FROM TIMBRE DECOMPOSITION TO MUSIC COMPOSITION

Frédéric Le Bel

Music Representation Team, IRCAM (UMR 9912 STMS), Paris, France;
Centre de recherche en informatique musicale, MUSIDANCE, Paris 8, Paris, France.
frdric.lebel@gmail.com | http://repmus.ircam.fr/lebel/

INTRODUCTION

This poster proposes to overview a work in progress that aims at developing a computer-aided-composition (CAC) approach to structuring music by means of audio clustering and graph search algorithms [Le Bel 2017]. Although parts of this idea have been investigated in order to achieve different tasks such as corpus-based concatenative synthesis [Schwartz, Beller et al. 2006], musical genre recognition [Peeters 2007] or computer-aided orchestration [Carpentier 2008] to name a few, the challenge remains to find a way of integrating these techniques into the composition process itself, not to generate material but to explore, to analyse and to understand the full potential of a given sound corpus (sound file database) prior to scoring a musical piece; being instrumental, acousmatic or mixed. As opposed to mainstream CAC tools, mostly focusing on generative methods, the following one proposes an analytical approach to structuring music based on auditory attributes and their physical correlates. The idea is then to use unsupervised machine-learning to extract hidden structures from sound corpuses' features space and to translate them into musical structures at different stages of the composition; from the micro-form to the macro-form. From another perspective, the idea is to elaborate musical works based, not on patterns proliferation or similar techniques, but rather on relationships that bring together different sound entities. Consequently, the goal here is to unfold the algorithmic structure through a prototype software and to reveal how it was used to compose two recent pieces of mine: '*Il ne s'agit pas de le recomposer*', for augmented string quartet and fixed media and 'Mais plutôt de trouver la, ou les justes relations accordant l'existence de tous ses éléments', for multichannel diffusion system.

STRUCTURAL OVERVIEW

Based on a typical unsupervised machine-learning architecture, the following algorithmic structure may be divided into three distinct stages. The first one focuses on data extraction (audio features extraction), the second one focuses on data analysis (features selection, clustering algorithm and evaluation criterion) and the third one focuses on data sorting (graph search algorithm). Essentially, the task is to deduce a musical structure from a sound file database (audio and not symbolic).

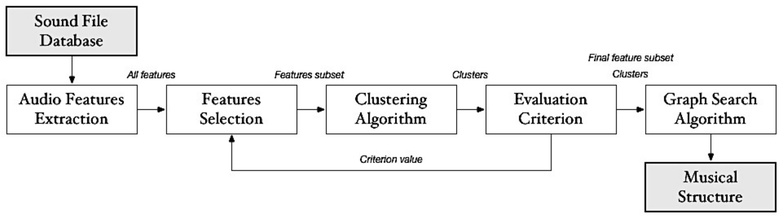

Figure 1. Algorithmic structure

FROM TIMBRE DECOMPOSITION…

In the frame of this work, the audio features extraction consists of decomposing each sound file from the database by mapping the signal's short-term Fourier transform (STFT) magnitude into a lower-dimensional domain that more clearly reveals the signal characteristics. Assumed to represent specific auditory attributes, these features inform different aspects of the temporal structure, the energy envelope, the spectral morphology and the harmonic content of sounds. From these low-level data models, the sound files are then compared, taken pairwise, on a dissimilarity/distance basis in order to generate what could be seen as a *n*-dimensional timbre-space upon which the clustering algorithm can be later applied. Briefly, the evaluation criterion for the features selection aims at maximizing the inter-clusters distances and at minimizing the intra-clusters' ones [Dy and Broadley 2004]. In this particular case, the timbre-space metaphor should be considered as an exploratory structure of dissimilarity data rather than a comprehensive perceptual coordinate system of timbre [Siedenburg, Fujinaga and McAdams 2016].

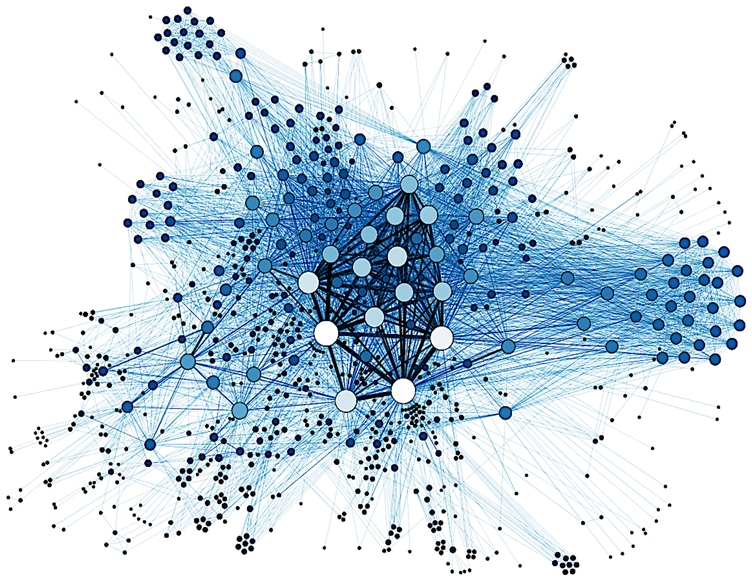

Figure 2. Two-dimensional timbre-space network, where the nodes represent the sounds and the edges the distances between them.

A SPECTROMORPHOLOGICAL APPROACH

As described earlier, the features selection and the evaluation criterion phases may seem relatively straight forward but it appears to be one of the most difficult and most crucial step of the process for the resulting timbre space to have a consistent psychoacoustic meaning. Indeed, many different strategies can be found in both the machine learning and the psychological literature but the propositions yet remain either task specific, simply incomplete or require a lot of human inference [Lagrange et al. 2015]. Because of that, different empirical approaches were deployed in order to serve adequately the exploratory purposes of the framework and one appeared to be particularly efficient. Being somehow not too generic but also not too specific, that is the spectromorphological approach [Smalley 1986]. In this case, a measure of the linear correlation *r* [Pearson 1895] between the averaged relative specific loudness (RSL) [Peeters 2004] of each pair of sounds is taken to assess their similarity.

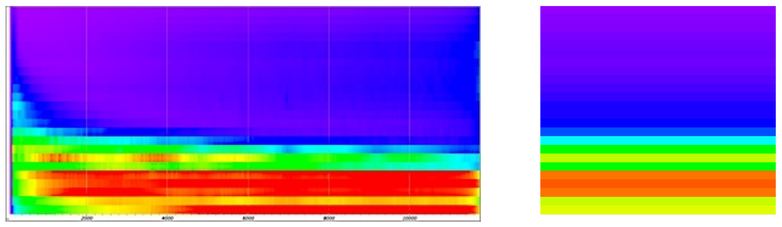

Figure 3. [left] Instantaneous RSL (x: time, y: bark bands, z: normalized specific loudness), [right] Global RSL (means/bark bands).

This method is interesting in its simplicity because it seems to handle two important yet psychologically unaddressed problems when comparing sounds. One being the temporal dimension and its effect on mental reconstruction of audio events, and the other one being the inherent interdependencies of audio features and their impact on how we describe tones.

…TO MUSIC COMPOSITION

Following an extensive data analysis stage (the features selection and the clustering algorithm), the graph search consists of finding path(s) within the timbre-space that could suggest way(s) of sequencing the different clusters of sounds and their components in a chronological order. Considering that the resulting network may be seen as a complete, undirected and weighted graph, multiple approaches could be used but spanning trees were investigated first. More precisely, the minimum spanning tree (MST) seemed to provide an interesting premise accordingly with the timbre-space's underlying dissimilarity data structure. A MST is a subset of the edges of an undirected and weighted graph that connects all the nodes together with the minimum possible total edge weight without any cycles. In other words, it is a tree (not a path) whose sum of edge weight is as small as possible [Graham and Hell 1985]. Hence, the MST does not represent an ordered sequence of clusters but rather some sort of an optimized road map through the timbre-space from which multiple paths may be drawn. However, the clusters and their components remain connected in a way that the global similarity is maximized. The Kruskal algorithm [Kruskal 1956] was used in the frame of this work.

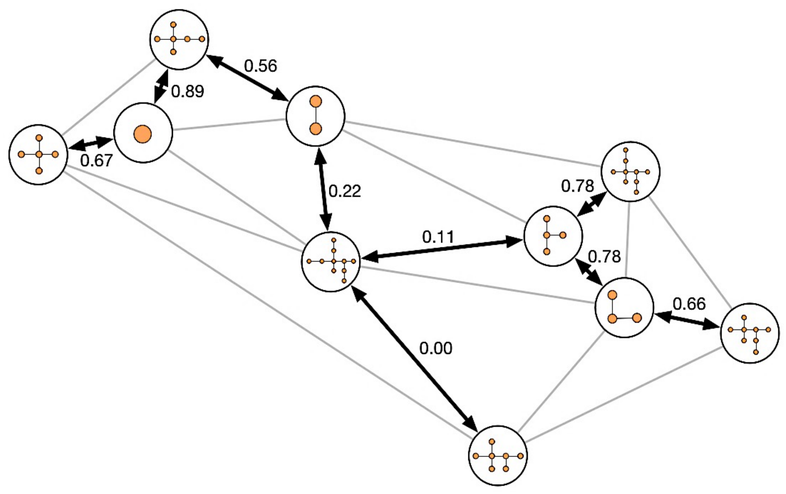

Figure 4. Two-layer MST extracted from the timbre-space, where the big nodes represent the clusters of sounds, the small nodes represent the sounds and the edges on both layer represent the transition/emission probabilities respectively.

The previous two-layer MST already suggests a rather clear musical structure, or a particular definition of it, but yet remains to be expressed in the time domain. For that, the model is translated into what could be seen as a hidden markov model (HMM) for which the edges, originally distance based, are converted into transition and emission probabilities respectively for each layer (global structure and local structures) according to the following principle: $1\text{-}d_i/\text{argmax}(d)$. In other words, the closer or the more similar are two sounds, or two clusters of sounds, higher is the probability to transit from one to the other and vice-versa. Finally, a customized polyphonic step sequencer is used to articulate the resulting probabilistic model on a timeline and let the musical structure be heard.

THE SEQUENCER

Based on a typical step sequencer, the timbre space exploration can be done in real-time manually or automatically by controlling or pre-setting various parameters such as the tempo, the number of steps per bars (one bar corresponding to one cluster or the transition period), the length of each steps (one step corresponding to one sound file or the emission period), the amount of rhythmic variation, the velocity and the envelope of each sound file, and the polyphonic density (maximum number of superimposed voices) among others. The spatialisation is done by assigning each sounds to one of 16 evenly spaced source positions on a circular plane in order to give a sense of spatial imprint to each cluster.

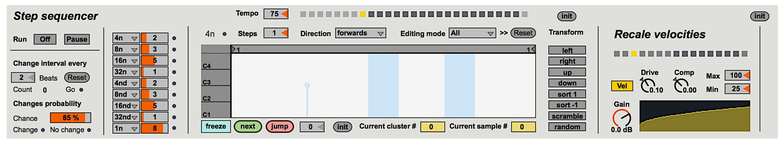

Figure 5. The sequencer's control interface

CONCLUSIONS

Many questions related to sound perception remain open despite solutions are put forward. Among those, the temporal dimension should be further investigated in order to have a better understanding on the effect of time (sound durations) through perception for measuring the similarity between sounds with more accuracy. Another one is the method for measuring the similarity itself. Using more than one approach simultaneously (magnitude, orientation and dependency) may be a fairly good solution but the problem of interpreting the results accurately remains open. More specifically when comparing a sound and its retrograde [Le Bel 2017]. Another problem is the shape of space implied by the dimensionality reduction and their impact on the shape of clusters. Although the Euclidean space seems to be well suited for achieving such tasks in general, this question is another one that should be further investigated from a perceptual angle. Another one is related to the graph exploration. Considering that the context of this work is art oriented, the graph search algorithms should be further investigated from a perceptual angle rather than an optimization one in order to exploit the full potential of these tools into the creative process. In this sense, these algorithms should be further evaluated for their musical potential rather than for their efficiency. In other words, the question is about the kind of musical structure the various graph search algorithms may lead to. This approach also comes with certain limitations. The first one is about using raw audio signal as main input. Contrary to mainstream CAC approaches, it may be an advantage but it is also its main disadvantage because the quality of the output is directly correlated to the quality of the input but also because the whole process depends on it. Another limitation is related to the use of low-level audio features. Although the resulting space of variables may quickly become very complex and give the impression of covering a very large array of sounds, the results remain interpretable on a low-level basis only, meaning that no aesthetical nor emotional affects may be considered using such an approach. Then, the clustering method is itself another notable limitation. Based on unsupervised learning, the method does not provide any information on the clusters components other than a similarity index. In other words, the results are simply quantitative and not qualitative. Finally, the complexity of this tool may be a limitation itself because an 'uneducated' user may end spending a lot of time understanding the multiple parameters of this approach.

REFERENCES

Carpentier, G. (2008). Approche computationelle de l'orchestration musicale: Optimisation multicritère sous contraintes decombinaisons instrumentales dans de grandes banques de sons. *Thèse de doctorat*, Université Paris VI – Pierre et Marie Curie, Paris, France.

Dy, J. G. and Broadley, C. E. (2004). Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research 5*, 845-889.

Graham, R. L., Hell, P. (1985). On the History of the Minimum Spanning Tree Problem. *Annals of the History of Computing*, Vol. 7, No. 1, 43-57.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, Vol. 7, 48-50.

Lagrange M., Lafay G., Défréville B. and Aucouturier J.-J. (2015). The bag-of-frames approach: A not so sufficient model for urban soundscapes. *The Journal of Acoustical Society of America*, Vol. 138, No. 5.

Le Bel, F. (2017). Structuring Music by Means of Audio Clustering and Graph Search Algorithms. *Proceedings of the Journées d'Informatique Musicales 2017*, Paris, France.

Peeters, G. (2007). A Generic System for Audio Indexing: Application to Speech/Music Segmentation and Music Genre Recognition. *Proc. Of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France.

Peeters G. (2004). A large set of audio features for sound description (similarity and classification) in CUIDADO project. unpublished version 1.0 (23 avril), Ircam.

Schwartz, D., Beller, G. et al. (2006). Real-time Corpus-based Concatenative Synthesis with Catart. *Proc. Of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada.

Siedenburg, K., Fujinaga, I. and McAdams, S. (2016). A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology. *Journal of New Music Research*, DOI:10.1080/09298215.2015.113273

Smalley D. (1986). Spectro-morphology and Structuring Processes. *The Language of Electroacoustic Music*, Palgrave Macmillan, London, pp. 61-93.

FOR MORE INFORMATION

http://repmus.ircam.fr/lebel/from-timbre-decomposition-to-music-composition