SYNTHESE D'INTERACTIONS MULTIMODALES PARCIMONIEUSES POUR L'ECRITURE DE L'ŒUVRE *IQUISME* ET L'ANALYSE DE SES 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

RÉSUMÉ

Par synthèse d'interactions multimodales, nous entendons resynthétiser des scènes multimodales via un corpus de données hétérogènes. Ce projet repose sur un framework développé en Max et Matlab qui constitue l'atelier d'écriture numérique d'*Iquisme*, une pièce pour soprano, neuf instruments, dispositif électroacoustique et vidéo. Le corpus est constitué des matériaux utilisés pour le premier mouvement de la version de concert d'*Iquisme*. Il s'agit de sons, de vidéos, de matrices de particules 3d et de paramètres de contrôles algorithmiques indexés au déroulement temporel de la pièce. Un apprentissage de dictionnaire par Sparse Coding du corpus établi un lexique d'interactions élémentaires à même de reproduire de manière parcimonieuse toute séquence du 1er mouvement. Munis de ce code, nous proposons alors de réinjecter et redéployer dans le déroulement de l'œuvre un flot de données destiné à contrôler les dispositifs électroacoustique et vidéo.

Une première expérimentation nous à permis de définir une méthodologie pour générer différents dictionnaires issus de l'analyse de l'audio, de la vidéo et de l'audio couplé à la vidéo. L'ensemble du corpus ainsi qu'un outil de visualisation des résultats projetés sur le déroulement temporel des 6 premières minutes du 1er mouvement d'*Iquisme* est disponible en téléchargement [1].

1. PRESENTATION

1.1. Ecriture

Iquisme [1] est le projet d'écriture d'une pièce de concert en trois mouvements pour soprano, neuf instruments, dispositif électroacoustique et dispositif de synthèse vidéo en temps réel.

Sa narration suggère l'évolution organique d'un univers intermédiatique. Cette évolution est le fruit d'une écriture générative dont l'une des problématiques porte sur l'orchestration d'interactions audiovisuelles.

Un ensemble d'applications temps réel réalisé avec Max 6 constitue l'atelier d'écriture numérique d'*Iquisme*.

Le premier mouvement de la pièce a été écrit de manière empirique. Les relations entre la musique et la vidéo sont le fruit de minutieux paramétrages déterminés sur la base de critères esthétiques et narratifs.

Le temps réel fut envisagé dès l'origine du projet afin de satisfaire un travail d'écriture intuitif au rendu immédiat et permettre par la suite une interprétation flexible des instrumentistes. D'autre part les dispositifs électroacoustiques et vidéo seront également déclinés sous forme d'installation interactive. Ces systèmes temps réel permettent une simulation immédiate d'écriture audiovisuelle.

Pour la narration d'*Iquisme*, le sujet porté à l'écran concerne une nuée de particules animées par divers algorithmes [2]. Donner vie à ces particules et les marier au sonore avec les contraintes esthétiques et narratives de la pièce est le fruit de longues recherches et compromis qui ont abouti à un panel de comportements adéquats. Le premier mouvement les expose à l'état embryonnaire. La synthèse d'interactions multimodales va permettre de les développer dans la suite de l'œuvre de manière organique

1.2. Analyse

Les données utilisées et générées par le 1er mouvement sont agrégées dans un corpus hétérogène.

L'analyse du corpus est effectuée en temps différé. Elle s'appuie sur une méthode d'apprentissage par Sparse Coding, nommée également «compress sensing» ou «codage parcimonieux» qui va identifier, classifier puis modéliser les interactions du corpus.

Les relations entre musique et vidéo du 1er mouvement d'Iquisme sont extraites pour constituer un lexique d'interactions multidimensionnel.

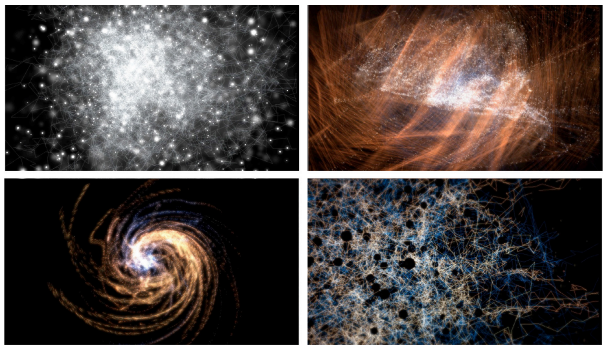

Figure 1. Différents états d'organisations des particules vidéo en 3d générées en temps réel.

1.3. Synthèse

La synthèse d'interactions multimodales s'appuie sur la modélisation des caractéristiques du corpus analysé.

Les interactions identifiées pourront être recombinées pour composer de nouvelles séquences à destination du 2e et 3e mouvement d'*Iquisme*.

Cette démarche devra offrir un degré d'écriture unifié des interactions ainsi que de nouvelles perspectives de corrélation et de convergence non linéaire de leurs paramètres.

L'objectif est de faire proliférer, maximiser et contrôler la direction des processus génératifs des dispositifs électroniques en explorant des possibilités de combinaisons cohérentes avec les équilibres paramétriques définis dans le corpus d'origine. Le lexique qui en résulte associé à un outil graphique permettra en temps réel de représenter et manipuler de grands ensembles de données hétérogènes en fonction des propriétés morphologiques du corpus. La Figure 2 schématise l'architecture du système.

2. DISPOSITIFS TEMPS REEL

Tous les outils pour l'écriture et la performance d'*Iquisme* ont été conçus avec Max[3] sous OSX. Les ressources CPU sont optimisées en répartissant les processus en neuf applications distribuées sur deux machines, l'une dédiée à la vidéo l'autre à l'audio. Les applications communiquent par réseaux et échangent des matrices Jitter. L'audio transit par routage interne à l'ordinateur. Les générateurs vidéo envoient leurs rendus dans une application dédiée à la projection, comme représenté en figure 3. Une application de contrôle global reçoit les ordres de déclenchement midi et gère la synchronisation entre l'audio et la vidéo. Des interfaces dédiées au son et à la vidéo offrent un contrôle intuitif sur les paramètres des programmes de génération temps réel lors des phases d'écritures.

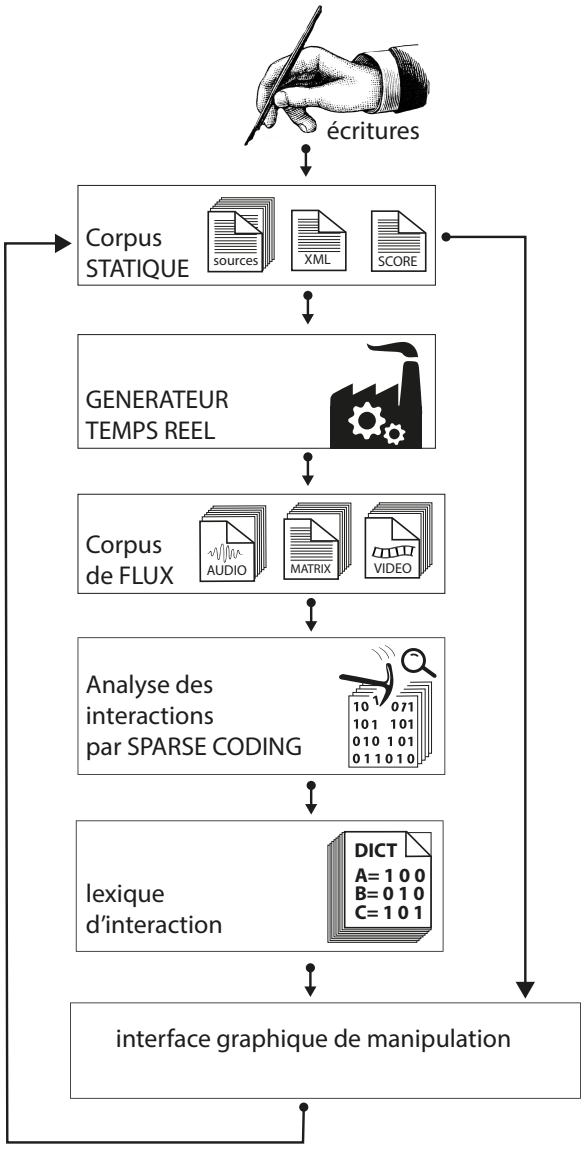

Figure 2. Workflow de la synthèse d'interaction multimodale

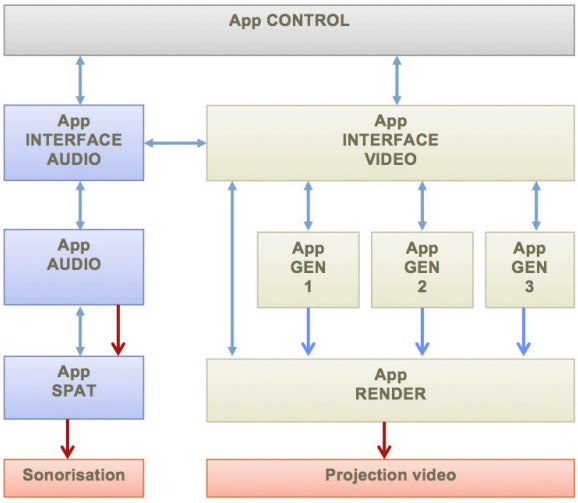

Figure 3. Synoptique des applications temps réel.

Les paramètres définis dans les interfaces graphiques sont inscrits dans des mémoires de présets (pattrstrorage).

Chaque changement de préset peut modifier une centaine de paramètres auxquels s'ajoutent des stratégies de contrôle paramétrique évolutif par l'intermédiaire de déclenchement d'enveloppe (bpf) et le contrôle continu par des descripteurs temps réel.

Dans le cadre d'une simulation, un séquenceur classique avec une démo audio des parties instrumentales permet de synchroniser en midi l'ensemble de l'environnement.

Pour une performance scénique, l'ensemble des processus en temps réel peut s'adapter au jeu des musiciens. Tous les processus algorithmiques et lecture d'échantillons sont conçus pour s'adapter à un tempo variable en concert et en répétition.

3. LE CORPUS

Le corpus est scindé en deux catégories de données.
L'analyse porte sur les flux issus des processus temps réel qui seront segmentés en unités sémantiques sur la base de leurs propriétés morphologique. Les résultats seront indexés par rapport aux données statiques.

3.1. Corpus de données statiques

Il concerne les données issues du travail d'écriture du compositeur. Il s'agit des données statiques qui comprennent la partition instrumentale, la partition électronique et les programmes (patch Max/MSP/Jitter) utilisés par les générateurs temps réel.

Les fichiers de contrôle sont au format Json et gouvernent les paramètres des générateurs. Cet ensemble de fichiers constitue la partition électronique, cette dernière devrait à l'avenir être adaptée pour utiliser Antescofo [4]. Ce corpus a été réalisé de manière empirique, mais respecte une structure permettant d'extraire du contenu de façon automatique.

3.2. Corpus de flux de données

La deuxième catégorie résulte des flux de production des générateurs temps réel. Nous disposons à partir de l'exécution d'une simulation, d'une matrice de particules présentée sous forme de Frame, de fichier audio et vidéo.

3.2.1. Vidéo

Les données vidéo sont agrégées dans une matrice enregistrée dans des fichiers csv en frames successives. Chaque ligne de la matrice est associée à une particule. Les colonnes renseignent sur sa position, direction, vitesse et sa couleur.

Les frames sont complétées par une description de l'image visible en sortie du générateur vidéo. Elle est caractérisée par la position de la caméra OpenGL, le centroid de l'image et l'intensité lumineuse moyenne. Les descripteurs perceptifs proviennent de la bibliothèque cv.jit [5].

La cadence de la génération vidéo temps réel est à 60fps (16ms par frame), mais l'enregistrement des frames est sous-échantillonné à 25fps (40ms par frame).

Les fichiers sont nommés ainsi :
XXX-F_XXXXX-B_XXXXX-P_XXXXX-T.csv

F pour numéro de Frame (spécifique à programme)
B pour nom de fichier de la banque de préset active
P pour *Preset Number and Name* (numéro de mesure)
T pour *Elapsed Time since last preset change*

La description du nom de fichier permet au système d'analyse de référencer chaque frame sur un index temporel en milliseconde, mesure musicale en bar.beat.unit et nom de programme utilisé.
Les frames sont des matrices où chaque colonne est l'atome d'un vecteur à n dimensions.

Pour chaque particule une ligne de la matrice informe sur sa position XYZ dans l'espace 3d, sa direction et sa couleur.

3.2.2. Audio

Les flux audio sont pré-analysés par des descripteurs perceptifs et spectraux issus des librairies pour Max/MSP, Zsa.Descriptors [6] et ircamdescriptor [7].

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

Ces données sont agrégées dans une matrice enregistrée en Frames successives de 40 ms. Les fichiers de chaque frame sont créés en respectant le même schéma que l'enregistrement des frames issus de la vidéo.

Les données audio sont également disponibles en fichiers wav et peuvent faire l'objet d'analyses plus approfondies dans Matlab. Il y a une piste pour chaque instrument, quatre pistes pour la partie électroacoustique et un mixage stéréo de l'ensemble.

4. SPARSE CODING

Le « Sparse Coding » permet d'identifier et de classifier les interactions induites par les différentes typologies d'écritures du son et de la vidéo.

Les Scènes multimodales (sons, vidéo, particules) sont synthétisées dans un espace abstrait qui code les événements majeurs en éléments de base. Ces éléments forment un dictionnaire permettant d'indexer toutes les scènes par rapport aux activités relatives de ces évènements trimodaux élémentaires (sons, vidéo, particules).

L'apprentissage du dictionnaire suit la double contrainte de reproduire au mieux chaque sous-séquence du corpus de l'oeuvre (critère des moindres carrés), tout en recrutant le moins possible les mots du dictionnaire construit (norme L1 du vecteur de recrutement des mots du dictionnaire). L'algorithme d'apprentissage est de la famille des algorithmes LASSO, qui une fois l'apprentissage effectué, permettent une projection assez rapide du corpus sur les mots appris. Le cardinal du dictionnaire est un des paramètres à optimiser, suivant le rendu souhaité.

5. RESULTATS ATTENDU

5.1. Lexique d'interactions

Le lexique d'interaction résulte de combinaisons entre le dictionnaire issu du Sparse Coding et les données paramétriques du corpus statique.

Il se distingue par son formatage et est accessible sous forme de base de données adaptée pour générer une représentation visuelle des interactions modélisées par le Sparse Coding. Les figures 4 et 5 illustrent son contenu.

```
"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. Exemple d'une entrée de la base lexicale.

Les champs sont multiples. Ils exposent la liste d'événements partageant des critères similaires.

ID : identifiant de l'unité lexicale dans le dictionnaire

N.TIME_START : liste d'offsets en temps absolu.

N.MEASURE_START: liste d'offsets en BAR.BEAT.UNIT (mesure)

N.TIME_STOP : liste des offset en ms de fin des événements

N.MEASURE_STOP: liste des offset en BAR.BEAT.UNIT de fin des événements

N.GROUPE_PART : fichier CSV comprenant un index des particules dans chaque frame qui partages les mêmes caractéristiques. format de matrice (ID FRAME, N.ID.PART,

N.DIST_CAM_GROUP : liste de FLOAT distance moyenne de chaque groupe par rapport à la camera OpenGL

N.FUND, N.SPREAD, N.LOUNDNESS sont des exemples de descripteur audio intégrable dans le lexique.

Figure 5. Descriptions des champs d'une entrée de la base lexicale.

5.2. Interface graphique de visualisation du lexique

La représentation du lexique expose les similarités entre des scènes multimodales. Celle-ci permet de retranscrire le corpus de manière symbolique en affichant l'enchaînement des unités lexicales de façon simple : A B A C D A etc….

L'identification de patterns résultant de ces transcriptions offre un champ d'utilisation étendu à l'analyse et à l'écriture musicale [8] [9].

L'interface graphique permettra à terme de manipuler le corpus en recombinant en de nouveaux arrangements la structure transcrite, de la micro à la macro forme.

Les interactions avec cette interface réinjecteront en temps réel dans les systèmes génératifs les paramètres indexés à la base lexicale [10]. Tous les contrôles pourront alors être intégrés à la partition numérique.

Cette interface informe sur la structure de l'œuvre du point de vue de ses relations audiovisuelles. Les descripteurs visuels et sonores indexés au lexique vont permettre de contraindre l'affichage suivant des critères définis.

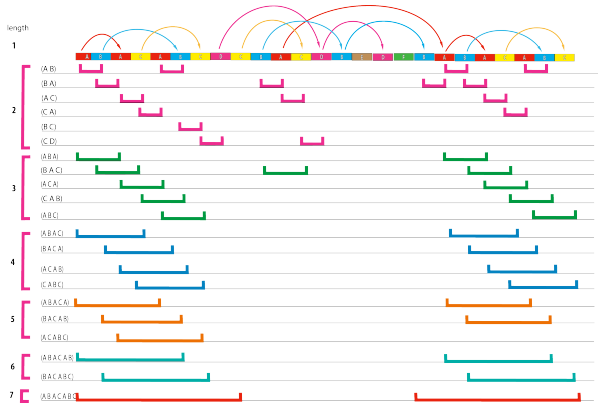

Figure 6. Représentation de l'interface de visualisation des unités lexicales sur un axe temporel.

Elle permettra d'opérer sur sa représentation des transformations et manipulations à la manière de la maquette d'OpenMusic [11].

L'objectif est de faire proliférer, maximiser et contrôler la direction des processus génératifs des dispositifs électroniques en explorant des possibilités de combinaisons cohérentes avec les équilibres paramétriques définis dans le corpus d'origine.

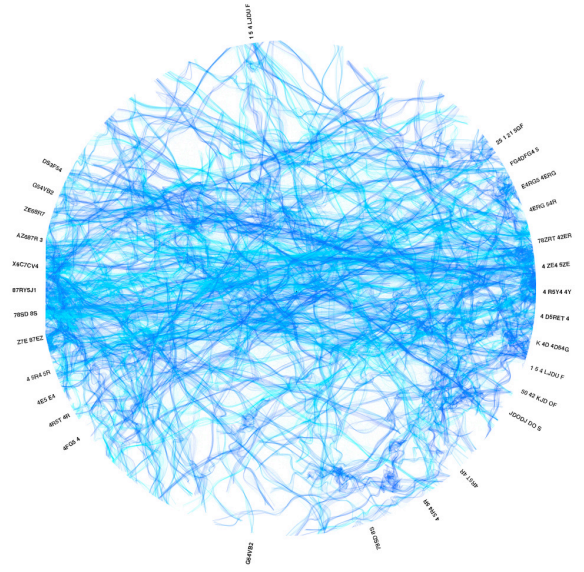

Figure 7. Représentation circulaire de la partition.

5.3. Cartographie et partition interactives

Une version spécifique d'*Iquisme* pour 8 écrans et 8 canaux audio est envisagée sous forme d'installation interactive. Il s'agit d'adapter les dispositifs génératifs électroacoustique et vidéo pour répondre aux besoins scénographiques et narratifs d'une partition navigable interactive. Celle-ci sera symbolisée au sol sous la forme d'un diagramme hyperbolique circulaire.

À la différence de la version de concert, le rapport à la narration est non linéaire. Le recours au lexique est constant pour redéfinir les règles multimodales d'interactions en fonction des mouvements et positions du public dans l'espace scénographique.

Les spectateurs navigueront dans une partition générative qui s'adapte à leurs comportements. Ils expérimenteront par eux-mêmes les logiques interactives qui ont cours dans la version de concert.

La Figure 7 illustre une ébauche de représentation de cette partition, sorte de carte de navigation dans les interactions du corpus. Chaque point à l'extérieur du cercle correspond à une entrée de la base lexicale. L'accès à la carte permet d'activer des réseaux de relations et de les interpréter dans l'espace de l'installation.

Les intersections des lignes sont autant de points de convergences et d'interpolations paramétriques que le public pourra activer.

L'approche gestuelle est similaire aux contrôles de la synthèse concaténative de CataRT [12], à la différence que le corpus concerne ici un ensemble de données paramétriques.

6. PREMIERE EXPERIMENTATION

Cette expérience porte sur l'analyse des six premières minutes du corpus du 1er mouvement d'*Iquisme*.

Pour juger de la cohérence du supervecteur à analyser, nous avons généré trois types de dictionnaires : audio, vidéo et audio associé à vidéo.

6.1. Entropie

Les séquences de mots pour chaque flux (audio, vidéo et audiovidéo) sont analysées suivant la distribution des activités des mots. La distribution de l'activité d'un mot représente la quantité d'information qu'il code suivant la théorie de l'information de Shannon. Nous avons donc calculé l'entropie (en base 2) de la distribution de chaque mot après avoir construit les densités de probabilité de leur activité sur 256 bins. L'entropie maximum possible

est donc 8 logons (ou bits). Il y a 23 x 3 mots en tout. Nous représentons les entropies de tous ces mots dans la figure 8 en les triant par ordre croissant d'entropie de manière indépendante sur chaque flux.

Nous voyons sur cette figure que les mots du flux vidéo sont en général de plus faible entropie que les mots des autres flux. Nous observons aussi que les mots des flux audio et audio-visuel sont d'entropie similaire, avec une tendance à la baisse pour l'audiovisuel, ce qui semble logique étant donné que le flux visuel est de faible entropie. Il faut rester prudent sur l'interprétation de ces valeurs. En effet les mots de plus faible entropie sont parfois des mots peu représentatifs, activés que quelques instants sur toute la composition.

Par contre les mots dès le 5e rang environ ont une fréquence d'activité significative, et nous pouvons assumer le fait que leur entropie représente en effet une mesure de l'information qu'ils portent dans la composition.

Deux usages de cette mesure pourraient alors être proposés :

a) une quantification des volumes d'informations produits et donnés au spectateur.
b) une aide à l'analyse des mots.

Nous avons en l'occurrence représenté les mots dans la suite de cette étude dans cet ordre croissant d'entropie (donc du plus informatif au moins informatif). Les discussions porteront donc surtout sur les mots de rang moyen (assez représentatifs et informatifs).

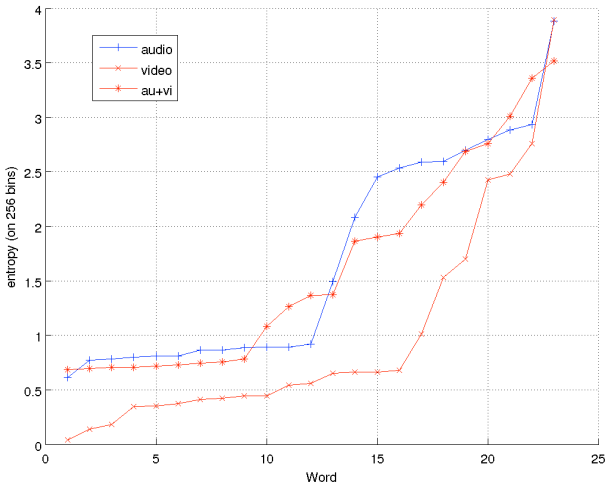

Figure 8. Entropies de chaque mot (23 mots par flux, triés par ordre croissant) pour audio, vidéo et audio couplé à vidéo.

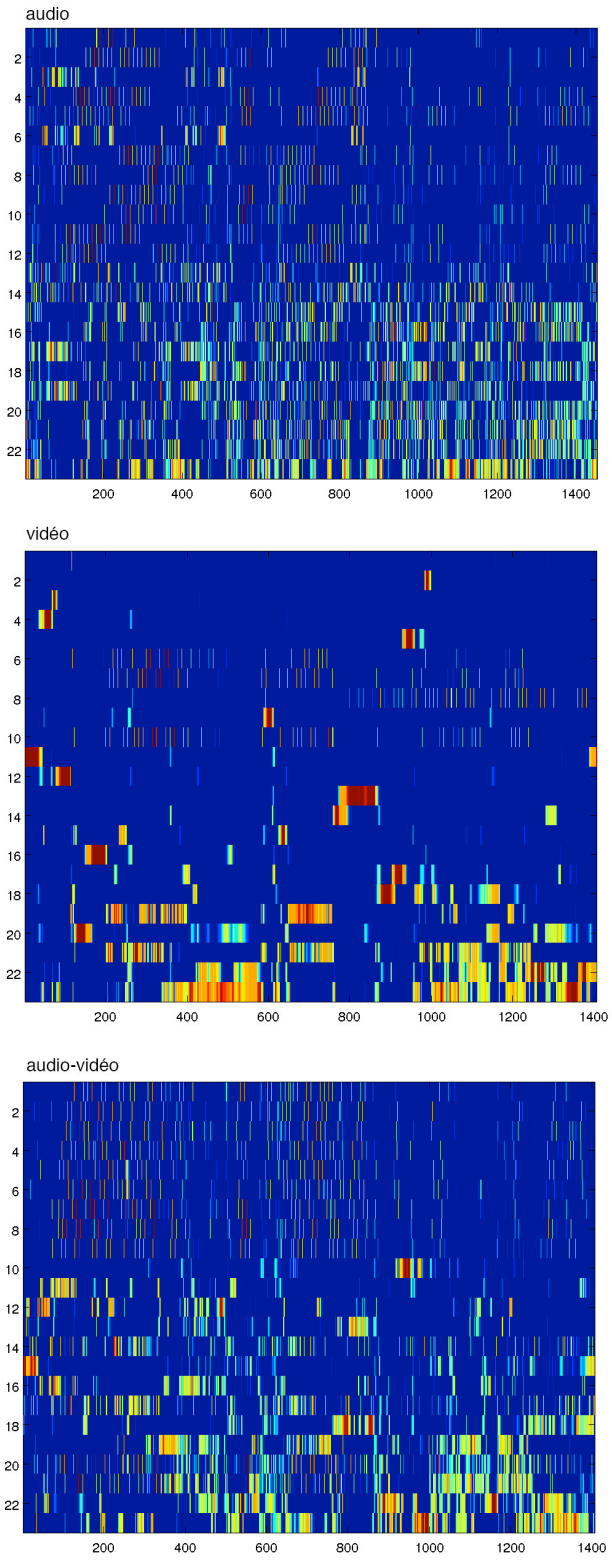

Figure 9. Projection temporelle des mots des dictionnaires audio, vidéo et audio-vidéo. En abscisse le temps en numéro de trame et en ordonnées les indices de mot ordonnés par entropie croissante.

6.2. Visualisation

Une application destinée à visualiser les résultats est disponible sur le site internet du projet [1]. Celle-ci synchronise la projection temporelle des mots des dictionnaires avec le rendu sonore et visuel de l'œuvre comme illustré sur la figure 10.

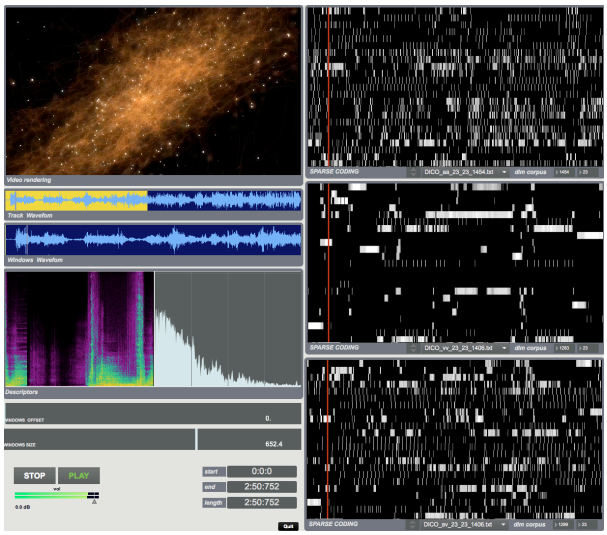

Figure 10. Interface de visualisation et de manipulation des matrices de Sparse Coding (dictionnaires)

Le Sparse Coding est plus proche de l'espace perçu que de l'espace écrit.
Les résultats sont conditionnés par le choix des descripteurs inclus dans le corpus qui peuvent nécessiter d'êtres ajustés et optimisés pour équilibrer l'incidence mutuelle de l'audio et la vidéo.

Néanmoins l'apparition des mots les plus singuliers sur la projection temporelle correspond à des classes d'événements sonores et vidéos en adéquation avec les volontés expressives de l'écriture. Ils offrent la sensation évidente de représenter des unités perceptuelles.

Ceci conforte le projet du lexique d'interactions qui sera mis en œuvre prochainement.

7. CONCLUSIONS

La première expérimentation permet de valider le concept de l'analyse du corpus par Sparse Coding.

Le projet de synthèse d'interactions multimodales est sur le point d'être mis en pratique pour la réalisation de la version d'installation d'*Iquisme*. Celle-ci offre un cadre d'expérimentation adéquate pour affiner la méthodologie de modélisation du corpus.

La modélisation du comportement de 20 000 particules, associé à une écriture musicale est un exemple qui pourra être transposé à l'analyse automatique de partitions.

Transcrire par codage parcimonieux un corpus de données hétérogènes vers une représentation symbolique peut répondre à de nombreux besoins dans les domaines de l'analyse et de l'écriture musicale.

8. REFERENCES

[1] M. Mercier, *Iquisme*, présentation complète, vidéo, sources & téléchargement de l'application de visualisations des dictionnaires pour 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