Machine listening intelligence

Carmine-Emanuele Cella[1]

[1]Ircam, Paris, France

This manifesto paper will introduce *machine listening intelligence,* an integrated research framework for acoustic and musical signals modelling, based on signal processing, deep learning and computational musicology.

Keywords: deep unsupervised learning, computational musicology, representation theory.

1 Introduction

1.1 Motivation

The relation between *signals* and *symbols* is a central problem for acoustic signal processing. Among different kind of signals, musical signals are specific examples in which there is some information regarding the underlying symbolic structure. While an impressive amount of research has been done in this domain in the past thirty years, the symbolic processing of acoustic and musical signals is still only partially possible.

The aim of this paper, grounded on our previous work [Cella, 2009], is to propose a manifesto for **a generalised approach for the representation of acoustic and musical signals called *machine listening intelligence* (MLI),** by integrating cognitive musicology insights, hand-crafted signal processing and deep learning methods in a general mathematical framework.

Among existing approaches that share similarities with ours, there are the *multiple viewpoint system* [Conklin, 2013] and *IDyOM* [Pearce and Wiggins, 2013]. While comparing differences and similiarities with these approaches could be interesting, we will not do this here and we mention them only for reference.

1.2 Scientific assessment

In the past twenty years, with the improvement of computers and the advancements in machine learning techniques, a whole field called *music information retrieval (MIR)* developed massively. This important domain of research brought impressive results and had been able to tackle problems that appeared to be unsolvable, such as the classification of very complex signals. Nonetheless, tasks that are relatively easy for humans are still hard and there are not general solutions. Apparently, these kind of tasks are often ill-defined or lack proper information to be correctly represented.

More recently, the uprise of deep learning techniques in computer vision created a real revolution in machine learning given the advancements they provided [Krizhevsky et al., 2012]. Deep convolutional networks, for example, provide state of the art classifications and regressions results over many high-dimensional problems [Le Cun et al., 2015]. Their general architecture is based on a cascade of linear filter weights and non-linearities [Mallat, 2016]; the weights are learned from massive training phases and generally outperform hand-crafted features. The switch to large scale problems that happened in the past few years in the computer vision community, moreover, proved the fact that we need to address more general problems for acoustic signals, where the domain of research is not defined by a single sample but by a whole *population*.

However, these complex programmable machines bring us to a very difficult and partially unknown mathematical world. We believe that *representation theory* is good candidate for such mathematical model. Signal representation methods can be thought as linear operators in vector space and representation theory studies abstract algebraic structures by representing their elements as linear transformations of vector spaces [Fulton and Harris, 2004].

Next sections will be as follows: section 2 will show some problems and applications that we would like to address with our approach. From section 3 we will go more into technical details describing what are representations for signals and reviewing their properties both from cognitive and mathematical standpoints. This will serve a background for section 4, that will present the general research approach for MLI.

2 Problems and outcomes

Among the large number of open problems in the field of signal processing, we would like to present here some interesting examples that could be treated in the context of machine listening intelligence. These problems refer particularly to music and creative applications, but we think that the developed methodology could be further used in different domains. As such, they must be considered just as examples of possible outcomes.

2.1 Semantic signal processing

A signal transformation is, in a general sense, any process that changes or alter a signal in a significant way. Transformations are closely related to representations: each action is, indeed, performed in a specific representation level. For example, an elongation performed in time domain gives inferior results (perceptually) to the same transformation performed in frequency domain, where you have access to phases. In the same way, a pitch shift operated in frequency domain gives inferior results to the same operation performed using a spectral envelope representation, where you have access to dominant regions in the signal. In the two cases discussed above we passed from a low-level representation (waveform) to a middle-level representation (spectral envelope). We could, ideally, iterate this process by increasing the level of abstraction in a representation thus giving access to specific properties of sound that are perceptually relevant; by means of a powerful representation it could therefore be possible to access a *semantic level* for transformations. We envision, therefore, the possibility in the future to have such kind of semantic transformations by accessing representations given by deep learning models.

2.2 High-level acoustic features

One of the most impressive example of creative application of deep learning is, in our opinion, style transfer. In a paper published in 2016 [Gatys et al., 2016], for example, the authors showed how it is possible, using the latent space of a deep network, to transfer high level features from one image to another. It is interesting to remark that the transferred features are not simple *effects* but real traits of the style of the image.

Unfortunately, such an impressive process is not possible on acoustic signals for the moment. In 2013 we achieved the construction of an advanced hybridisation system for sounds [Cella and Burred, 2013]. The basic idea of our approach was to represent sounds in term of classes of equivalences and probabilities, then mix the classes of one sound with the probabilities of another one. While the perceptual results were satisfying for us, the limitation of the representation method used, didn't permit to access *high-level acoustic features*.

We believe that deep musical style transfer can be considered as a generalisation of hybridisation and we strongly believe that this kind of processing could be achieved by MLI.

3 Signal representations

Defining a representation for music and musical signals involves the establishment of essential properties that must be satisfied. Many years of research have been devoted to such a complex task in the field of cognitive musicology, a branch of cognitive sciences focused on the modelling of musical knowledge by means of computational methods [Laske and al., 1992].

Among important properties for musical representations found from the literature in the field, there are milestones such as *multiple abstraction levels*, *multi-scale* and *generativity*. Interestingly enough, these properties are not only specific to music but can be applicable to different kind of acoustic signals, as we will see.

Usually, the description of acoustic signals select a particular degree of abstraction in the domain of the representation. In general, low-level representations are generic and have very high dimensionality. These representations evolve fast and the only transformations that are possible at this level are geometric (translations, rotations, etc.) and are mostly defined on continuous (Lie) groups [Mallat, 2016]. In the middle, there are families of representations (often related to perceptual concepts) that have a medium level of abstraction and a not so huge dimensionality and allow for transformations on specific concept (variables), usually defined on discrete groups [Lostanlen and Cella, 2016]. On the other end, very abstract representations are pretty much expressive and have a low dimensionality; in a sense, these representations deal with almost-stationary entities such as musical ideas and unfortunately it is very difficult to know which mathematical structure stays behind. As an example, we could think to low-level representations as signals (used by listeners), to middle-level as scores (used by performers) and to high-level as musical ideas (used by composers). Figure 1 depicts the outlined concepts.

Representations can be considered linear operators that need to be invariant to sources of unimportant variability, while being able to capture discriminative information from signals. As such, they must respect four basic properties; being x a signal and Φx its representation:

- *discriminability:* $\Phi x \neq \Phi y \implies x \neq y$;
- *stability:* $\|\Phi x - \Phi y\|_2 \leq C\|x - y\|_2$;
- *invariance (to group of transformations G):* $\forall g \in G, \Phi g.x = \Phi x$;

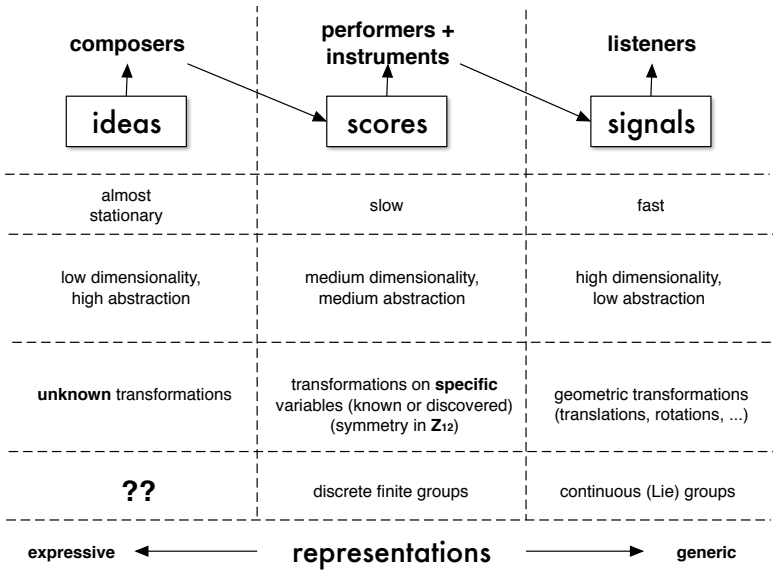

Figure 1: Different abstraction degrees in representations.

- *reconstruction:* $y = \Phi x \iff \tilde{x} = \Phi^{-1} y$.

Discriminability means that if the representations of two signals are different than the two signals must be different. Stability means that a small modification of a signal should be reflected in a small modification in the representation and vice-versa. Invariance to a group of transformation G, means that if a member of the group is applied to a signal, than the representation must not change; reconstruction, finally, is the possibility to go back to a signal that is *equivalent* to the original (in the sense of a group of transformations) from the representation. It is possible to divide representations in two major categories: *prior* and *learned*.

3.1 Prior and learned representations

In prior representations, signals are decomposed using a basis that has been defined mathematically, in order to respect (some of) the properties given above. The general model of a prior representation is a decomposition of a signal x into a linear combination of expansion functions: $x = \sum_{k=1}^{K} \alpha_k g_k$, where K is the dimensionality, the coefficients α_k are weights derived from an analysis stage an functions g_k are fixed beforehand and are used during a synthesis stage. The choice of the decomposition functions is dependent on the particular type of application needed and the more compact (sparse) the representation is, the more the functions are correlated to the signal.

In learned representations, the decomposition functions used to describe a signal are learned by a training on some examples that belong to a specific problem. The training can be done in two different ways: *supervised* or *unsupervised*.

Supervised learning is a high-dimensional interpolation problem. We approximate a function $f(x)$ from q training samples $\{x_i, f(x_i)\}_{i \leq q}$, where x is a data vector of very high dimension d. A powerful example of supervised learned representation are convolutional neural networks.

In unsupervised training, on the other hand, there is not a target function to approximate and other mathematical constraints are applied, such as sparsity or variance reduction. Typical

examples of unsupervised representations are sparse coding and auto-encoders.

3.2 Importance of unsupervised learning

Recent advancements showed that supervised learning is able, if given the right conditions, to outperform both unsupervised and prior representations. There are problems, however, where this family of representations cannot be applied and the only possible approach for learning a representation is using unsupervised methods:

- *lack of labeled data*: copyright issues can make impossible to deploy a database of labeled music of a significative size to be used a reference case for reproducible research;
- *cost of data gathering*: some real-life problems are related to contexts in which is impossible to gather large amount of data (such as biomedical recordings);
- *conceptual disagreement*: in some cases, it is very difficult or even impossible to assign labels to acoustic signals given their inherent ambiguity (music is often such a case).

4 Research approach

The discussion given above outlines, in our opinion, the necessity of a general framework able to integrate the different approaches to the representation of musical and acoustical signals into a common perspective.

Machine listening intelligence aims at being such a framework, by integrating cognitive musicology insights, established signal processing techniques and more recent advancements in deep learning in the context of *representation theory*.

Prior representations are defined by mathematical models but fail to achieve the same expressivity of learned representations. On the contrary, deep learning proved to be valuable in incredibly different domains and showed that some learning techniques are indeed general. One of the issues, in the case of acoustic and musical signals, is that there is not a common and established mathematical model for this kind of methods. Moreover, supervised learning is not always possible for the reasons outlined in section 3.2. Therefore, we envision a research approach based on the following main factors:

- **deep unsupervised learning methods**: only with deep architectures it is possible to create multi-scale representations that have different abstraction levels; using unsupervised learning, it is possible to address difficult problems that lack labelled data;
- **representation theory**: by interpreting learning with linear operators, it is possible to create a common mathematical language to compare and study its properties;
- **large scale problems**: using large datasets (that usually embody difficult problems) will impose the research of scalable and general learning methods, that can be transferred to many different domains;
- **multi-disciplinarity**: putting together several sources of knowledge such as psychoacoustics, cognitive musicology, computational neurobiology, signal processing and machine learning is the key for future development of the so-called *intelligent* machines.

While prior representations formally define the level of abstraction, they cannot reach the same level of aggregate information gathered by deep learning networks. These networks, on the other hand, are not capable of explaining the concepts they discover. For such reasons, it is interesting to make a bridge between these two approaches by immersing both in a more general framework that could be found in representation theory. Machine listening intelligence aims at filling this gap.

References

C. E. Cella. Towards a symbolic approach to sound analysis. In *proceedings of MCM*. Yale, USA, 2009.

C. E. Cella and J. J. Burred. Advanced sound hybridization by means of the theory of sound-types. In *proceedings of ICMC*. Perth, Australia, 2013.

D. Conklin. Multiple viewpoint systems for music classification. In *Journal of New Music Research*, volume 42, no. 1, pages 19–26, 2013.

W. Fulton and J. Harris. Representation theory. In *proceedings of MCM*, volume 129. NY: Springer New York, 2004.

L. Gatys, A. Ecker, and M. Bethge. Style transfer using convolutional neural networks. In *proceedings of CVPR*, 2016.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *proceedings of NIPS*, pages 1090–1098, 2012.

O. Laske and al. Understanding music with a.i.: perspectives on music cognition. AAAI Press, 1992.

Y. Le Cun, Y. Bengio, and G. Hinton. Deep learning. In *Nature*. 215, 2015.

V. Lostanlen and C. E. Cella. Deep convolutional networks on the pitch spiral for musical instrument recognition. In *proceedings of ISMIR*. New York, USA, 2016.

S. Mallat. Understanding deep convolutional networks. In *Philosophical transactions A*, 2016.

M. T. Pearce and G. A. Wiggins. Auditory expectation: The information dynamics of music perception and cognition. In *Topics in Cognitive Science*, volume 4, no. 4, pages 625–652, 2013.