MULTIMODAL INTERACTIONS SYNTHESIS BY SPARSE CODING FOR THE WRITING OF IQUISME AND ANALYSIS OF ITS PERCEPTS

Maxence Mercier [1], Joseph Razik [2,3], Hervé Glotin [2,3,4]

[1] Association Otra, www.o-tra.net
[2] Aix Marseille Université, CNRS, ENSAM, LSIS UMR 7296, 13397 Marseille, France
[3] Université de Toulon, CNRS, LSIS UMR 7296, 83957 La Garde, France
[4] Institut Universitaire de France, IUF, Paris 75005

maxence.mercier@o-tra.net, razik@univ-tln.fr, glotin@univ-tln.fr

ABSTRACT

By multimodal interactions synthesis, we mean resynthesizing multimodal scenes via a corpus of heterogeneous data. This project is based on a framework developed in Max and Matlab which constitutes the digital composer's desk of Iquisme, a piece for soprano, nine instruments, an electroacoustic device and video. The corpus is made of material used in the first movement of Iquisme concert version. It is comprised of sounds, videos, 3D particle matrices and algorithmic control parameters indexed to the temporal progress of the piece. A dictionary learning of the corpus by Sparse Coding produces a lexicon of elementary interactions capable of reproducing parsimoniously any sequence of the first movement. Provided with this code, we then intend to reinject and deploy back into the progress of the piece a flow of data whose aim is to control the electroacoustic and video devices.

A first experimentation enabled us to define a methodology for generating different dictionaries derived from the analysis of audio, video and audio coupled to the video. The whole corpus as well as a visualizing tool of the results applied to the temporal progress of the first six minutes of Iquisme first movement is available for download [1].

1. PRESENTATION

1.1. Writing

Iquisme [1] is a writing project of a concert piece in three movements for soprano, nine instruments, electroacoustic device and real-time video synthesis device.
Its narrative is suggestive of the organic evolution of an intermediatic universe. This evolution is the fruit of a generative writing whose one issue focuses on the orchestration of audiovisual interactions.

A set of real-time applications realized with Max 6 forms the digital writing workshop of Iquisme.
The piece's first movement is written empirically. The association between the music and the video is a result of thorough settings on the basis of esthetical and narrative criteria.

Real-time has been considered from the origin of the project in order to satisfy an intuitive writing process with an immediate rendering and to allow eventually a flexible interpretation by the instrumentalists. Besides, the electroacoustic and video devices will also be proposed as an interactive installation. These real-time systems allow for an immediate audiovisual writing simulation.

Concerning Iquisme's narration, the topic brought to the screen is a cloud of particles animated by various algorithms [2]. Bringing these particles to life and combine them with the sound within the aesthetic and narrative constraints of the piece is the fruit of extensive researches and compromises which lead to a panel of suitable behaviors. The first movement exposes them in embryonic stages. The synthesis of multimodal interactions will enable to develop them organically in the continuation of the work.

1.2. Analyse

The data used and generated in the first movement are incorporated in a heterogeneous corpus.

The corpus analysis is not performed in non real-time. It is base on a learning method by Sparse Coding, also called "compress sensing" or "parsimonious coding" which identifies, classifies and then models the corpus' interactions.

The relations between music and video of *Iquisme*'s first movement are extracted in order to compose a lexicon of multidimensional interactions.

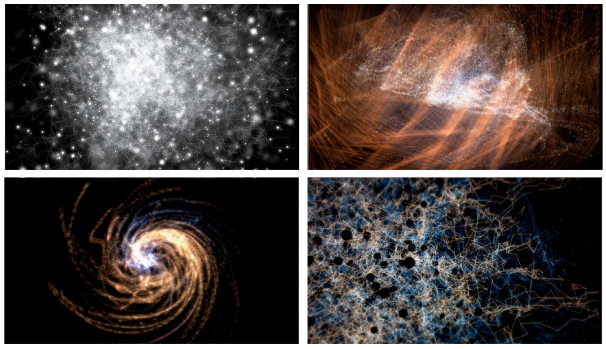

Figure 1. Various 3D video particles organizing arrangements generated in real-time.

1.3. Synthesis

The multimodal interactions synthesis is based on the modeling of the properties of the analyzed corpus.

The identified interactions will eventually be recombined in order to arrange new sequences for the second and third movement of Iquisme.

This approach will need to offer a unified standard in the writing of interactions as well as new perspectives and non linear convergences in its parameters.

The objective is to make proliferate, maximize and to control the direction of the generative processes of the electronic devices. This shall be done by exploring the different possibilities of coherent combinations with the parametric balances as defined in the original corpus. The result lexicon, combined with a graphic tool, will enable to manipulate and represent in real-time vast sets of heterogeneous data according to the morphologic properties of the corpus. (Figure 2 for a schematization of the architecture of the system).

2. REAL-TIME DEVICES

Every tool for the writing and performing of Iquisme has been designed with Max [3] under OSX. The CPU resources have been optimized by distributing the processes to nine applications allocated on two machines, one dedicated to the video, the other to the audio. The applications communicate by means of networks and exchange Jitter matrices. The audio passes through an internal routing within the computer. The video generators send the rendering to an application dedicated to the projection as shown in figure 3. A global control application receives the MIDI triggering orders and manages the synchronization between audio and video. Interfaces dedicated to the sound and to the video offer an intuitive control over the parameters of the real-time generation software during the writing phases.

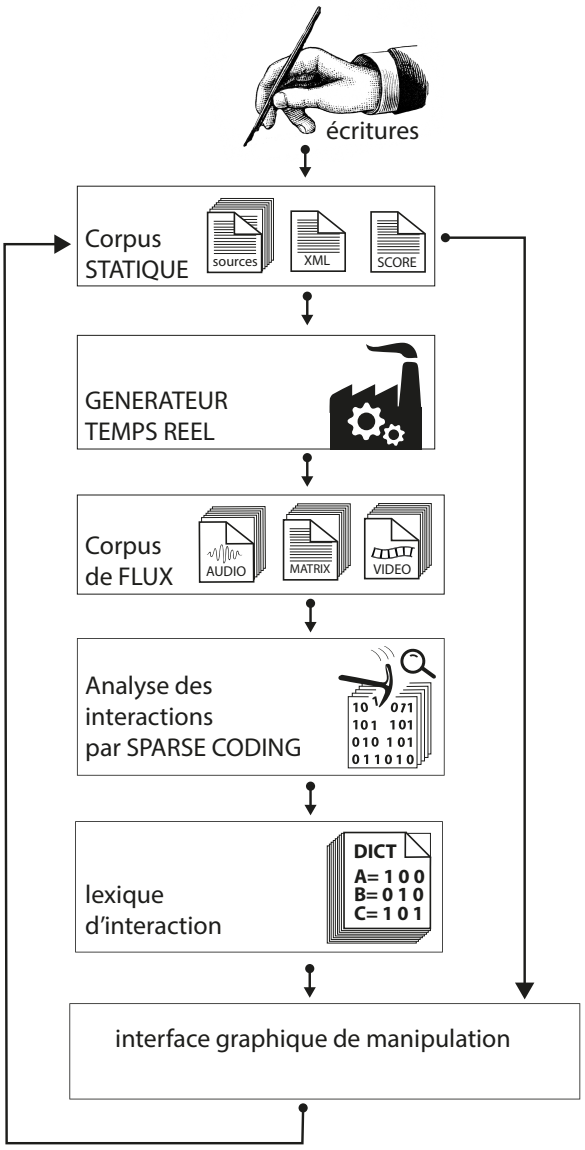

Figure 2. Workflow of the multimodal interaction synthesis.

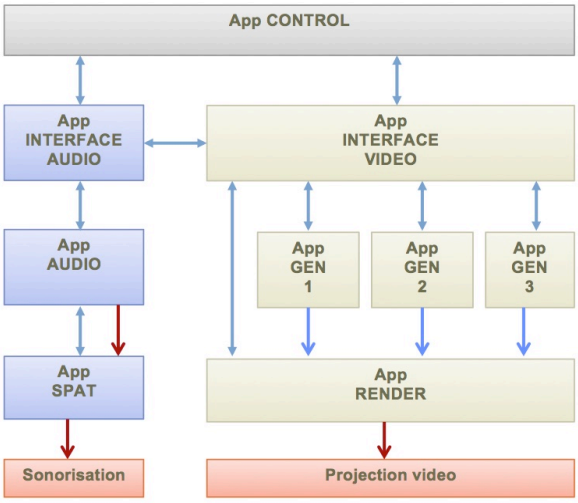

Figure 3. Synoptic of the real-time applications.

The parameters defined in the graphic interfaces are recorded in preset memories (pattern storage).

Every preset change can modify around a hundred parameters, added to which are parametric control strategies through envelope triggering and continuous control through real-time descriptors.

As part of a simulation, a regular sequencer with an audio demo of the instrumental parts enables a MIDI synchronization of the whole environment.

On the occasion of a stage performance, all the real-time processes can adapt to the musicians interpretation. All the algorithmic processes and samples readings are designed to adapt to a changing tempo during live performances or rehearsals.

3. THE CORPUS

The corpus is split into two data categories.
The analysis rests on flows stemming from real-time processes which are to be segmented into semantic units on the basis of their morphological properties. The results will be indexed over static data.

3.1. Static data corpus

It pertains to the data stemming from the composer's writing. This data comprises the instrumental score, the electronic score and the software (patch Max/MSP/Jitter) used by the real-time generators.

Control file are in Json format and command the parameters of the generators. This set of parameters constitutes the electronic score which is to be adapted in the future in order to accommodate the use of Antescofo [4]. This corpus has been conceived empirically but follows a structure that allows the extraction of content in an automatic manner.

3.2. Data flow corpus

The second category comprises the flows produced by the real-time generators. The completion of a simulation provides us with a particles matrix available in Frame and in audio and video files. The completion of a simulation provides us with a particles matrix available in Frame and audio and video files.

3.2.1. Video data

The video data is aggregated into a matrix recorded in csv files in successive frames. Every line of the matrix is dedicated to a particle. The columns inform on its position, direction, speed and color.

The frames are finalized with a description of the image rendered at the output of the video generator. It is characterized by the position of the OpenGL camera, the image's centroid and the average luminous intensity. The perceptive descriptors come from the cv.jit library.

The real-time video generation rate is set at 60fps (16ms per frame) but the frames recording is downsampled to 25fps (40ms per frame).

The files are designated as follows:
XXX-F_XXXXX-B_XXXXX-P_XXXXX-T.csv

F for the Frame number
B for the filename of the active preset bank
P for the Preset Number and Name
T for Elapsed Time since last preset change

The filename's description allows the analysis system to reference each frame on a temporal index in milliseconds, musical bar in bar.beat.unit and name of the software used.
Frames are matrices where each column is the atom of a vector with n dimensions.

For every particle, a line of the matrix informs on its position XYZ in the 3D space, its direction and its color.

3.2.2. Audio

The audio flows are pre-analyzed by perceptive and spectral descriptors obtained from the libraries for Max/MSP, Zsa.Descriptors [6] and ircamdescriptor [7].

(Loundness, Spread, MFCC, SpectralCentroid, SpectralSpread, SpectralVariation, NoiseEnergy, TotalEnergy, Loudness, FundamentalFrequency)

This data is aggregated in a matrix recorded in successive 40 ms frames. The files of each frame are created according to the same method used for the recording of the frames stemming from the video.

The audio data is also available in wav files and can be subjected to more in-depth analysis in Matlab. There is a track for every instrument, four tracks for the electroacoustic part and an overall stereo mixing.

4. SPARSE CODING

"Sparse coding" allows to identify and to classify interactions induced by the different typologies of sound and video writings.

The multimodal Scenes (sound, video, particles) are synthesized in an abstract space which codes the major events in basic elements. These elements form a dictionary allowing to index every scene according to the activities of these trimodal events (sound, video, particles).

The learning of the dictionary follows the double constraint of best reproducing every sub-sequence of the work's corpus (least squares criterion), while recruiting as few words from the assembled dictionary as possible (norm L1 of the vector of recruitment of the words of the dictionary). The learning algorithm is from the LASSO algorithms family which, once the learning completed, allows quite a fast projection of the corpus on the words that have been learnt. The cardinal of the dictionary is one of the parameters that should be optimized, depending on the desired result.

5. EXPECTED RESULTS

5.1. Interaction lexicon

The interaction lexicon stems from combinations between the Sparse Coding dictionary and the parametric data of the static corpus.

It differs in its format and is accessible in the form of a database adapted in order to generate a visual representation of the Sparse Coding modelized interactions. Figures 4 and 5 illustrate its content.

```
"ID" : [ 45 ],
"N.TIME_START" : [ 5000, 14000, 25000, 35000, 55000, 90000 ],
"N.MEASURE_START" : [ 2.2.1, 7.1.1, 13.1.1, 17.3.1, 28.1.1, 49.1.1 ]
"N.TIME_STOP " : [ 2500, 1500, 3255, 4850, 1730, 4250 ]
"N.MEASURE_STOP" : [ 4.1.1, 8.1.1, 16.1.1, 19.1.1, 30.1.1, 52.1.1 ]
"N.PROGRAMME" : [ RD, RD, MM, MM, AP, RD ]
"N.ID_GROUPE" : [ ./GroupPart/ID_LEX_X.csv ]
"N.DIST_CAM_GROUP" : [12.54, 13.25, 4.65, 1.80, 15.25 ]
"N.LUM" : [ 0.8, 0.6, 0.84, 0.5, 0.7, 0.6 ]
"N.LOUNDNESS" : [ 0.56, 0.45, 0.68, 0.52, 0.58, 0.63 ]
"N.SPREAD" : [ 0.56, 0.45, 0.68, 0.52, 0.58, 0.63 ]
"N.FUND" : [ 256, 360, 145, 215, 180, 140 ]
```

Figure 4. Example of an entry of the lexical database.

The fields are manifold. They display the list of events sharing similar criteria.

ID : identity of the lexical unit in the dictionary

N.TIME_START : list of offsets in absolute time

N.MEASURE_START: list of offsets in BAR.BEAT.UNIT (bar)

N.TIME_STOP : list of events' end offsets in ms

N.MEASURE_STOP: list of events' end offsets in BAR.BEAT.UNIT

N.GROUPE_PART : CSV files comprising an index of the particles in each frame that shares the same characteristics. Matrix format (ID FRAME, N.ID.PART

N.DIST_CAM_GROUP : FLOAT list. Average distance of each group from the OpenGL camera

N.FUND, N.SPREAD, N.LOUNDNESS are examples of audio descriptors that can be integrated in the lexicon

Figure 5. Field Descriptions for an entry in the lexical base.

5.2. Graphic interface of lexicon visualization

The representation of the lexicon brings to light the similarities between multimodal scenes. It allows to transcribe the corpus symbolically by displaying in an easy way the sequence of the lexical units: A B A C D A etc.

The patterns identifications stemming from these transcriptions offer a scope of application extending to musical analysis and writing [8] [9].

The graphic interface will enable the manipulation of the corpus by recombining the transcribed structure into new arrangements, ranging from a micro form to a macro one.

The interactions with this interface will reinject in real-time the parameters indexed to the lexical database into the generative systems[10]. Every control will then be ready to be integrated to the digital score.

This interface informs on the work's structure in what pertains to its audiovisual relations. The visual and audio

descriptors indexed to the lexicon will allow to constraint the display following defined criteria.

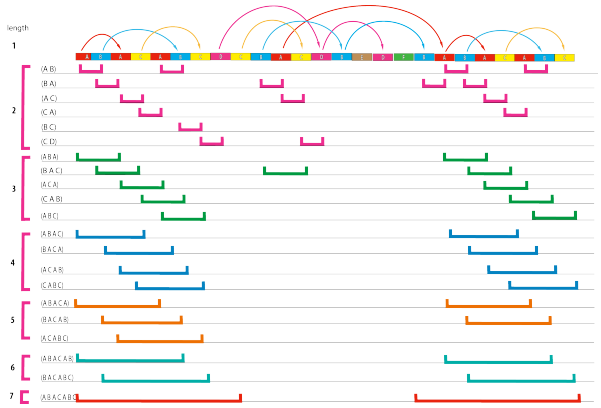

Figure 6. Representation of the visualization interface of the lexical units on a temporal axis.

It will allow to perform on its representation transformations and controls as the maquette of OpenMusic [11].

The objective is to make proliferate, maximize and to control the direction of the generative processes of the electronic devices. This shall be done by exploring the different possibilities of coherent combinations with the parametric balances as defined in the original corpus.

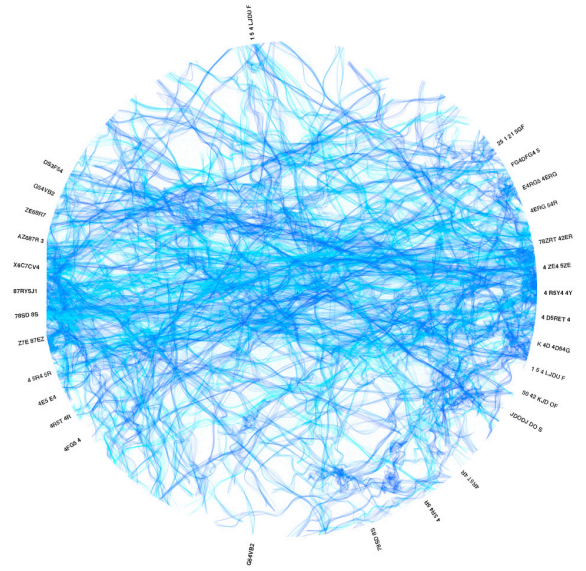

Figure 7. Circular representation of the score.

5.3. Interactive cartography and score

A specific version of Iquisme for 8 screens and 8 audio channels is considered in the form of an interactive installation. The intention is to adapt the electroacoustic and video generative devices in order to meet the scenographic and narrative requirements of a navigable interactive score. This score will be symbolized on the ground in the form of a hyperbolic circular diagram.

Unlike the concert version, the relation to the narration is non linear. Resorting to the lexicon consistently redefines the multimodal rules of interaction depending upon the movements and the positions of the audience in the scenographic space.

The spectators will navigate in a generative score which adapts to their behaviors. They will experiment by themselves the interactive rationales which have currency in the concert version.

Figure 7 illustrates a draft representation of this score, like a kind navigation map depicting the interactions of the corpus. Every point outside the circle corresponds to an entry of the lexical database. Accessing the map enables the activation of relations networks and their interpretation in the space of the installation.

The intersections of the lines are all so many points of convergence and of parametric interpolations that the audience will be able to activate.

The gestural approach is similar to the controls of CataRT concatenative synthesis [12], with the small difference that, in this case, the corpus deals with a set of parametric data.

6. FIRST EXPERIMENTATION

This first experience focuses on the analysis of the six first minutes of the first movement of Iquisme.

In order to judge the coherence of the supervector that needs to be analyzed, we generated three kinds of dictionaries: audio, video and audio coupled with video.

6.1. Entropy

Words sequences for each flow (audio, video and audiovideo) are analyzed depending on the distribution of the activities of the words. The distribution of the activity of a word represents the quantity of information it codes according to Shannon's information theory. Therefore we calculated the entropy (in base-2) of the distribution of each word after having established the probability densities of their activity in 256 bins. The maximum possible entropy is therefore 8 logons (or bits). In all, there are 23 x 3 words. We represent the entropies of all these words in figure 8 by sorting them by ascending order of entropy independently for each flow.

We can see in this figure that the words from the video flow generally have weaker entropy than words from the other flows. We also observe that words from the audio and audiovisual flows have similar entropy, with a tendency to decrease for the audiovisual one, which makes sense considering that the visual flow has weaker entropy. Indeed words with weaker entropy are sometimes not very representative, being activated for only shorts periods during the whole composition.

On the contrary, words starting from the 5th rank approximately have a significant frequency of activity, and it can be assumed that their entropy indeed represents a measure of the information they carry in the composition.

Two purposes for this measure can then be proposed:

a) a quantification of the volumes of information produced and presented to the spectator.
b) a support for analyzing the words.

As it happens, we represented the words in the rest of this study in an ascending order of entropy (from the most informative to the least informative). The discussions will focus essentially on words from the middle ranks (representative and informative enough).

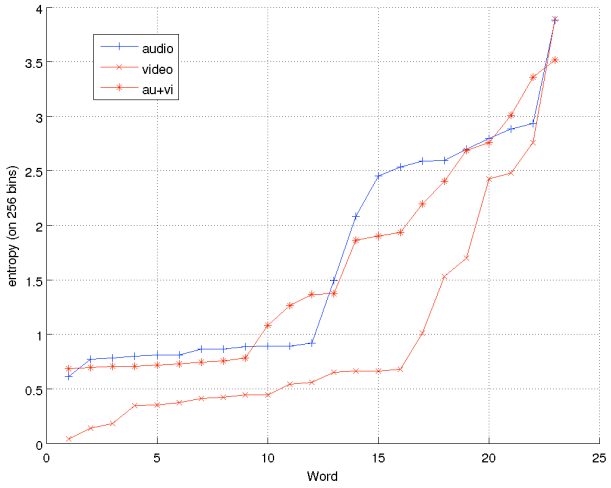

Figure 8. Figure 8. Entropy of each word (23 words per flow, sorted by ascending order) for audio, video and audio coupled with video.

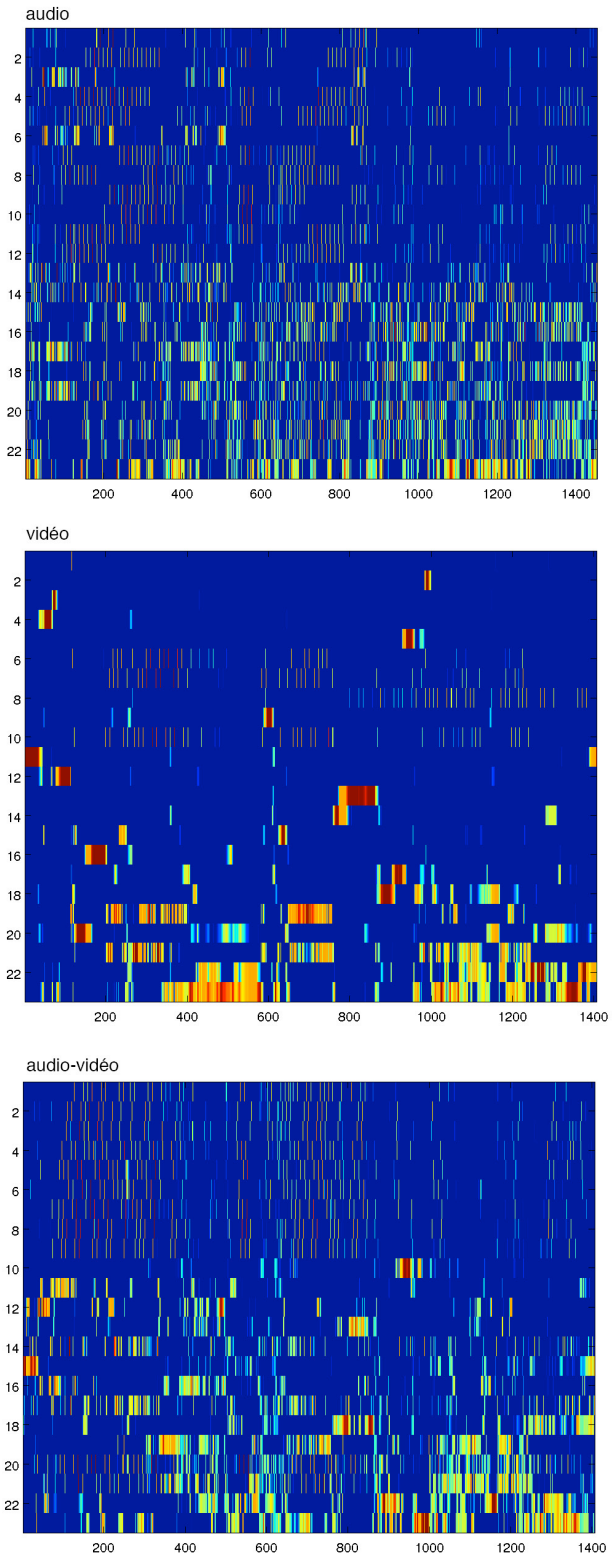

Figure 9. Temporal projection of words from the audio, video and audio-video dictionaries. The abscissa indicates time in frame numbers and the ordinate clues of words sorted by ascending entropy.

6.2. Visualization

An application designed to visualize the results is available on the project's website [1]. It synchronizes the temporal projection of the words of the dictionaries with the sound and visual rendering of the work as shown in figure 10.

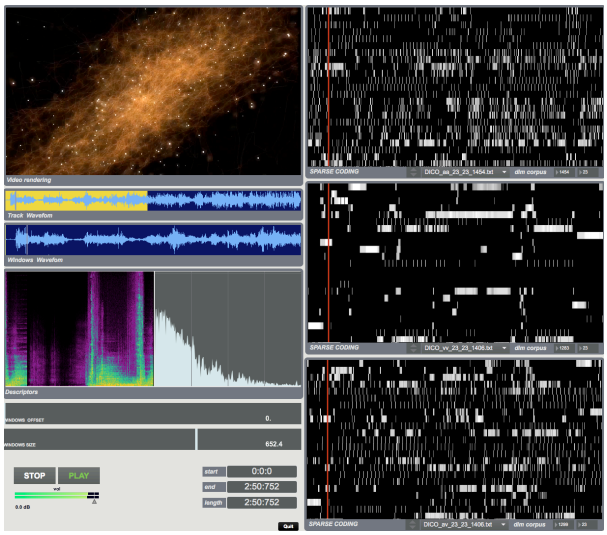

Figure 10. Visualization and control interface of the Sparse Coding matrices (dictionaries)

Sparse Coding is closer to the perceived space than to the written space.
Results are determined by the choice of descriptors included in the corpus that can need to be adjusted and optimized in order to balance the mutual influence of the audio and video.

However the apparition of most uncommon words on the temporal projection corresponds to classes of sound and video events in line with the expressive choices peculiar to the writing. They offer the obvious sensation of representing perceptual units.

This bolsters the project of the interactions lexicon which will be implemented in the future.

7. CONCLUSIONS

The first experimentation substantiates the validity of the concept of analysis of the corpus by Sparse Coding.

The project of multimodal interactions synthesis is about to be implemented as part of the installation version of Iquisme, which offers an experimentation framework suitable for refining the methodology to model the corpus.

Modelling the behavior of 20 000 particles, coupled to a musical writing, will set an example which could be transposed to automatic analysis of scores.

Transcribing with Sparse Coding a heterogeneous data corpus into a symbolic representation can meet many needs in different fields of analysis and of musical writing.

8. REFERENCES

[1] M. Mercier, Iquisme, full presentation, video, sources and download of the dictionary visualizing application fox osx : http://www.maxencemercier.com/iquisme

[2] D. Shiffman, *Nature of code*, book & website, 2012 http://natureofcode.com/book/

[3] Max 6 et Jitter, http://www.cyling74.com

[4] J. Echeveste, J.L. Giavitto, A. Cont, A Dynamic Timed-Language for Computer- Human Musical Interaction. [Research Report] RR-8422, 2013. <hal-00917469>

[5] J.M. Pelletier, *cv.jit computer vision for jitter 1.7* 2010, http://jmpelletier.com/cvjit/

[6] M. Malt M. & E. Jourdan, *Zsa.descriptors: a library for real-time descriptors analysis*. SMC 2008

[7] G. Peeters, *A large set of audio features for sound description (similarity and classification) Cuidado projet report*, (IRCAM), 2004

[8] J. Baboni-Schilingi, & F. Voisin, *Morphologie : Documentation OpenMusic*, Ircam, 1999.

[9] F. Voisin, *dissemblances et espace compositionnels*, JIM 2011

[10] B. Levy, *Visualising OMax* , Paris 6 [DEA ATIAM], 2009

[11] C. Agon, *"OpenMusic : Un langage visuel pour la Composition Assistée par Ordinateur,"* Ph.D. dissertation, Universite´ Pierre et Marie Curie – Paris 6, France, 1998.

[12] D. Schwarz, R. Cahen, S. Britton, *Principles and apllications of Interactive corpus-based concatenative* synthesis, 2006

[13] Kowalski, *Codage parcimonieux* in ERMITES 2011, Glotin Ed, http://glotin.univ-tln.fr/ERMITES