Sound-based music style modelling, for a free improvisation musical agent

Andrea Bolzoni
School of Computing and Communications
The Open University
Milton Keynes
MK7 6AA, UK
andrea.bolzoni@open.ac.uk

Abstract

This paper presents the first stages of development of an improvising musical agent capable of interacting with a human musician in a free improvised music context. This research aims to explore the creative potential of a co-creative system that draws upon two approaches of music and sound generation: the well-established practice of modelling musical styles with Markov-based models and recent developments in neural-network-based audio synthesis. At this preliminary stage, the focus is on the definition of style in a sound-based music context, and the outline of a formal evaluation framework for style imitation systems.

1. Introduction

The field of sound-based music has been one of the most hindered in Computational Creativity. The reason why lies in the wide range of unusual sounds employed, and the challenges in classifying them and modeling musical structures based on them. In particular, the missing link is in the development of a formal evaluation framework for sound-based music style modelling. In fact, even though some sound-based music style modelling systems have been developed (Tatar & Pasquier, 2017; Bernardes, Guedes, & Pennycook, 2013), it is unlikely to find a study that objectively evaluates their capabilities to do so.

Sound-based music style modelling can draw on the recent developments in the field of neural-based sound synthesis. In fact, neural-based sound synthesis would allow the expansion of the sound palette available. This will save time in building a big audio database to retrieve the sound samples from, as well as space to store it. In addition, I claim, it will facilitate the emergence of novel pieces through the exploration of a given style.

At this stage, I have implemented and tested already existent approaches to automatic sound-based composition. The aim was to gain experience for developing a musical agent that will be able to compose in real-time – improvising – along with a human musician.

2. Background

In the following subsections I will outline the context for this research. It embraces the fields of sound-based music, concatenative sound synthesis, statistical style modelling, and evaluation.

2.1. Sound-based music

Landy (2007) defines sound-based music as “the art form in which the sound, that is, not the musical note, is its basic unit”. In sound-based music, the individual entities that constitute a piece of music are commonly referred to as sound objects. Ricard and Herrera (2004) define a sound object as “any [sonic] entity perceived as having its own internal properties and rules”. Roads (2002) defines a sound object as "a basic unit of musical structure, generalizing the traditional concept of note".

2.2. Concatenative Sound Synthesis

Concatenative sound synthesis draws on two other synthesis techniques: granular sound synthesis, where sound synthesis is performed through the generation of very short synthetic sonic grains (Roads, 1988); and granulation, where an audio corpus is segmented in tiny grains that are reassembled through time-domain-based operations (Roads, 2002). In concatenative sound synthesis the audio corpus is segmented into units (Schwarz, 2006). Each unit is a short sound segment of variable length. Sonic features - such as pitch, duration, or audio descriptors - are extracted from the units. The resynthesis is performed through an algorithm that looks into the audio corpus for the closest units, most of the time in terms of Euclidean distance, in relation to a feature target.

Thanks to its potential, this synthesis technique has been applied in various forms and to various systems. CataRT (Schwarz, Beller, Verbrugghe, & Britton, 2006) allows the exploration of a sound corpus through a user interface where the segmented corpus is shown in a 2D space. MATConcat (Sturm, 2006) offers an implementation of adaptative concatenative sound synthesis where the feature target is controlled by the audio descriptors extracted from an audio file. Similarly, AudioGuide (Hackbarth, Schnell, Esling, & Schwarz, 2013) aims to extract morphologies from an audio file to generate new sonic material through concatenative sound synthesis. A different approach is offered in earGram (Bernardes et al., 2013), where temporal modelling is used to generate new sonic outcomes with a similar style to the audio corpus, for example in relation to the harmonic or timbral content.

More recently, thanks to neural network based generative techniques, a different approach has been proposed. Training a variational autoencoder, it is possible to learn a probabilistic distribution of the units, called a latent space. This is a continuous invertible space. Therefore, it is possible to synthesise units that match the feature target even if they were not in the audio corpus (Bitton, Esling, & Harada, 2020).

2.3. Statistical Style Modelling

Music can be seen as "organised sound" (Varése & Wen-chung, 1966). Therefore, in principle, a music corpus treated as a sequence of organised events can be represented through a model. There are two main approaches to define such models: explicitly code the stylistic rules or infer the rules through statistical analyses (Conklin & Witten, 1995). Belonging to the latter approach, Markov-based models are widespread in music style modelling for their ability to model music patterns (Pachet, 2003).

If we look at the two main shortcomings of Markov Models, we will see that: 1) if the order is low they can model patterns, but they can't properly model the structure of a piece (Pachet, 2003); 2) if the order is high they generally overfit the piece (Papadopoulos, Roy, & Pachet, 2014). Indeed, Pachet (2003) states that an interactive system could benefit from the ability of Markov models to learn patterns, while the definition of the structure can be left to the human musician interacting with the system. In this way, there are no drawbacks from their inability to learn long-term structures. Another way to compensate the lack of ability to model long-term structures is proposed in Improtek (Nika & Chemillier, 2012), a developed version of Omax (Assayag, Bloch, & Chemillier, 2006). Here, the interaction between the human musician and the system is based on a predefined dynamic score.

2.4. Evaluation

In the context of Musical Computational Creativity, evaluation is at the same time one of the most important aspects and one of the most overlooked. Evaluation is necessary to show the progress and contributions to the field (Jordanous, 2012), but only a small number of the systems presented in conferences offer a formal evaluation. Tatar and Pasquier (2019) provide a clear example in their typology and exposition of the state of the art of musical agents. Here, they show that, out of 78 presented systems, only 17 had undergone an evaluation process. An approach to evaluating free improvisation in proposed by Linson, Dobbyn, Lewis, and Laney (2015).

This lack of evaluation could be due to a highly fragmented field, where many systems have been developed for specific creativity needs of their developers. Hence, the difficulty of objectively evaluate them (Gifford et al., 2018). Nevertheless, trying to reduce this fragmentation might not be a solution: even though it might give more opportunities to develop more solid evaluation frameworks, the specificity of the tasks carried out by the systems can increase their success (Truax (1980) cited in Pasquier, Eigenfeldt, Bown, and Dubnov (2016)). As a consequence of this fragmentation, the tasks that the systems are asked to carry out have not a "yes or no answer"; and, the evaluation of their outcomes very often depends on the subjective preferences of the users or the audience (Pasquier et al., 2016).

Looking at the bigger picture, the lack of consensus on what creativity is - human and artificial - makes the evaluation of artificial agents' creativity a non-trivial task (Jordanous, 2012). Although a number of evaluation frameworks have been proposed in the last few years, the differences among systems might result in the necessity of tailoring evaluation strategies "to specific research goals in ways that are relevant and have integrity" (Pasquier et al., 2016).

3. Research perspectives

In Computational Creativity, statistical modelling has been widely implemented to model and generate note-based music (Assayag et al., 2006; Conklin, 2003; Pachet, 2003). Sound-based music has been studied to some extent, primarily employing concatenative sound synthesis along with Markov-based style modelling (Tatar & Pasquier, 2017; Bernardes et al., 2013).

Even though note-based musical agents can only generate notes, those notes can be played and interpreted in a variety of ways through synthesised sounds or human musicians. Sound-based musical agents rely on an audio corpus. Therefore, their output is limited to the sonic material present in the audio corpus. This material can be expanded through sound processing techniques, but the audio quality could, nevertheless, easily degrade (Schwarz, 2006).

The development of neural-based synthesis techniques in the last few years opened new possibilities for sound-based musical agents. These techniques can model an audio corpus as an invertible space. Therefore, they give the opportunity to synthesise sounds that were not present in the audio corpus (Bitton et al., 2020).

We will provide a system that will draw on mature work from the field of statistical modelling merged with the expressivity of neural-based sound synthesis. The aim is to contribute to the study of musical computational creativity through a system capable of provoking novel interactions in a free improvised context.

4. Discussion

Style imitation is defined by Pasquier et al. (2016) as: "Given a corpus C = C1, ... Cn representative of style S", style imitation is the generation of "new instances that would be classified as belonging to S by an unbiased observer (typically a set of human subjects)". As a general definition, style means "a particular manner or technique by which something is done, created, or performed"[1]. Among the meanings of music style offered by Dannenberg (2010), we find that use of musical texture could be an aspect of a given style. Nevertheless, musical texture is a difficult component to define. From a sound-based musical point of view, it can certainly be related to the spectromorphological approach proposed by Smalley (1997). From a computational perspective, the spectromorphological approach has been implemented to model an audio corpus - and its style (Tatar & Pasquier, 2017; Bernardes et al., 2013). But, from an evaluation perspective it is still an open question how to define style in sound-based music - and, therefore, how to define the parameters to be evaluated.

In note-based style modelling the use of MIDI notes let us use a symbolic representation of music that can be resynthetised for the purpose of evaluation. As an example, Collins, Laney, Willis, and Garthwaite (2016) use a synthesised piano to reproduce MIDI files in order to evaluate the stylistic success of computationally generated mazurkas. The basic component used to model the style - the note - is detached from the sonic rendering of the music.

In sound-based music, to some extent the sounds used in the audio corpus define themselves the style. And, as exposed in the Background section, sound-based musical agents usually generate their outputs retrieving sounds from the same audio corpus they modelled. Therefore, another open question is to what extent the stylistic success of the music generated by such models is based on the modelling technique implemented, or on the sounds that constitute the audio corpus.

[1] https://www.merriam-webster.com/dictionary/style, accessed on 7 April 2022.

5. Acknowledgments

Thanks to my supervisors Dr Robin Laney, Dr Alistair Willis, Prof David Sharp, and Dr Adam Linson for their suggestions and support. This research is funded by The Open University.

6. References

Assayag, G., Bloch, G., & Chemillier, M. (2006). OMax-Ofon. In *Sound and music computing (smc)*. Marseille, France.

Bernardes, G., Guedes, C., & Pennycook, B. (2013). EarGram: An Application for Interactive Exploration of Concatenative Sound Synthesis in Pure Data. In *International symposium on computer music modeling and retrieval* (pp. 110–129). Retrieved from `http://link.springer.com/10.1007/978-3-642-41248-6_7` doi: 10.1007/978-3-642-41248-6_7

Bitton, A., Esling, P., & Harada, T. (2020). Neural Granular Sound Synthesis. *CoRR*, *abs/2008.0*, arXiv preprint arXiv:2008.01393. Retrieved from `https://arxiv.org/abs/2008.01393` doi: https://doi.org/10.48550/arXiv.2008.01393

Collins, T., Laney, R., Willis, A., & Garthwaite, P. H. (2016, feb). Developing and evaluating computational models of musical style. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *30*(1), 16–43. Retrieved from `https://www.cambridge.org/core.https://www.cambridge.org/core/product/identifier/S0890060414000687/type/journal_article` doi: 10.1017/S0890060414000687

Conklin, D. (2003). Music Generation from Statistical Models. In *Proceedings of the aisb 2003 symposium on artificial intelligence and creativity in the arts and sciences* (pp. 30–35).

Conklin, D., & Witten, I. H. (1995, mar). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*(1), 51–73. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/09298219508570672` doi: 10.1080/09298219508570672

Dannenberg, R. B. (2010). Style in Music. In *The structure of style* (pp. 45–57). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `http://link.springer.com/10.1007/978-3-642-12337-5_3` doi: 10.1007/978-3-642-12337-5_3

Gifford, T., Knotts, S., McCormack, J., Kalonaris, S., Yee-King, M., & D'Inverno, M. (2018, jan). Computational systems for music improvisation. *Digital Creativity*, *29*(1), 19–36. Retrieved from `https://www.tandfonline.com/doi/full/10.1080/14626268.2018.1426613` doi: 10.1080/14626268.2018.1426613

Hackbarth, B., Schnell, N., Esling, P., & Schwarz, D. (2013, feb). Composing Morphology: Concatenative Synthesis as an Intuitive Medium for Prescribing Sound in Time. *Contemporary Music Review*, *32*(1), 49–59. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/07494467.2013.774513` doi: 10.1080/07494467.2013.774513

Jordanous, A. (2012, sep). A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative. *Cognitive Computation*, *4*(3), 246–279. Retrieved from `http://link.springer.com/10.1007/s12559-012-9156-1` doi: 10.1007/s12559-012-9156-1

Landy, L. (2007). Introduction. In *Understanding the art of sound organization* (pp. 1–20). MIT Press.

Linson, A., Dobbyn, C., Lewis, G. E., & Laney, R. (2015, dec). A Subsumption Agent for Collaborative Free Improvisation. *Computer Music Journal*, *39*(4), 96–115. Retrieved from `http://dx.doi.org/doi:10.1162/COMJa00323https://direct.mit.edu/comj/article/39/4/96/106783/A-Subsumption-Agent-for-Collaborative-Free` doi: 10.1162/COMJ_a_00323

Nika, J., & Chemillier, M. (2012). Improtek: integrating harmonic controls into improvisation in the filiation of OMax. In *International computer music conference (icmc)* (pp. 180–187). Ljubljana, Slovenia. Retrieved from `https://hal.archives-ouvertes.fr/hal-01059330`

Pachet, F. (2003, sep). The Continuator: Musical Interaction With Style. *Journal of New Music Research*, *32*(3), 333–341. Retrieved from `http://www.tandfonline.com/doi/abs/`

10.1076/jnmr.32.3.333.16861 doi: 10.1076/jnmr.32.3.333.16861

Papadopoulos, A., Roy, P., & Pachet, F. (2014). Avoiding plagiarism in Markov sequence generation. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 28, pp. 2731–2737).

Pasquier, P., Eigenfeldt, A., Bown, O., & Dubnov, S. (2016, dec). An Introduction to Musical Metacreation. *Computers in Entertainment*, *14*(2), 1–14. Retrieved from https://dl.acm.org/doi/10.1145/2930672 doi: 10.1145/2930672

Ricard, J., & Herrera, P. (2004). Morphological sound description: computational model and usability evaluation. In *Audio engineering society 116th convention.*

Roads, C. (1988). Introduction To Granular Synthesis. *Computer Music Journal*, *12*(2), 11–13.

Roads, C. (2002). *Microsound.* MIT Press.

Schwarz, D. (2006). Concatenative sound synthesis: The early years. *Journal of New Music Research*, *35*(1), 3–22.

Schwarz, D., Beller, G., Verbrugghe, B., & Britton, S. (2006). Real-Time Corpus-Based Concatenative Synthesis with CataRT. In *9th international conference on digital audio effects (dafx)* (pp. 279–282).

Smalley, D. (1997, aug). Spectromorphology: explaining sound-shapes. *Organised Sound*, *2*(2), S1355771897009059. Retrieved from http://www.journals.cambridge.org/abstract_S1355771897009059 doi: 10.1017/S1355771897009059

Sturm, B. L. (2006). Adaptive Concatenative Sound Synthesis and Its Application to Micromontage Composition. *Computer Music Journal*, *30*(4), 46–66.

Tatar, K., & Pasquier, P. (2017). MASOM: A Musical Agent Architecture based on Self-Organizing Maps, Affective Computing, and Variable Markov Models. In *Proceedings of the 5th international workshop on musical metacreation (mume 2017).* Atlanta, Georgia, USA. Retrieved from https://musicalmetacreation.org/mume2017/proceedings/Tatar.pdf

Tatar, K., & Pasquier, P. (2019, jan). Musical agents: A typology and state of the art towards Musical Metacreation. *Journal of New Music Research*, *48*(1), 56–105. Retrieved from https://www.tandfonline.com/doi/full/10.1080/09298215.2018.1511736 doi: 10.1080/09298215.2018.1511736

Varése, E., & Wen-chung, C. (1966). The Liberation of Sound. *Perspective of New Music*, *5*(1), 11–19. Retrieved from https://www.jstor.org/stable/832385