Perception, Cognition, and Generation of Music: a Case for Human-Like Understanding of Temporal Structures

Tillman Weyde

City, University of London, t.e.weyde@city.ac.uk

Abstract. Long-term temporal structures as exhibited in music are difficult to model in machine intelligence and learning. Although symbolic and sub-symbolic models have been created, the problem is still far from solved. The integration of signal analysis and higher-level symbolic cognition, which is natural to humans, is still a challenge in computational systems. Approaches combining generation and analysis have recently attracted more attention and shown positive potential. The case is made for music perception, cognition, and generation as a phenomenon and modelling task that exemplifies the need for human-like computing and has the advantage of presenting it in a form that is largely independent of external reference systems.

Keywords: artificial intelligence, machine learning, music, composition, perception, cognition

1 Introduction

An important feature of human perception, cognition, and action is that they are temporal processes and that their objects typically have temporal dimensions. In some areas, like much of mathematics, the temporal dimension is less relevant, and in others, like natural language, relatively little modelling of time is needed, as much of syntax and semantics is time independent. In music, however, the specific temporal design and performance are essential to the experience of music and its syntactic structure, while there is very little referential semantics. Music thus presents the interesting problem of modelling temporal structure that is mainly determined by human perception, cognition and generation.

Music analysis and generation have long been a topic of interest to machine intelligence research, since early works such as the *Illiac Suite* of 1957 [10]. Especially in the last 15 years much progress has been made in audio-based music information retrieval (see [11] for an overview). Although classification techniques and recommender systems have been developed and refined, their performance has hit so-called glass ceilings and there are many aspects of music that current automatic systems do not capture successfully [13], in particular the temporal structure of repetition, similarity, and variety and their development. In music generation, this problem of generating temporal musical structure became already apparent since the 1960s, as summarised in [2], and is still considered generally an open problem [9].

2 Symbols and Sub-symbolic Models

Human perception, cognition, and generation of music involves the signal level, the perception of sound, and the symbolic level, the cognition in terms of notes, chords and temporal structures. For human music listeners and players, the integration of signal and symbol level is natural and intuitive. On the other hand, the automatic extraction of symbolic musical elements is still considered an open problem [1]. Some models of dependencies between these levels have been developed (e.g. [12]), but there is no generally agreed approach.

In this context, it is an interesting observation that most music listeners have no musical training and are typically unable to write music or describe music in symbolic representations. This suggest that a symbolic representation may not be necessary to perform some musical tasks. This approach has been followed in much of Music Information Retrieval, where the recognition of high-level features, such as genre or artist is often modelled based on engineered feature values that are extracted from the audio, without any symbolic representation [5]. More recently methods have been proposed that work without engineered features and instead operate directly on a time-frequency representation or even the raw audio signal [3]. This approach has very recently also been applied to the generation of audio signals from a machine learning system [4]. However, the glass ceilings still apply for many tasks, which may be related to the lack of symbolic information and background knowledge.

3 Analysis and Generation

Humans with musical experience can produce music that has recognisable temporal structure so that listeners familiar with a style can intuitively recognise the themes, variations, and development during a piece of music. It is a characteristic property of human learning and intelligence that knowledge and performance develop in the interplay between perception, cognition, and production of music and other human activities, such as language, sports, or visual art. Therefore we need to consider the relation of generation and analysis and their interaction to enable intelligent systems to learn relevant concepts and structure from limited data, which often the case in music, in particular for longer temporal structures. An interesting related approach in machine learning is that of adversarial generative nets [6], where a model generates adversarial examples, which can improve classification models [7]. This approach is successful at improving classifiers and generators, but it still remains to be see it could be effective for long term temporal structures.

Specifically music generation has recently attracted attention, e.g. in Googles project Magneta [8], acknowledging the problem of long-term temporal structure. However, the long-term structure is typically still created using existing templates, which clearly falls short of the intelligence and creativity exhibited by humans.

4 Conclusions

Human-like performance by machines analysing and generating temporal structures requires the development of flexible models for hierarchical structure that can connect short-term and long-term aspects. The integration of sub-symbolic and symbolic, logic-based modelling may be required to capture the depth of human understanding of temporal structures and that music is an ideal testbed, as it is a specifically human phenomenon, lower dimensional than other domains such as video, and largely independent of semantic considerations and constraints.

References

1. Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., Klapuri, A.: Automatic music transcription: challenges and future directions. Journal of Intelligent Information Systems 41(3), 407–434 (2013)
2. Conklin, D.: 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. Citeseer (2003)
3. Dieleman, S., Schrauwen, B.: End-to-end learning for music audio. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6964–6968. IEEE (2014)
4. Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
5. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Transactions on Multimedia 13(2), 303–319 (2011)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
8. Google: Magenta (2016), https://magenta.tensorflow.org
9. Herremans, D., Chew, E.: Morpheus: Automatic music generation with recurrent pattern constraints and tension profiles. Tech. rep., Queen Mary University of London (2016), https://qmro.qmul.ac.uk/xmlui/handle/123456789/13599
10. Hiller, L., Isaacson, L.: McGraw-Hill, New York (1959)
11. J. Stephen Downie, Andreas F. Ehmann, M.B.M.C.J.: The music information retrieval evaluation exchange: Some observations and insights. In: Advances in Music Information Retrieval, pp. 93–115. Springer (2010)
12. Sigtia, S., Benetos, E., Boulanger-Lewandowski, N., Weyde, T., Garcez, A.S.d., Dixon, S.: A hybrid recurrent neural network for music transcription. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2061–2065. IEEE (2015)
13. Sturm, B.L.: Classification accuracy is not enough. Journal of Intelligent Information Systems 41, 371 (2013)