Sujet de stage Master ATIAM 2015-2016
Représentation symbolique de musique orchestrale pour les méthodes d'apprentissage automatiques

PHILIPPE ESLING ET LÉOPOLD CRESTEL
IRCAM
8 Février 2016

Keywords: Deep learning, Representation learning, embedding space, Music generation

Contexte

Depuis 2015, l'Institut de Recherche et Coordination Acoustique/Musique (IRCAM) est engagé au côté de la Schulich School of Music de l'université McGill à Montréal (SSM) et de la Haute École de Musique de Genève (HEM) dans un projet de recherche dont le but est de construire une théorie moderne de l'orchestration musicale. L'orchestration musicale peut se définir de manière générale comme l'art d'écrire des partitions musicales pour des ensembles d'instruments. L'un des objectifs de ce projet est de développer un système d'orchestration projective automatique dans lequel une partition pour piano est *projetée* sur une partition d'orchestre. Ainsi, les différentes notes présentes dans la partition pour piano sont affectés à différents plan de timbre [1] en conservant la structure harmonique, rythmique et mélodique.

Les récentes avancées en apprentissage automatique laissent entrevoir des résultats prometteurs (génération de séquences vidéos avec contexte [2]). Parmi la grande famille des réseaux de neurones, certains modèles probabilistes génératifs offrent un cadre élégant pour appréhender ce problème et les outils adaptés pour le résoudre. Si le terme réseaux de neurones regroupe un nombre important de modèles en pratique diffèrent, tous ont pour principe fondateur l'inférence statistique de régularités (i.e. corrélations) dans un ensemble de données. Dans le cadre de l'orchestration automatique il s'agit de comprendre les mécanismes sous-jacent en observant des exemples d'orchestrations réalisées par de grands compositeurs. Ces exemples d'orchestrations sont appelés dans un cadre plus général *base d'entraînement*. La qualité du modèle dépend grandement de la qualité et du nombre d'exemples contenus dans la base d'entraînement.

En informatique musicale, on oppose souvent données symboliques (la partition d'orchestre) et données du signal (*e.g.* enregistrement audio). Si nous pensons qu'un système d'orchestration ne peut se contenter uniquement de l'une des deux modalités, nous nous focalisons dans un premier temps sur un modèle entraîné avec des données purement symboliques. Les données musicales symboliques peuvent être représentées de nombreuses manières. En informatique et dans un cadre algorithmique, le format traditionnellement utilisé est celui du *piano-roll* (*cf* figure 1). La série temporelle formée par les colonnes du *piano-roll* peut être utilisée comme donnée d'entraînement pour un réseau de neurone.

Problématiques : manque de données et parcimonie

La collecte de données est essentielle en vue de construire un modèle d'orchestration automatique performant. Deux difficultés surgissent lors de la construction d'une base de données

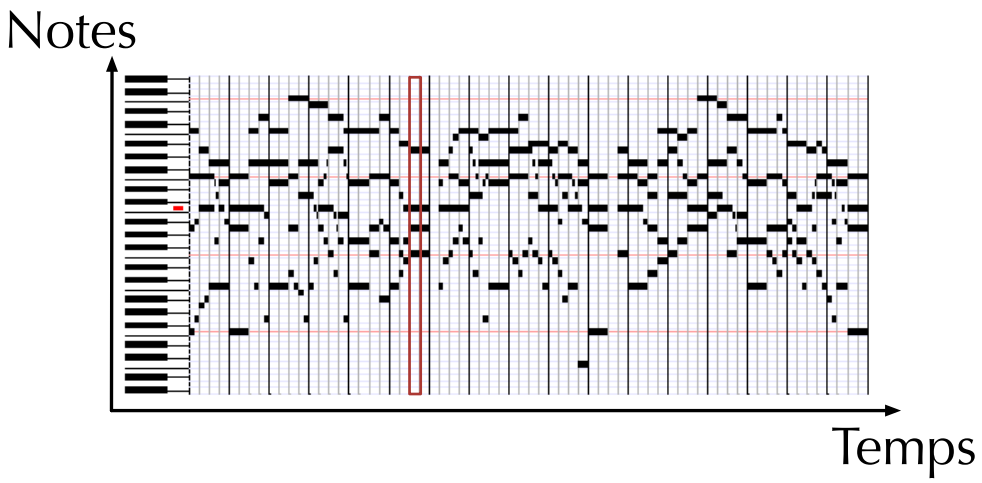

Figure 1: Représentation *piano-roll*. Le temps est discrétisé selon un quantum de temps de référence (typiquement une fraction de la noire). Les fréquences sont discrétisées selon le tempérament égal en vigueur dans la musique classique occidentale (12 notes de DO à SI et une octave). Le *piano-roll* est une matrice P binaire dont la valeur $P(n,t)$ est égale à 1 si la note n est jouée à l'instant t et à 0 sinon.

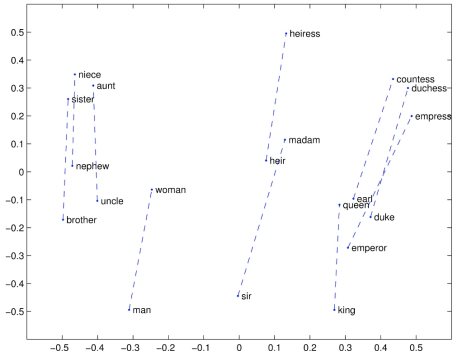

Figure 2: Visualisation de la projection sur 2 dimensions d'un espace de plongement (*embedding)* pour des vecteurs de mots selon des axes associés au genre [3]

symboliques pour l'orchestration : la taille des bases de données et la nature parcimonieuse des vecteurs manipulés.

Plus le nombre de données d'entraînement est grand et meilleur sera le modèle. Recopier au format numérique (MIDI ou XML) des partitions d'orchestre est fastidieux et peu de compositeurs se sont attelés à cette tâche. La première difficulté est donc la relative paucité de base de donnée de qualité et taille suffisantes.

Par ailleurs, la représentation en *piano-roll* découle d'une discrétisation du temps et de la hauteur des fréquences certes naturelle, mais peu adaptée aux techniques d'apprentissage que nous souhaitons utiliser. Si on considère un d'orchestre composé d'une vingtaine de pupitres différents, chacun pouvant jouer 40 hauteurs de notes différentes, on obtient à chaque instant un vecteur de taille 800 ne contenant qu'une vingtaine de coordonnées non nulles. Seul une faible partie de l'ensemble possible de ces vecteurs sera contenu dans la base de départ.

Sujet de stage

Nous proposons de mener au cours de ce stage une réflexion sur les différentes manières de représenter l'information contenue dans une partition pour orchestre. Les deux problématiques soulevées précédemment engagent ainsi deux axes de recherches principaux :

Data augmentation : Fréquemment employée en reconnaissance d'image, cette technique consiste à appliquer des transformations et dégradations sur les données de la base d'entraînement afin d'en augmenter artificiellement la taille. En image, ces transformations sont généralement des obstructions (mise à zéro d'une partie des pixels de l'image), changement d'échelle, transvections (*shear mapping*). Quelles seraient ces transformations pour la musique orchestrale ?

Chord-embedding : une représentation "naïve" des mots en traitement du langage naturel consiste à prendre un vecteur de la taille du dictionnaire étudié et associer chaque dimension à un mot (*one-hot encoding*). En anglais, pour un dictionnaire restreint de 30000 mots, chaque mot est alors représenté par un vecteur de taille 30000 constitué uniquement de zéro sauf pour une coordonnée égale à 1, ce qui est extrêmement parcimonieux. Une solution consiste alors à encoder les vecteurs dans un espace de dimension plus petit que l'espace de départ. Cet espace est appelé dans la littérature *embedding* [4], et peut être recherché de manière automatique en fixant comme objectif la prédiction du contexte entourant le mot (cf figure 2). Les espaces ainsi construits exhibent généralement d'intéressantes relations entre transformations géométriques et valeur sémantique. Pour de nombreuses raisons (similitude entre les différents instruments, intensité des notes, simultanéité de leurs occurrences dans la musique polyphonique) l'analogie avec le langage est limitée, et il ne s'agira pas d'effectuer une simple adaptation des techniques utilisées en traitement du langage naturel, mais de proposer une méthode spécifiquement ciblée à la nature des données orchestrales.

Déroulement du stage et compétences requises

Le stage s'étend sur une durée de 2 mois à plein-temps, aménageable sur 6 mois. Le stage s'articulera ainsi autour de points de travail suivants :

- Data augmentation (2 semaines)
  - Bibliographie des méthodes de *data augmentation* utilisées en traitement d'image.
  - Interview avec des compositeurs (transformations admissibles ?)
  - Proposition et implémentation d'un ensemble de transformations spécifiques à l'orchestration
- Chord-embedding (1 mois et demi)
  - Bibliographie des techniques employées en traitement du langage naturel [4, 5]
  - Développement et implémentation de plusieurs algorithmes de construction des espaces de enchâssement (*embedding*) adaptés à l'orchestration.
  - Évaluation qualitative, en étudiant notamment l'effet de diverses transformations géométriques dans l'espace de enchâssement
  - Évaluation quantitative sur une tâche de prédiction à court-terme

Le candidat doit avoir de solides bases en informatique et mathématique. Une connaissance des méthodes d'apprentissage automatique est souhaitable. Une bonne connaissance de la musique et si possible de la composition orchestrale est essentielle. Le langage de programmation utilisé est laissé libre, avec une préférence pour Matlab ou Python.

References

1. Stephen McAdams. Timbre as a structuring force in music. In *Proceedings of Meetings on Acoustics*, volume 19, page 035050. Acoustical Society of America, 2013.

2. Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032. ACM, 2009.

3. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543, 2014.

4. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

5. Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *CoRR*, abs/1506.06726, 2015.