system in the context of human-machine co-creation. Even though the results are not perfect, we believe that this approach has valuable potential, especially to be utilised in multi-modal music generation systems. Also, due to the artistic form of this graphical music representation, we think that sonifying visuals in a defined sonic and visual space is valuable from a computational creativity perspective, as it might allow to further pieces with rich textures referencing a variety of visual abstractions and reflecting complex styles of composers.

In our future work, we plan to improve the quality of the generated material as well as the generalisation capability of the model by further experimenting with the architecture and applying data augmentation both in visual and audio domains. Also, besides our own subjective evaluation, we will introduce numerical metrics which can evaluate the closeness of generated audio material to given graphic scores in the context of Artikulation. To improve the match between a given graphic score excerpt and its corresponding generated audio, we plan to experiment with introducing various inductive biases to the model, which might ease the learning process and allow the model to learn a mapping between the graphic score and audio more effectively. Besides Artikulation, we will experiment with other contemporary classical music pieces with graphic scores using our approach. Additionally, we are interested in using this technique in other combinations of data modalities as well, such as audio-MIDI. Moreover, we would like to build an online tool based on this system, which can generate music using graphic score excerpts specifically created by the users. Furthermore, we plan to utilise this system in a scenario where an audio excerpt in the style of Artikulation is provided and the model is expected to generate its corresponding graphic score (i.e., the reverse direction to the inference workflow discussed here), which enhances the creative potential of this approach.

Author Contributions

The majority of the technical and experimental work for this paper was undertaken by the first author, Berker Banar. The second author, Simon Colton, contributed some image processing code and advice on the direction of this work. The paper was written by Berker Banar with some contributions and editing by Simon Colton.

Acknowledgments

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

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv:1809.11096*.

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 Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

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

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv:1609.03499*.

Payne, C. 2019. Musenet. https://openai.com/blog/musenet/. Accessed: 2022-04-20.

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*, 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.

Spencer, M. 2015. Art and music collide in these 20 stunning graphic scores. https://www.classicfm.com/discover-music/latest/graphic-scores-art-music-pictures/. Accessed: 2022-04-20.

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.

Wang, B., and Yang, y.-h. 2019. PerformanceNet: Score-to-audio music generation with multi-band convolutional residual network. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:1174–1181.