TISMIR

Cella, C.-E. (2020). Music Information Retrieval and Contemporary Classical Music: A Successful Failure. *Transactions of the International Society for Music Information Retrieval,* 3(1), pp. 126–136. DOI: https://doi.org/10.5334/tismir.55

OVERVIEW ARTICLE

Music Information Retrieval and Contemporary Classical Music: A Successful Failure

Carmine-Emanuele Cella

This paper is about the story of my relationship, as a contemporary music composer, with computational tools that are situated in the areas of signal processing, machine learning and music information retrieval (MIR). I believe that sharing this story can be useful to the MIR community since it illustrates the problems that can arise when you try to use these techniques in the context of contemporary music creation. Since this is a personal story, I will refer to experiences that I had during about fifteen years of usage of MIR-related technologies. I will show how these technologies tried to (unsuccessfully) shape my musical thinking and why I believe that some of them have come to an end. Finally, I will propose new possible directions for the future of MIR.

Keywords: MIR; contemporary classical music; computational creativity; assisted composition

1. Introduction

My journey, as a composer, in the music information retrieval domain started about fifteen years ago and went through many related topics such as signal processing and machine learning. This paper is about the story of this journey and about how the technologies that I learned and used affected my musical creation. I hope that sharing this story with the MIR community will foster a reflection on the relationship between computers and musical creativity.[1]

1.1 Composers and computational tools

There is a close connection between music creation and computational tools (Fernández and Vico, 2013). Since the origins of Western music, composers are known to use computational support during musical writing; among some of the most notable examples are Mozart's dice games or Schoenberg's twelve-tone matrices. In the last few decades, with the development of computer technologies and the advances of artificial intelligence and machine learning, these techniques have become an important area of research and development (Briot et al., 2019) and have converged towards the field of *computational creativity* (Wiggins et al., 2009; Cardoso et al., 2009).

Music information retrieval (MIR), while initially focused on extracting high-level information from music, has also evolved (among other directions) towards the development of computational tools to support music creativity (Humphrey et al., 2013; Andersen and Knees, 2016). Within this context there is an active area of research, particularly useful for composers, known as *Computer-Aided Composition* (Deserno, 2015) that focuses on the development of software to assist musical composition.

In the next sections, I will show how these tools and technologies impacted my life as a composer and why I believe it is important to keep a close connection between contemporary classical music and MIR research.

2. A Composer's Perspective

I like to think about myself as a music composer: a person whose main activity is writing music with paper and pencil. However, being a *contemporary* music composer with a scientific background, I have had, in the past fifteen years or so, long-term relationships with signal processing, machine learning and music information retrieval (Cella, 2011a).

The main needs that motivated these relationships are, roughly speaking, three:

- I wanted to use computers to better understand my music and the music of other contemporary composers: often times, I am not totally aware of the processes that I put in place when I write music;
- I wanted some assistance from computers to accelerate and simplify my own compositional work;
- I wanted to use computers to create new interfaces to express and control my musical processes.

These needs correspond, to some degree, to my personal interpretation of *analysis*, *synthesis* and *human-computer interaction* (HCI) respectively. In order to clarify what I mean by these concepts, I need to give some preliminary information about the context in which I apply them: *Western contemporary classical music.*

CNMAT/Music, University of California, Berkeley, CA, US
carmine.cella@berkeley.edu

2.1 On contemporary classical music

Although giving an exhaustive definition of Western contemporary music is beyond the scope of this paper,[2] it will be useful to provide some information of what is my understanding of it (as a composer, not as a musicologist).

Western contemporary classical music can be thought as the music created in the past fifty/sixty years, mostly in Europe and the US, by a number of composers such as (in no particular order): Pierre Boulez, Gyorgy Ligeti, Karlheinz Stockhausen, Milton Babbitt, Helmut Lachenmann, Luciano Berio, Pauline Oliveros, Linda Bouchard, Kaija Saariaho, Stefano Gervasoni, Yan Maresz, Gerard Grisey, Jonathan Harvey, Chaya Czernowin, Rebecca Saunders, Edmund Campion or Olga Neuwirth, just to name a few.

It is a heterogeneous world that gravitates around a community of classically-trained composers who have interests in sound-design, computer music, mixed music (instruments with electronics) and extended instrumental techniques. While this community is fairly small if compared to contemporary pop music, it has quite an active life and it motivates a good amount of research, especially in France (Vinet, 2008). Generally speaking, contemporary classical music belongs to the classical tradition (as opposed to pop or jazz traditions, for example) and is often characterised by some of the following features.

- **Heterogeneity of languages.** Each composer defines his or her own poetics, and that is typically very different from other composers' poetics (much more than the difference between Bach's and Beethoven's, for example). In some cases, this heterogeneity can be so extreme that no common semantics or even techniques can be found.
- **Different interpretation of technical elements.** Given the large variability of contemporary musical languages, the employed technical elements also have very different interpretations. The concept of harmony for Boulez, for example, can differ very much from the same concept in Lachenmann's or Ligeti's music. Along the same lines, the technical interpretation of chord, melody, rhythm and form vary a great deal among composers.
- **Multiplicity of instrumental techniques.** Each composer creates new instrumental techniques for classical instruments (such as violin or flute) and occasionally even creates new instruments. These techniques (sometimes called *extended*) can vary a lot among composers, ranging from noise to silence, and their notation on the score is not standardised for the most part. The tuning of instruments can also be altered (for example, to be different from the tempered system) and the same concept of *pitch* can be redefined.
- **Electronic medium.** A part of what I consider contemporary classical music is made exclusively with electronic sounds, without physical instruments. Examples of this kind of music are J. Chowning's Stria (1977) or Turenas (1988), B. Truax's Wave Edge (1983) or Riverrun (1986), J. C. Risset's Sud (1985) and many others. This kind of music is essentially defined in the timbre space and concepts such as harmony and rhythm assume a very different meaning.

My work, as a composer, happens within this kind of artistic and cultural environment.

2.2 On analysis, synthesis and HCI

Now that the context of my work is hopefully more clear, I can try to define better what I mean by analysis, synthesis and human-computer interaction.

2.2.1 Analysis

When I need to analyse contemporary music I could not care less if a piece is in *C* major, if at bar seven there is a chord of F^7, what is the BPM or where is the chorus. I do not care simply because, in contemporary classical music, these concepts **do not exist** or have a more generalised interpretation. What I care, instead, is understanding the hierarchical structure of *musical elements*, their relative importance, their harmonic layout (not made of major or minor chords but of complex aggregates of pitches that can be also defined outside the tempered system), their density or their temporal evolution.

For musical elements, here, I mean a generalisation of the typical concepts found in pop music: for example, melodies become streams of sounds, chords become spectra, rhythms become textures and so on. For the lack of a better expression I will call these elements **sculptures of timbre**,[3] since they are directly defined in the timbre space (McAdams and Giordano, 2016). The idea of sculptures of timbre is, to some extent, close to Smalley's concept of *spectromorphology*: the description of perceived morphological evolutions of sound spectra over time (Smalley, 1997). Smalley's idea, however, is more a descriptive analysis of existing sounds. Sculptures of timbre, on the other hand, can also have a *prescriptive* interpretation and can be actively used during the compositional process. In either case, both concepts are directly defined in the timbre space.

For this reason, I have always thought that in order to analyse contemporary music we need to start from *raw* signals (such as the recording of an orchestral piece) and not from symbolic representations (such as scores or MIDI files) since many of the things I look for are only defined there. This is why I needed MIR techniques in my work.

The difficulty of analysing music in the timbre space and the potential of MIR techniques for this task have also been discussed by Klien et al. (2012).

2.2.2 Synthesis

My interpretation of synthesis is, instead, twofold. From one side, I mean all the usual techniques to generate sound with computers, such as FM synthesis, granular synthesis and any techniques of sound modification (digital audio effects, advanced transformations and so on).

From another side, however, I mean the support that computers can give to composers during the *act* of composing. These techniques are often grouped under the expression *computer assisted composition* (CAC). Extensive research has been devoted to CAC, especially at

IRCAM (Paris), and pioneering and powerful software has been produced (Assayag et al., 1999).

However, most of this research focused on the symbolic level and provided support to manipulate pitch classes or rhythms in a purely quantitative way (for example, to generate the transposition of a chord by *n* semitones, or to variate a rhythm by adding durations in specific positions). Given my interest in musical elements as sculptures of timbre, however, I always looked for tools that could help me in a more *qualitative* way in order to define semantic transformations in the timbre space. An example of such a type of tool is the project *Orchidée* (Carpentier et al., 2007). In this work, the authors create a mapping between the physical world of timbre and the symbolic world of scores by creating an *assisted orchestration* of a target sound; later sections will discuss this in more detail.

2.2.3 Human-computer interaction

In order to use computational tools in my creative work, I need to situate them in my compositional process. To do this, I need to be able to design an expressive interface that handles my musical processes in a convenient way.

The key feature that such interface must have is, for me, **composability**: when I design a new synthesis algorithm, for example (in either one of the two meanings given above) I need to use it in a complex composition process, where I organise time, hierarchy and form. As such, these tools must be able to become part of a larger system as modular and programmable components.

A powerful example of composability can be found in the dual nature of the scripts for sound synthesis in *Csound*, a music programming language originally conceived at MIT by Barry Vercoe and based on the MUSIC-N family of languages (Boulanger, 2000). In Csound, the code is split into two different conceptual blocks: the *orchestra* and the *score*. The former is dedicated to the creation of the algorithms to be used for synthesis (called *instruments* in Csound jargon), while the latter is used to define the temporal behaviour of such algorithms. In other words, with the score the user can *compose* the algorithms in time. In the more recent iteration of the language, moreover, the user can interact in real time with the algorithms to change their behavior by means of different interfaces.

2.3 Why contemporary music is important for the MIR community

After the discussion given above, the reader may be left under the impression that the problems and the perspective I discussed are only relevant for me. Moreover, since contemporary classical music is a *niche* compared to other types of music, it could seem that the MIR community should not be interested in it.

I want to advocate, instead, that the perspective I presented transcends my own work and that it is very important that the MIR community keeps a close relationship with composers of contemporary classical music. I will present two main motivations to support my statement.

- **Generality.** Several of the concepts that I presented when I discussed analysis and synthesis are actually shared by many contemporary composers. The idea of sculptures of timbre, for example, is general and expressive enough to be used for the music of very representative composers such as Ligeti, Xenakis or Lachenmann. The same idea of using computers to help the composition process, moreover, has motivated a lot of research that has converged in a new area called computer-aided composition. Several institutions around the world, among which IRCAM in Paris, the MIT MediaLab in Boston or the Center for New Music and Audio Technologies (CNMAT) in Berkeley, to name a few, contributed in a significative manner to this type of research.
- **Mutual exchange.** While the amount of music produced in the contemporary classical context is smaller than the amount produced in other contexts, contemporary music presents new challenging problems to the MIR community that can foster the development of new research. In other words, if the community of contemporary classical composers would certainly benefit from the interaction with the MIR community, the converse is also true. This will be more clear in the following sections.

Hence, to put the discussion into a more general perspective, it is important to understand that my personal needs as a composer are actually grounded on fundamental underlying questions that have been of primary importance for several communities (Vinet, 2003):

- "Which kind of connections can we establish between the physical space of *timbre* and the symbolic space of musical scores?" (McAdams, 1999).
- "Is it possible to formalise the creative process followed by composers?"
- "What is a valid methodology to establish a collaboration between researchers and artists?"
- "How is it possible to artistically evaluate the outcomes of MIR research?"

Despite an impressive amount of work done in the past twenty years, I believe that these questions remain still open.

It must be noted, finally, that the research motivated by contemporary music is done across multiple domains and multiple communities. From one side there is the computer-music community, principally made by artists and musicians with some scientific background, interested in creative applications of technology. This community gathers around several specific conferences such as the International Computer Music Conference (ICMC, http://www.computermusic.org). From another side, there is the MIR community, principally made of researchers and engineers who are more interested in problems related to signal processing, classification and analysis. The reference conference for this community is the International Society for Music Information Retrieval Conference (ISMIR, https://ismir.net). In between, there are communities more focused on the human factor, that gather around conferences such as the New Interfaces for

Musical Expression conference (NIME, https://www.nime.org). As we will see, each community has different needs and often requires different theoretical infrastructures.

The MIR community, being concerned with labels and categories as a way of defining musical information, was my main reference for problems related to analysis. The computer-music community, on the other hand, was my primary reference for all questions related to synthesis and musical writing. Finally, the NIME community, whose focus is more on communication and transfer of information (mapping), was important for my idea of control of musical processes.

As a contemporary composer, I had to position myself in between this heterogeneous world (**Figure 1**). I passed several years trying to understand and use methods developed by people belonging to different communities and trying to develop my own to share with other composers.

The following sections will present an overview of these efforts, showing to which extent specific techniques impacted my musical thinking.

3. Low-Level Features

My first professional contact with state-of-the-art MIR techniques happened at IRCAM in 2007. In the analysis-synthesis team, I had the chance to work with gifted researchers to develop *IrcamDescriptor* (Burred et al., 2008), a library for real time computation of low-level features (Peeters, 2004).

It was a common belief, at the time, that these kinds of descriptions for sound were possibly the way to go to analyse and represent contemporary music (one of my long-standing problems). I was working under the assumption that, in order to produce more conceptual and deep descriptions of music, I could aggregate low-level features and climb up the ladder of abstraction through a sequence of intermediate representations designed to exploit specific properties of musical timbre.

I tried, for example, to analyse the complex structure of some contemporary compositions (and to classify the musical elements present therein) by combining features such as spectral centroids, spectral spreads and MFCCs. In the same way, I designed systems that were able control the synthesis and the transformation of sounds for my own composition, by reacting to specific perceptual features within a paradigm that I used to call *synthesis by analysis*. The next section will clarify this concept.

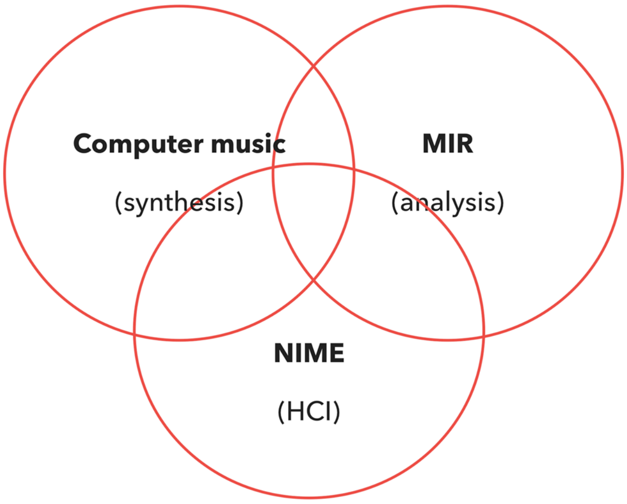

Figure 1: The heterogeneous world of the different communities interested in the study of musical signals.

3.1 A personal contribution

I believe that my main technical contribution, in this context, is a representation framework for music and sound called *sound-types* (Cella, 2011b): the idea behind this is to be able to describe raw signals with a pyramid of representations with increasing level of abstraction. I used this framework myself for a musical production at IRCAM. In 2013, I was commissioned for a piece for large orchestra and live electronics by Orchestre Philharmonique de Radio France; I used sound-types to *listen* to the orchestra in real time during the performance and *learn* classes of sounds that I used later in the piece to generate in real time sound hybridisations at abstraction levels higher than the signal (Cella and Burred, 2013). I think that this example of using sound-types for musical creation illustrates what I mean by analysis, synthesis and human-computer interaction:

- music is analysed as a raw signal by machine (in this case, the real time performance of the orchestra);
- some high-level concepts are produced (in this case, a hierarchy of classes);
- the machine, using the information acquired during the analysis stage and responding to some human control, generates back some signals that represent high-level semantic transformations of the original content.

While this work has been used in other major musical productions, it did not have an impact on the MIR community. One interesting feature of sound-types was the layered architecture designed to represent different abstraction levels, similar in the intention to deep learning networks (section 4). In retrospect, however, I think that one of the reasons that prevented the diffusion of the sound-type representation framework within the MIR community was the lack of an appropriate language. I had not been able, at the time, to find a correct communication with other researchers. This difficulty in communication generated several methodological problems.

3.2 Methodological problems

I eventually realised that my problems were somehow ill-posed for the MIR community. Around 2008, an important part of the community was in fact focusing on different types of music than contemporary music, such as pop. I will refer to this as the *standard MIR approach*: I am aware that there is non-pop related MIR research that is really interesting, such as the *CompMusic* project led by MTG that focuses on non-Western music traditions,[4] but here I refer to the attitude that the majority of the community had around that time. This is not a bad thing *per se*, but it generated two main methodological problems for me.

For one, trying to convey information about a specific music that was not mainstream (or sufficiently known) was painfully difficult. Each time I tried to discuss with a researcher in MIR about the problems that I wanted to tackle, I felt that we were not really understanding each other. In my music, for example, the concepts of rhythm, chord and melody simply do not exist and are replaced by more general (and often fuzzy) categories such as density or texture; just to explain this fact could literally take days of work. This was no one's fault: the way of thinking that stands behind musical composition is pretty far from the mindset of an engineer trying to analyse a signal. It is therefore understandable that trying to define a common language is not easy and can require major efforts.

The other problem generated by the fact that the MIR community was focusing on pop music was, instead, more scientific. A typical strategy in standard MIR to tackle a problem is by simplifying it in order to develop techniques to be subsequently transferred to a more complex case (Müller, 2015). Unfortunately, the developed techniques often *make assumptions* about the problem itself and cannot really be transferred to other cases. In other words, does it make sense to analyse the Beatles corpus to develop techniques to be used later on other types of music such as, for example, Stockhausen's music?

Generally speaking, I felt that state-of-the-art research was not applied to many important musical problems that were specifically present in contemporary classical music: as such, I started thinking that including more contemporary classical music into MIR would have helped in extending the scope of the research. The problem of automatic chord estimation (*ACE*) will serve as an example.

The common two-step feature extraction/classification scheme for ACE has been successful for pop music (Papadopoulos and Peeters, 2007, 2011) but it is very difficult to transfer the adopted techniques to other contexts. Beat-synchronous features do not always make sense in the context of contemporary classical music, in which most of the language develops around different concepts of time. Similarly, chroma features are not really useful when analysing different tuning systems, often used in electronic music.

For these reasons, while low-level features proved to be very effective in many MIR-related problems, I did not manage to use them efficiently to support my creative work. Similar issues in using MIR techniques for representing musical information have been also discussed by Wiggins in a more philosophical manner (Wiggins, 2009).

4. The Deep Revolution

The situation changed, for me, a few years later with the massive reintroduction of neural networks in machine learning.

Neural networks have been studied since 1950 and have gone through alternate phases (Nilsson, 2009). In 2012, however, a new type of artificial neural network with many convolutional layers (hence *deep*) outperformed other algorithms in the ImageNet competition (Krizhevsky et al., 2012): the result was so impressive that it attracted the attention of the scientific community immediately.

4.1 A neural zoo

Deep networks provide state-of-the-art results in many problems (LeCun et al., 2015). Their general architecture, made of a cascade of linear filters and non-linearities, usually outperforms hand-crafted features.

The artificial vision community reacted immediately to this stimulus and produced an enormous amount of research based on similar architectures (Srinivas et al., 2016): a whole *zoo* of networks developed,[5] going from the 8 layers of AlexNet to the incredibly deep *ResNet* with more than 150 layers (He et al., 2016).

The audio and MIR communities also reacted to these results. While there have been early examples of neural networks in the MIREX[6] contest (Lacoste and Eck, 2005), it is only between 2012 and 2014 that the consistent adoption of these tools started to develop (Schlüter and Böck, 2014). Humphrey et al. (2015) presented an important overview of deep learning techniques for music informatics.

I believe that an important advancement for the MIR community happened when time started to be considered into these architectures: recurrent models such as LSTM (long-short term memory) networks and RNNs started to be used to model time series and evolving signals (Sak et al., 2014; Choi et al., 2017).

While these approaches were designed to analyse signals, deep architectures have also been proposed to synthesise them. The work done by Dieleman and Schrauwen (2014) and the work done at DeepMind on *WaveNet* (van den Oord et al., 2016) with the subsequent *SampleRNN* model (Mehri et al., 2015) opened the way to the so called *end-to-end* architectures: complete frameworks for analysis and synthesis of signals starting from raw data.

This was an important step towards my own idea of analysis and synthesis: I could now train these machines to learn the high-level concept of sculptures of timbre and I could use them to support my creative work.

4.2 Problems in representing musical knowledge

Undoubtedly, from a composer's perspective, these techniques represented good progress: we now had the possibility to use learning algorithms that could generalise better on the salient elements of contemporary classical music. Some remarkable artistic work has been produced using this technology. In *La fabrique des monstres* (Ghisi, 2017), for example, the composer explores the possibilities of a musical machine able to learn from different corpora of sounds and creatively reproduce the learned patterns using SampleRNN-like networks.

These complex machines, however, bring us to a very difficult and partially unknown world (Mallat, 2016). The problem of representing musical knowledge with these architectures can be daunting and, sometimes, intractable (Cella, 2017). For a few years (roughly from 2015 to 2017) I was constantly in this situation: I wanted to use neural networks for creative musical purposes, but I did not really manage to train these machines for my specific problems. Even when I managed to train them to some extent, I did not know how to interpret the results. I tried to communicate with other researchers or composers (with all the possible difficulties that I outlined in previous sections), but I was not aware of many other persons working on similar issues.

At this stage, I was asking myself questions like: "What do the learned features represent?" or "How can I constrain these features to a specific musical meaning?" and so on.

4.2.1 Supervised versus unsupervised

A particular problem that I found in my own musical applications of deep learning has been the difficulty of using supervised learning. Supervised learning can be thought of as the learning of a high-dimensional function $y = f(x)$ which maps some input features x to the respective output labels y. Unsupervised training, instead, is not based on labels but it tries to find topologies or patterns from raw data. Convolutional neural networks are a good example of supervised learning while auto-encoders, typically, implement unsupervised learning.

To be able to apply supervised learning, you need a database with labels: this can be a serious issue in contemporary music. Often times, for the reasons discussed in section 2.1, there are *conceptual disagreements* in contemporary music and it is very difficult to assign labels to musical signals given their inherent ambiguity. It is possible, indeed, that if you ask two different composers to analyse a passage of contemporary music, they will give somewhat different interpretations. Moreover, copyright issues can make it difficult to deploy such a database to be used for reproducible research.

For such reasons, it appeared to me that the best approach for my problems was unsupervised learning with auto-encoders and their variants. I experimented with them for a while (Gabrielli et al., 2018) and even if I was really fascinated by these unsupervised architectures from a scientific standpoint, I never managed to use them for creation given the unsatisfying results I obtained.

4.3 A methodological conundrum

At this point, even if I was clearly aware of the potential of deep neural networks to tackle my problems, the difficulty in handling musical knowledge and the generally unknown world in which they were bringing us put me off. From one side, there was a plethora of theories and techniques such as Fourier transform, hand-crafted low-level features, MFCC, chroma vectors, etc. that were very well understood and had had important applications for MIR that were ultimately unable to tackle my original problems. From another side, instead, there was a zoo of neural networks and unintelligible machine-crafted features with a huge potential both for analysis and synthesis that we did not understand.

As far as I knew, most of the research in this context was not done to understand neural networks better but rather to find new architectures for specific problems. In my eyes, part of the research methodology had been shaped to something different after the new AI wave: paradoxically, the usual pattern *hypothesis-experiment-conclusion* had become something like *data-experiment-more data*.

Everybody was looking for answers in the data, but how did we know that for all problems the data *contained* the answer? From a musical creation point of view, it did not look right; unfortunately, it was very difficult for me to explain why.

5. Convergences

It was at this point that I started a post-doc position at École normale supérieure (ENS), in Paris, in the *DATA* team. Among the topics studied in this group, there was the investigation of mathematical models of deep networks. One of the pillars of such investigation is the work done by Stéphane Mallat on invariant representations, culminating in the definition of the *scattering transform* (Mallat, 2012): a cascade of wavelet transforms and modulus operators that is spatially averaged. This representation builds local invariance to translation thanks to the averaging but preserves discriminative capacities thanks to the layererd architecture. Scattering networks are a type of convolutional network whose filters are defined as wavelets and not with a learning algorithm: these networks have solid mathematical foundations and are well understood (Bruna and Mallat, 2013), as opposed to other deep networks.

Before the deep network revolution, the scattering transform was having state-of-the-art results in many high-dimensional problems such as handwritten digit classification (Bruna and Mallat, 2013), sound classification (Andén and Mallat, 2014), texture analysis (Sifre and Mallat, 2013) and quantum energy regression (Hirn et al., 2017). I believed (and still believe) that the scattering transform was the best representation that we can possibly have without learning.

This experience changed my approach on signal analysis, machine learning and, ultimately, MIR. I had found a group of people that were interested in problems similar to mine: for example, they also wanted to figure out *why* neural networks work and not only *how*. Working with them, I could find a way to connect the well defined world of Fourier and hand-crafted features with the confusing (to me) zoo of deep networks. Was my methodological conundrum (section 4.3) close to an end?

5.1 Results versus knowledge

With the help of a colleague, I worked on a problem that was interesting both musically and scientifically: musical instrument recognition. We created a classification system made by a convolutional neural network based on a known acoustic model (Lostanlen and Cella, 2016). We tried different weight sharing strategies for deep convolutional networks and we provided an acoustical interpretation of these strategies within the source-filter framework of quasi-harmonic sounds with a fixed spectral envelope, typical of musical notes.

I believe that the key point of our work is indeed that we managed to make such interpretation thanks to a *previously known* model. The knowledge that we can acquire on a problem by using known techniques is indeed very useful. While hand-crafted features are outperformed by learned features, the former provide an understanding level of the problem that is still very valuable. The idea that we should create end-to-end systems that magically provide answers is a mistake, in my opinion.

When I look at the design of MFCC, for example, I am always amazed. The researchers that managed to create such a representation, put in it years of experience and an incredible amount of knowledge on the problem they were solving (Mermelstein 1976; Davis and Mermelstein, 1980).

I think that we should never allow an algorithm, however powerful, to replace our knowledge; we should instead *integrate* it with our knowledge. The result alone is never the answer, the acquired knowledge is (section 5.3).

5.2 Representations and communities

One important thing I realised during my work at ENS is how representations and communities are related. When we design a representation for a problem, we actually focus on different *mathematical properties* (Mallat, 2012), depending on the nature of the problem. Since different communities have different problems, they look for different properties.

The MIR community is often interested in *invariance* and *unicity*: in cover song detection, for example, we want our representation to be insensitive to specific transformations (such as different performances of the song) but at the same time we want to discriminate between two different classes (two different songs). For the NIME community, instead, *stability* can be very useful: in gesture detection, we need a representation that does not vary too much if the gesture changes just a little; this is essential in order to produce a consistent interaction model (mapping). The computer-music community is often times interested in *invertibility*: since the possibility of generating sounds is essential, we need a representation to be converted back to a signal. This allows us to transform the content in the semantic level of the representation and produce a new sound.

The mathematical properties of representations are fundamental entities to define the identity of each community. Each community uses all of these properties to some extent and a researcher or a composer must understand them correctly in order to connect with that community. Understanding this has been really important for me as a composer, since it has been the key to better communicate my needs.

5.3 Logic meets learning

In the last few years, I had the privilege to work on one of the projects that I always considered one of the best examples of connection between signals and symbols: computer-assisted orchestration (Cella and Esling, 2018). Target-based assisted orchestration can be thought of as the process of searching for optimal combinations of sounds to match a target sound, given a similarity metric and a set of musical constraints. A solution to this problem is a proposed orchestral score that maximises the similarity, in some feature space, between a target sound and the mixture of audio samples corresponding to the notes in the score. Assisted orchestration remains relatively unexplored because of its high complexity, requiring knowledge and understanding of both mathematical formalisation and musical writing (Maresz, 2013).

The current model is made of two principal components: a combinatorial optimisation algorithm (that computes and evaluates a large number of combinations of musical notes from a database) and a constraint solver (that assures the satisfaction of musical rules). The large number of candidate combinations,[7] the difficulty of creating a good embedding space for timbre (Carpentier et al., 2010) and the complex and sometimes contradictory constraints make this problem fairly difficult but very interesting for a contemporary composer.

While I still do not have a complete solution to this problem, I know today that we cannot solve such a problem with learning algorithms only. The complex nature of musical rules in orchestration (which instruments must be used, which playing styles or dynamics are allowed, which combinations are not musically possible, what is physically playable by a musician, etc.) does not fit easily within a statistical learning framework and requires articulated *logical rules* capable to capture such complexity. I believe that music cannot be simplified to basic probabilistic models since it is intrinsically made of an articulated hierarchy of musical elements that evolve over time. This hierarchy represents high-level relationships and can only be modelled by relational operators, usually provided by logic descriptions. Often times, the kind of problems that I have during the creation and the analysis of contemporary music cannot be solved by analysing multiple instances of something. Often, these problems appear only *once*: the importance of a musical element in a composition does not only depend on how many times it occurs.

While in my current approach on the problem of computer-assisted orchestration I use a number of different techniques together, among which are neural networks (Gillick et al., 2019), I believe that a more general representation framework, able to integrate statistical learning and logical rules, would bring great benefit to this research. A possible candidate for such a framework is *statistical relational artificial intelligence* (StarAI), where learning and logical rules coexist. I discussed this idea, together with another researcher in MIR, in a recent perspective paper that focuses on the potential that StarAI has for modelling complex musical problems (Crayencour and Cella, 2019).

5.3.1 Examples versus rules

In a recent talk given at the *Brains, Minds and Machines* symposium in occasion of MIT's 150th birthday party, Noam Chomsky has been reported to have derided researchers in machine learning who only use statistical methods to model behaviours but do not try to understand the *meaning* of those behaviours.[8] Professor Chomsky thinks that statistical models have had engineering success but are irrelevant to science as long they do not provide insights. I also believe that engineering success is not a good measure for the progress of science; however, science and engineering develop together and the former often provides evidence to the latter. If we want to tackle state-of-the-art musical problems we must take into account the complex relational nature of musical elements and we cannot simplify musical problems to statistical models alone: examples and rules should be *integrated* in a common language that provides insights on our problem.

6. Where to Go from Here

This paper discussed the story of my relationship with music information retrieval and related technologies in about fifteen years of research and creation. I hope that sharing this story will foster a productive discussion in the MIR community.

During these years I have had many fundamental insights that made me grow as an artist. I know now that while hand-crafted features are outperformed by deep learning, they provide important insights on the problem that we are trying to solve. I know that, in order to communicate with a research community, I need to correctly understand and define the mathematical properties of my representations. Finally, I know that statistical learning alone, however deep, is not enough to solve problems in state-of-the-art contemporary music and we also need logical rules to model high-level relationships between musical elements.

Deep learning made us forget many things that we learned in the past years. The desire of finding end-to-end machines that magically solve all our problems brought us on the wrong path. My statement, here, is that deep learning intended as a purely brute-force statistical learning strategy is coming to an end. We will soon realise that many of the results that we have had in recent years did not actually bring any new knowledge about our problems. This is why I believe that we should expand MIR by including a different set of features.

MIR research should increase its focus on contemporary classical music. By **keeping contemporary composers in the loop** and trying to answer their questions, MIR will manage to create significant insights on the nature of music.

The amazing interaction between the contemporary composer Pierre Boulez and the gifted researcher Giuseppe Di Giugno created a brand new world in which state-of-the-art technology managed to solve problems presented by state-of-the-art music (Lipp, 1996). This interaction is at the origin of the creation of IRCAM and, ultimately, provided essential contributions in the creation of the MIR field itself. The reason for which MIR has focused more on pop music is not only practical (easier to analyse), but also economical. Most MIR research projects are funded by institutions and companies that hope to make music navigation tools for a large audience; as such, databases are made of greatest hits. This imposes heavy structural constraints to the models in a vicious cycle that we could break by creating new branches of MIR closely related to contemporary music creation.

MIR, moreover, should integrate mathematical representations that we understand, large-scale deep learning algorithms and logical rules:

- the mathematical properties of our representations must be appropriate to the problem we are solving;
- hand-crafted features must be used to help our intuition while modelling problems;
- deep learning must be used as a powerful tool that works *inside* a larger context framed by our knowledge;
- complex musical rules must be handled at a *logical level* with a formal language that is expressive enough to model them.

In order to create such integration we need a powerful conceptual infrastructure. I believe that StarAI, where statistical learning and logical rules coexist, is a good candidate to build this integration but other ways are possible too. In addition to logic, for example, another important element is *memory*. There are already efforts in incorporating memory into deep learning and there are promising results (Ycart and Benetos, 2017).

I have great respect for the MIR community and for the amazing successes obtained in the last twenty or more years. From a purely musical standpoint, however, we are still far from what I consider to be a real support to musical creativity. I believe that focusing on the complex world of contemporary classical music could foster, in this sense, important scientific advancements.

Notes

[1] This paper is in a similar vein of the talk given at ISMIR 2018 in Paris by Rebecca Fiebrinck (a recording of which is available on YouTube: https://www.youtube.com/watch?v=QflI-ewRJ6o, accessed on June 30, 2020), but focuses more on the needs of a person who wants to write music on *paper* and not necessarily create it in real time.

[2] See, for more information, the work of Donin and Feneyrou (2017).

[3] A German word that has a similar meaning is *Klangbild*.

[4] For more information see https://compmusic.upf.edu, accessed on June 30, 2020.

[5] For more information see http://www.asimovinstitute.org/neural-network-zoo/, accessed on June 30, 2020.

[6] For more information see https://www.music-ir.org/mirex/wiki/MIREX_HOME, accessed on June 30, 2020.

[7] Typically, this number is of the order of 2^{30000} (power set of all possible instruments, notes, dynamics and playing styles) making the problem of assisted orchestration NP-complete.

[8] A transcript of his intervention can be found here: http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html, accessed on June 30, 2020.

Competing Interests

The author has no competing interests to declare.

References

Andén, J., & **Mallat, S.** (2014). Deep scattering spectrum. *IEEE Transactions on Signal Processing*, *62*(16), 4114–4128. DOI: https://doi.org/10.1109/TSP.2014.2326991

Andersen, K., & **Knees, P.** (2016). Conversations with expert users in music retrieval and research challenges for creative MIR. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 122–128.

Assayag, G., **Rueda, C.**, **Laurson, M.**, **Agon, C.**, & **Delerue, O.** (1999). Computer-assisted composition at IRCAM: From PatchWork to OpenMusic. *Computer Music Journal*, *23*(3), 59–72. DOI: https://doi.org/10.1162/014892699559896

Boulanger, R. C., editor. (2000). *The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming*. MIT Press.

Briot, J.-P., **Hadjeres, G.**, & **Pachet, F.** (2019). *Deep Learning Techniques for Music Generation*. Computational Synthesis and Creative Systems Series. Springer Verlag. DOI: https://doi.org/10.1007/978-3-319-70163-9

Bruna, J., & **Mallat, S.** (2013). Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1872–1886. DOI: https://doi.org/10.1109/TPAMI.2012.230

Burred, J. J., **Cella, C.-E.**, **Peeters, G.**, **Roebel, A.**, & **Schwarz, D.** (2008). Using the SDIF sound description interchange format for audio features. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 427–432.

Cardoso, A., **Veale, T.**, & **Wiggins, G. A.** (2009). Converging on the divergent: The history (and future) of the International Joint Workshops in Computational Creativity. *AI Magazine*, *30*(3), 15–22. DOI: https://doi.org/10.1609/aimag.v30i3.2252

Carpentier, G., **Tardieu, D.**, **Assayag, G.**, & **Saint-James, E.** (2007). An evolutionary approach to computer-aided orchestration. In M. Giacobini, editors, *Applications of Evolutionary Computing: EvoWorkshops 2007*, volume 4448, pages 488–497. Springer. DOI: https://doi.org/10.1007/978-3-540-71805-5_54

Carpentier, G., **Tardieu, D.**, **Harvey, J.**, **Assayag, G.**, & **Saint-James, E.** (2010). Predicting timbre features of instrument sound combinations: Application to automatic orchestration. *Journal of New Music Research*, *39*(1), 47–61. DOI: https://doi.org/10.1080/09298210903581566

Cella, C.-E. (2011a). *On symbolic representations of music*. PhD thesis, University of Bologna.

Cella, C.-E. (2011b). Sound-types: A new framework for symbolic sound analysis and synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 179–184.

Cella, C.-E. (2017). Machine listening intelligence. In *Proceedings of the International Workshop on Deep Learning for Music*, pages 50–55.

Cella, C.-E., & **Burred, J. J.** (2013). Advanced sound hybridizations by means of the theory of soundtypes. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 39–46.

Cella, C.-E., & **Esling, P.** (2018). Open-source modular toolbox for computer-aided orchestration. In *Proceedings of Timbre 2018: Timbre is a Many-Splendored Thing*, pages 93–94.

Choi, K., **Fazekas, G.**, **Sandler, M.**, & **Cho, K.** (2017). Convolutional recurrent neural networks for music classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396. DOI: https://doi.org/10.1109/ICASSP.2017.7952585

Crayencour, H.-C., & **Cella, C.-E.** (2019). Learning, probability and logic: Toward a unified approach for content-based music information retrieval. *Frontiers in Digital Humanities*, *6*(6). DOI: https://doi.org/10.3389/fdigh.2019.00006

Davis, S. B., & **Mermelstein, P.** (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. DOI: https://doi.org/10.1109/TASSP.1980.1163420

Deserno, S. (2015). Algorithmic composition: An overview of the field, inspired by a criticism of its methods. Seminar topics in computer music, RWTH Aachen University.

Dieleman, S., & **Schrauwen, B.** (2014). End-to-end learning for music audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968. DOI: https://doi.org/10.1109/ICASSP.2014.6854950

Donin, N., & **Feneyrou, L.**, **editors.** (2017). *Théories de la composition musicale au xx^{e} siècle*. Symétrie.

Fernández, J. D., & **Vico, F.** (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research*, *48*(1), 513–582. DOI: https://doi.org/10.1613/jair.3908

Gabrielli, L., **Cella, C.-E.**, **Vesperini, F.**, **Droghini, D.**, **Principi, E.**, & **Squartini, S.** (2018). Deep learning for timbre modification and transfer: An evaluation study. In *Proceedings of the Audio Engineering Society (AES) Convention 144*.

Ghisi, D. (2017). *Music Across Music: Towards a Corpus-Based, Interactive Computer-Aided Composition*. PhD thesis, IRCAM.

Gillick, J., **Cella, C.-E.**, & **Bamman, D.** (2019). Estimating unobserved audio features for target-based orchestration. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 192–199.

He, K., **Zhang, X.**, **Ren, S.**, & **Sun, J.** (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

Hirn, M., **Mallat, S.**, & **Poilvert, N.** (2017). Wavelet scattering regression of quantum chemical energies. *Journal of Multiscale Modeling Simulation*, *15*(2), 827–863. DOI: https://doi.org/10.1137/16M1075454

Humphrey, E. J., **Bello, J. P.**, & **LeCun, Y.** (2015). Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent*

Information Systems, *41*(3), 461–481. DOI: https://doi.org/10.1007/s10844-013-0248-5

Humphrey, E. J., **Turnbull, D.**, & **Collins, T.** (2013). A brief review of creative MIR. In *International Society for Music Information Retrieval Conference (ISMIR), Late-Breaking News and Demos*.

Klien, V., **Grill, T.**, & **Flexer, A.** (2012). On automated annotation of acousmatic music. *Journal of New Music Research*, *41*(2), 153–173. DOI: https://doi.org/10.1080/09298215.2011.618226

Krizhevsky, A., **Sutskever, I.**, & **Hinton, G. E.** (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pages 1097–1105.

Lacoste, A., & **Eck, D.** (2005). Onset detection with artificial neural networks. In *Music Information Retrieval Evaluation eXchange (MIREX)*, pages 1097–1105.

LeCun, Y., **Bengio, Y.**, & **Hinton, G.** (2015). Deep learning. *Nature*, *521*, 436–444. DOI: https://doi.org/10.1038/nature14539

Lipp, C. (1996). Real-time interactive digital signal processing: A view of computer music. *Computer Music Journal*, *20*(4), 21–24. DOI: https://doi.org/10.2307/3680412

Lostanlen, V., & **Cella, C.-E.** (2016). Deep convolutional networks on the pitch spiral for music instrument recognition. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 612–618.

Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, *65*(10), 1331–1398. DOI: https://doi.org/10.1002/cpa.21413

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065).

Maresz, Y. (2013). On computer-assisted orchestration. *Contemporary Music Review*, *32*(1), 99–109. DOI: https://doi.org/10.1098/rsta.2015.0203

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, *23*(3), 85–102. DOI: https://doi.org/10.1080/07494467.2013.774515

McAdams, S., & **Giordano, B. L.** (2016). The perception of musical timbre. In S. Hallam, I. Cross & M. Thaut, editors, *The Oxford Handbook of Music Psychology (2nd ed.)*. Oxford University Press. DOI: https://doi.org/10.1162/014892699559797

Mehri, S., **Kumar, K.**, **Gulrajani, I.**, **Kumar, R.**, **Jain, S.**, **Sotelo, J.**, **Courville, A.**, & **Bengio, Y.** (2015). SampleRNN: An unconditional end-to-end neural audio generation model. In *Proceedings of the International Conference on Learning Representations (ICLR)*. DOI: https://doi.org/10.1093/oxfordhb/9780198722946.013.12

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, *116*, 374–388.

Müller, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer.

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511819346

Papadopoulos, H., & **Peeters, G.** (2007). Large-scale study of chord estimation algorithms based on chroma representation and HMM. In *Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 53–60. DOI: https://doi.org/10.1109/CBMI.2007.385392

Papadopoulos, H., & **Peeters, G.** (2011). Joint estimation of chords and downbeats. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 138–152. DOI: https://doi.org/10.1109/TASL.2010.2045236

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, IRCAM.

Sak, H., **Senior, A.**, & **Beaufays, F.** (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 338–342.

Schlüter, J., & **Böck, S.** (2014). Improved musical onset detection with convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6979–6983. DOI: https://doi.org/10.1109/ICASSP.2014.6854953

Sifre, L., & **Mallat, S.** (2013). Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1233–1240. DOI: https://doi.org/10.1109/CVPR.2013.163

Smalley, D. (1997). Spectromorphology: Explaining sound-shapes. *Organised Sound*, *2*(2), 107–126. DOI: https://doi.org/10.1017/S1355771897009059

Srinivas, S., **Sarvadevabhatla, R. K.**, **Mopuri, K. R.**, **Prabhu, N.**, **Kruthiventi, S. S. S.**, & **Babu, R. V.** (2016). A taxonomy of deep convolutional neural nets for computer vision. *Frontiers in Robotics and AI*, *2*, 36. DOI: https://doi.org/10.3389/frobt.2015.00036

van den Oord, A., **Dieleman, S.**, **Zen, H.**, **Simonyan, K.**, **Vinyals, O.**, **Graves, A.**, **Kalchbrenner, N.**, **Senior, A.**, & **Kavukcuoglu, K.** (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Vinet, H. (2003). The representation level of music information. In *Proceedings of the International Symposium on Computer Music Modeling and Retrieval (CMMR)*, pages 193–209. DOI: https://doi.org/10.1007/978-3-540-39900-1_17

Vinet, H. (2008). Science and technology of music and sound: The IRCAM roadmap. *Journal of New Music Research*, *36*(3), 207–226. DOI: https://doi.org/10.1080/09298210701859313

Wiggins, G. A. (2009). Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. In *Proceedings of the IEEE International Symposium on Multimedia*, pages 477–482. DOI: https://doi.org/10.1109/ISM.2009.36

Wiggins, G. A., Pearce, M. T., & **Müllensiefen, D.** (2009). Computational modelling of music cognition and musical creativity. In R. T. Dean, editor. *The Oxford Handbook of Computer Music*, pages 383–420. Oxford University Press.

Ycart, A., & **Benetos, E.** (2017). A study on LSTM networks for polyphonic music sequence modelling. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 421–427.

How to cite this article: Cella, C.-E. (2020). Music Information Retrieval and Contemporary Classical Music: A Successful Failure. *Transactions of the International Society for Music Information Retrieval,* 3(1), pp. 126–136. DOI: https://doi.org/10.5334/tismir.55

Submitted: 29 February 2020 **Accepted:** 02 July 2020 **Published:** 01 September 2020

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS