Classification of 1950 to 1960 Electronic Music Using the VGGish Neural Network and Random Forest

Maurício do V. M. da Costa[1], Florian Zwißler[1], Philip Schwarzbauer[1] and Michael Oehler[1]

[1] Music Technology & Digital Musicology Lab (MTDML), Institute for Musicology and Music Pedagogy, Osnabrück University, Germany
michael.oehler@uos.de

Abstract. This paper presents an approach to extend an ontological database concept aimed at the systematization of Electronic Music. Machine Learning techniques are used to test the significance of empirical investigations on the "output layer" of the production process, namely finished compositions of Electronic Music. As an example, pieces from the era of 1950 to 1960 are being examined, representing the aesthetics of *Musique Concrète* from Paris and *Elektronische Musik* from Cologne. The experiments performed using state-of-the-art techniques suggest the confirmation of measurable differences in the musical pieces from different studios for electronic music that were motivated by aesthetically divergent approaches.

Keywords: electronic music, musique concrète, Elektronische Musik, VGGish, random forest

1 Introduction

1.1 Analysis and systematization of Electronic Music

Despite Electronic Music having existed for many decades, it is still lacking tools to reliably systematize it, the most striking being a shortage of a clear terminology capable of describing the phenomena themselves as well as the processes used to produce them. In most cases, analogies to the strong and established terminologies of instrumental music and sound production [1-5] are being taken as a solution to this problem, not facing the problem that electronic sound production implies a fundamentally different potential that needs to be addressed [6]. This issue is continued in the field of music analysis: only a few attempts have been made to present universally valid tools that allow musicologists to get significant insights into the structure of a piece of Electronic Music. The most valuable source of information at hand is represented by [7, 8] and

the recently revised EMDoku[1], a huge database of Electronic Music that gives insights into all the results of composing with electronically produced sound. A systematization that comprises the conditions of the production of these results is yet to be found. Recently, this topic has received more attention, for example, with regard to *Musique Concrète* [9].

The PRESET research project, which was presented at CMMR 2019, has set out to do basic work to make progress in this direction: a database is being put together collating information from an in-depth survey of several studios for Electronic Music. Exploring their informational resources and bringing them together will open new lines of insight into the nature and relations of the processes involved. To address this issue, it was decided to use a semantic web database with an underlying ontology as a structural and terminological foundation [10]. In connection with the methods of actor-network theory [11] and theories from the field of information systems [12, 13], the working processes within the single studios as well as the connection in between them will display a new perspective on the field.

1.2 Electronic Music in the 1950s: *Musique Concrète* and *Elektronische Musik*

The early period of electronic music was characterized by a vivid debate between two quite different approaches of composing music within the context of an electronic studio. The *Musique Concrète,* which originated from Paris with its founder Pierre Schaeffer and since 1958 organised in the *Groupe de Recherches Musicales* (GRM), and the approach called *Elektronische Musik* (electronic music), which was pursued at the West German Radio in Cologne, most prominently represented by its then leader Herbert Eimert and Karlheinz Stockhausen. The *Musique Concrète* originally set out their experiments from recorded sound, thus integrating the production medium (records and, later on, magnetic tape) within the very first steps of working on sound. The repertoire of sound to create a piece was gained by very simple means of manipulation such as cutting the tape, reversing it, changing its speed, and building loops to generate rhythmic structures. This results in an empirical approach on dealing with sound as a medium to work on, also leading to an elaborate theoretical concept of the nature of sounds that Schaeffer formulated in his *Traité des objets musicaux* [14]. In Cologne, on the other hand, the idea was rather to construct the sound following a pre-structured concept devised by the composer. This strategy, in turn, was strongly connected to the concept of serial music, which favored a view on composition as a formal organization of sets of parameters [15]. It is evident that this view found a perfect fit in the new possibilities of sound creation and organization in an electronic studio.

These two approaches, of course, did not exist separately from one another, and there was a vital interest in each other's musical results. The opposing views on concepts of composition have been broadly discussed [16-18] and have led to the view that there was a remarkable aesthetic difference in these approaches.

Apart from the discussion to what extent this holds true, we decided to take the diverging concepts to an empirical test with the use of Machine Learning techniques.

[1] www.emdoku.de

2 Method

The methods presented in Section 1.1 basically represent a top-down model of systematization. Connecting our efforts to the existing potentials of databases such as EMDoku, we decided to add a bottom-up method of information retrieval in analyzing datasets representing the actual "output" of the studios in Cologne and Paris within a time window ranging from 1950 to 1960 – a period where the aesthetically divergent approaches were most prominent [18]. In doing so, we will try to test methods of empirical analysis and check them for their significance

The experiment consists in using a pre-trained Deep Neural Network (DNN) to convert the audio samples into semantically meaningful embeddings and then training a classifier to learn to identify material from both classes (GRM and WDR) using such high-level embeddings as input features. This way, we propose to empirically assess the existence of differences between recordings of such groups in purely acoustic features. Although this approach does not indicate what those differences are, we intend to pursue an indirect demonstration of their existence, for the only information provided for classification is related to audio content.

The VGGish [19] model was used for the computation of the embeddings. This network is based on the VGG [20] model, which is one of the most used DNN architectures for image recognition, and produces embeddings of 128 samples. In order to prepare the audio data to be processed by this network, first, the audio input signal is collapsed to mono and band-limited to 8 kHz. Then, its spectrogram is computed using the short-time Fourier transform, with a Hann analysis window of 25 ms and a hop of 10 ms. After that, a mel-spectrogram with 64 frequency bands (125 - 7500 Hz) is obtained by remapping the spectrogram time-frequency bins. This mel-spectrogram is then framed into non-overlapping examples of 0.985 s, each example covering the 64 mel bands and 96 time frames of 10 ms each. Finally, this process produces the embeddings for all audio files available by computing the network's outputs and stores them in text files with the same names as their audio counterparts.

Then, the random forest algorithm was used to classify the embeddings produced. To avoid having excerpts of the same musical piece both in the training and test sets by treating the embeddings as independent samples, all the embeddings of each piece were either assigned to the training set or to the test set. For this purpose, a random selection of the pieces was performed with a probability of 70% of each piece being selected as training data and 30% as test data. Since their variability in length is large (ranging from less than a minute to several minutes), considerable differences occur in the actual train/test proportion. This same classification experiment was repeated 10 times and both the average and the standard deviation of the results were computed to illustrate the classification performance. We used the implementation present in the "Scikit learn"[2] framework for the random forest algorithm, set to train an ensemble of 400 decision trees and using its default settings. Smaller number of trees were tested and provided slightly lower performance. Nevertheless, yielding high classification performance and providing a detailed analysis regarding the classification problem itself are not the objective of this paper.

[2] scikit-learn.org

In order to assess the performance, the majority vote of the embeddings within each musical piece was taken to assign the piece's classification. This way, each piece accounted for one sample, instead of their group of embeddings, i.e. pieces from which more than 50% of the embeddings were correctly estimated are considered to be one correctly estimated sample, despite its duration.

The actual lists of pieces used for the analysis were determined through the following: as a first step, all output from both studios within the chosen time interval was identified following the data resources provided by EMDoku, which represents the most reliable resource available. After that, only works that purely consist of electronically produced sounds were selected, thereby excluding all pieces that use sound resources from outside the production processes in discussion. It was also decided to exclude all sorts of functional compositions (e.g. music for radio plays) within those lists to again ensure the validity of the data as examples of the two aesthetic directions. The next step was to retrieve the actual audio material of the pieces. From the list of pieces from the studio of the WDR, it was possible to obtain about 75% of the pieces in question (57 files), making up a total duration of 3.5h. The examples available from the GRM made up a fairly larger amount, with 94 files, totaling roughly 6h of audio material.

It should be noted that we only compared the audio content of these pieces with no regard to spatialization, so from all the pieces, also those that exist in multichannel versions, only mono-mixdown versions were used, due to the characteristics of the architecture adopted for the classification task.

3 Results

The results obtained from this procedure are summarized in Table 1, which shows the average and standard deviation for accuracy, precision, recall and F-measure. Despite the small dataset available, the results suggest that the classifier was capable of identifying differences in the acoustic features related to each aesthetic approach.

Table 1. Overall results.

Measure	Average	Std.
Accuracy	0.82	0.08
Precision	0.89	0.08
Recall	0.66	0.20
F-measure	0.74	0.14

A histogram that represents classification accuracy of the embeddings within each musical piece, i.e. the proportion of correct votes for each class within each piece, is illustrated in Figure 1. As can be observed, the distributions obtained have different characteristics: the classifier was more successful in identifying excerpts from GRM, with voting proportion more concentrated towards 100% than from WDR, which had more diluted classification of the embeddings. In total, the GRM pieces were classified as 82% GRM and 18% WDR, whereas the WDR pieces were estimated to be 47% WDR and 53% GRM.

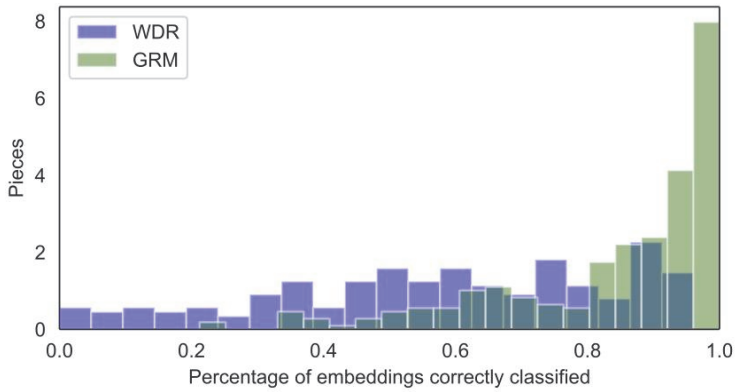

Fig. 1: Histogram of classification accuracy within each musical piece.

The distributions obtained suggest that the classes may have a significant overlap, as expected, but the classification system tended to have a bias towards the GRM class despite all data imbalancement compensation techniques. This may indicate that the GRM pieces might have less variation within the acoustic features of interest for the classification system, whereas the WDR pieces may show a wider variety in such dimensions. Besides, the pieces have a considerable amount of excerpts where the audio material present may severely interfere in this analysis, like background noise or long reverb tails. Nevertheless, the results are informative and serve the purpose of empirically assessing the differences present in sound.

4 Conclusion

The experiments presented in this paper served the initial goal to widen the focus of a database still under construction that aims at facilitating a valid and significant systematization of Electronic Music. The Machine Learning techniques employed to analyze the two specific sets of compositional results of studio work have displayed a specific difference within these sets. A possible consequence of this outcome in interaction with a future ontological database could be to check the technical equipment used within the specific time interval for correspondences and differences, as well as to investigate possible interdependencies of personnel involved. The inclusion of this “bottom up”- method is therefore likely to provide valuable insights and to bring up crucial questions to constantly improve the structure of the database as a whole.

The experimental setup was comprised of two different Machine Learning techniques: a pre-trained deep neural network (VGGish), which uses as input mel-spectrograms of the audio signal and outputs a sequence of high-level embeddings, followed by a random forest classifier, which was trained to differentiate embeddings from both classes under analysis. The musical pieces were then classified using the criterion of majority vote of the classes estimated for their embeddings. The train and test sets were randomly generated from piece selection and the experiment was performed 10 times. No embeddings from the same musical piece were used for both training and testing.

Although the results are not particularly outstanding for a music genre classification task (see Table 1), they do show that there are indeed noticeable differences in the acoustic features extracted from the pieces in these groups. This provides empirical evidence for what was only discussed theoretically in earlier studies.

It is worth mentioning that the VGGish network does not encompass long-term temporal interdependencies of acoustic events, which are a fundamental part of music structure and may reveal hidden patterns that could improve this intricate classification task. For this purpose, we intend to expand this experiment in future work tackling this specific problem by considering the whole sequence of embeddings using a different downstream model, instead of purely classifying each one independently, or even using a deep neural network that takes into account the temporal dimension.

References

1. Hornbostel, E. M. von, Sachs, C.: Systematik der Musikinstrumente. Ein Versuch. Zeitschrift für Ethnologie, 46: pp. 553–590 (1914).
2. Kartomi, M. J.: On Concepts and Classifications of Musical Instruments. The University of Chicago Press, Chicago (1990).
3. Simon, P.: Die Hornbostel/Sachs'sche Systematik der Musikinstrumente: Merkmalarten und Merkmale. Eine Analyse mit zwei Felderdiagrammen. Verlag Peter Simon, (2004).
4. MIMO – Musical Instrument Museums Online, *http://www.mimo-db.eu*, retr. 21/02/28.
5. Montagu, J.: Origins and Development of Musical Instruments. Scarecrow Press, (2007).
6. Kolozali, S., Barthet, M., Fazekas, G., Sandler, M. B.: Knowledge Representation Issues in Musical Instrument Ontology Design. In ISMIR: pp. 465-470 (2011).
7. Davies, H. (Ed.): Répertoire International des Musiques Electroacoustiques: International Electronic Music Catalog. Groupe de Recherches Musicales de l'ORTF (1967).
8. Hein, F., Seelig, T.: Internationale Dokumentation Elektroakustischer Musik. Pfau (1996).
9. Godøy, R. I.: Perceiving sound objects in the musique concrète. In: Frontiers in Psychology, 12, 1702 (2021).
10. Abdallah, S., Raimond, Y., Sandler, M.: An ontology-based approach to information management for music analysis systems. In: AES Convention 120 (2006).
11. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford (2005).
12. Goldkuhl, G.: Design Theories in Information Systems-a Need for Multi-Grounding. In: Journal of Information Technology Theory and Application (JITTA) 6(2): 7 (2004).
13. Gregor, S.: A Theory of Theories in Information Systems. In: Information Systems Foundations: Building the Theoretical Base: 1-20 (2002).
14. Schaeffer, P.: Traité des Objets Musicaux, Essai Interdisciplines. Le Seuil, Paris (1966)
15. Morawska-Buengeler, M.: Schwingende Elektronen. Tonger, Cologne (1988).
16. v. Blumröder, C.: Die elektroakustische Musik: Eine kompositorische Revolution und ihre Folgen. In: Signale aus Köln: Beiträge zur Musik der Zeit, vol. 22. Der Apfel, Wien (2017)
17. Frisius, R.: Musique concrete*. http://www.frisius.de/rudolf/texte/tx355.htm*, retr. 21/06/14.
18. Eimert, H., Humpert, H.U.: Das Lexikon der elektronischen Musik. Bosse (1973).
19. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Proc. of the Int. Conf. on Learning Representations, 2015, pp. 1–14 (2015).
20. Hershey, S., Chaudhuri, S., Ellis, D., Gemmeke, J., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R., Seybold, B., Slaney, M., Weiss, R., and Kevin Wilson: CNN Architectures for Large-Scale Audio Classification. In: Proceedings of ICASSP 2017, pp. 131-135 (2017).