Techniques de traitement automatique du langage naturel appliquées aux représentations symboliques musicales

Mikaela Keller, Kamil Akesbi, Lorenzo Moreira, Louis Bigo

TECHNIQUES DE TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL APPLIQUÉES AUX REPRÉSENTATIONS SYMBOLIQUES MUSICALES

Mikaela Keller
Univ. Lille, Inria
CNRS, Centrale Lille
UMR 9189 CRIStAL
F-59000 Lille, France
mikaela.keller@univ-lille.fr

Kamil Akesbi *Lorenzo Moreira*
Univ. Lille
CNRS, Centrale Lille
UMR 9189 CRIStAL
F-59000 Lille, France

Louis Bigo
Univ. Lille
CNRS, Centrale Lille
UMR 9189 CRIStAL
F-59000 Lille, France
louis.bigo@univ-lille.fr

RÉSUMÉ

La discipline du Traitement Automatique du Langage Naturel (TALN) a connu d'importants progrès ces dix dernières années grâce aux avancées de l'intelligence artificielle et en particulier des réseaux de neurones profonds. Ces techniques sont aujourd'hui largement utilisées pour l'extraction de connaissances, l'analyse de sentiments ou encore la traduction automatique de textes. Les similarités structurelles et conceptuelles entre le langage naturel et la musique ont motivé de nombreuses initiatives de recherche visant à adapter les outils du TALN pour le traitement de données musicales symboliques ou audio. Ces démarches ont fourni des résultats prometteurs, notamment dans les domaines de l'analyse et la génération automatique de musique. Au-delà de leur performance, le présent projet vise à étudier le fonctionnement interne de deux de ces modèles, les *plongements de mots* et les *transformeurs*, ainsi que leur aptitude à s'adapter à des données musicales plutôt que textuelles. Ces expériences bénéficient à notre maîtrise de ces outils adaptés à la musique et contribuent à clarifier de nombreux parallèles entre le langage naturel et le langage musical.

1. INTRODUCTION

1.1. Traitement Automatique du Langage Naturel et informatique musicale

Un parallèle existe entre le texte et la partition vus comme une séquence d'éléments constituant un tout expressif. Une mélodie constituée de notes, ou une séquence d'accords de jazz, peuvent être comparés à une phrase constituée de mots. Il est courant en musique de parler de *langage tonal*, de *narration*, de *fin de phrase*, par exemple en comparant les différents types de cadences aux éléments de ponctuation dans le texte, notamment la virgule et le point [1]. Concevoir la musique comme un langage suggère des apports mutuels potentiels entre les recherches en Traitement Automatique du Langage Naturel (TALN ou *NLP* pour *Natural Language Processing*) et en informatique musicale. Les domaines du TALN et de l'informatique musicale ont d'ailleurs des applications communes, comme la génération automatique de contenu ou encore la classification stylistique par époque, genre, auteur/compositeur. Le transfert de style (transformer un contenu dans le style d'un autre) a suscité des travaux en TALN [2] comme en musique [3, 4].

En TALN comme en musique, ces applications se distinguent par une prise en compte variable de la complexité des relations entre les éléments dans les séquences de données. Par exemple, la catégorisation de textes par thèmes est souvent abordée à l'aide de modèles qui prennent simplement en compte la co-occurrence des mots alors que la détection de sentiments nécessite des abstractions plus profondes. De même en musique, la détection de style peut généralement s'étudier sur des fenêtres temporelles moins large [5] que la détection de la structure qui nécessite le recours à des modèles capturant les dépendances à plus long terme entre les notes [6].

Certains domaines d'applications du TALN, comme la traduction automatique, la génération de résumé, ou la détection d'humour et d'ironie sont plus difficilement interprétable dans le domaine musical. La recherche de ces parallèles ouvrent toutefois la voie à d'intéressantes réflexions.

Une grand nombre de techniques modernes en TALN visent à représenter une phrase dans un espace abstrait rendant compte de sa sémantique. Ce type d'espace a un intérêt pour la traduction automatique étant donné qu'une phrase et sa traduction dans une autre langue sont censées y être représentées par un même point. Ces espaces font aussi apparaître le principe d'analogie, avec par exemple un même vecteur permettant de se déplacer du mot *reine* au mot *roi* et du mot *femme* au mot *homme*. Ces principes s'appliquent moins naturellement dans le domaine musical, où la notion de sens est plus subjective. Si le terme de *langage* est largement utilisé pour qualifier la musique, la question de savoir si deux phrases musicales distinctes ont un sens similaire ouvre elle aussi de nombreuses questions.

En texte comme en musique, le *contexte* d'un élément d'une séquence, c'est à dire les éléments qui l'environnent,

contribuent généralement au *sens* de cet élément. Cette propriété est particulièrement vrai en musique où la fonction d'une note ou d'un accord pris de manière isolée change radicalement selon la tonalité dans laquelle il se trouve.

1.2. Représentations séquentielles

L'application de techniques dédiées au texte sur des données musicales ne se fait toutefois pas de manière directe. En effet, les algorithmes de TALN sont adaptés en premier lieu à des représentations textuelles consistant en des séquences de mots. Malgré sa nature temporelle, la musique ne se représente pas systématiquement sous la forme de séquences. La polyphonie de la musique fait en effet apparaître des notes qui sont simultanées et qui se chevauchent. L'utilisation de méthodes de TALN sur des données musicales nécessite donc soit une adaptation des modèles afin de leur permettre de considérer des structures de données plus complexes que des séquences, soit une approximation de l'information musicale afin de la rendre strictement séquentielle.

Une partition pour plusieurs instruments peut par exemple être réduite en une séquence de symboles d'accords qui fait abstraction du rythme et de la conduite des voix, ou en une séquence de tranches de notes qui fait abstraction des notes tenues. Il est intéressant de remarquer que dans le domaine du TALN, les linguistes effectuent eux-aussi des abstractions en regroupant les différentes formes d'un mot (genre, nombre, temps) par une forme prototypique appelé *lemme*.

Cet article décrit deux expériences visant à analyser le comportement de techniques de TALN lorsqu'on les utilise sur des représentations symboliques musicales. La première expérience porte sur les *plongements de mots* et la seconde sur le mécanisme d'attention dans les réseaux de neurones.

2. TRAVAUX EXISTANTS

L'utilisation de techniques d'algorithmique du texte, et plus généralement de traitement du langage, sur des données musicales a fait l'objet de nombreux travaux en informatique musicale. L'algorithme de Mongeau Sankoff [7] propose une mesure de similarité musicale se basant sur un calcul de distance entre séquences d'éléments élaboré à l'origine pour des séquences textuelles. Récemment, les scores TF-IDF (*term frequency-inverse document frequency*) permettant d'évaluer l'importance des mots contenus dans un document, ont par exemple été utilisés pour l'analyse des modes dans le plain-chant [8] et pour l'étude du style musical arabo-andalou [9].

D'importants progrès ont récemment été réalisés dans le domaine du TALN grâce à l'élaboration de méthodes d'intelligence artificielle faisant notamment intervenir des réseaux de neurones profonds. L'adaptation de ces techniques dans un cadre musical fait partie des sujets très attractifs du moment dans la communauté de l'informatique musicale. Cette discipline fait l'objet d'un workshop dédié (NLP4MusA : *Natural Language Processing for Music and Audio* [1]) dont la première édition [2] a eu lieu en 2020 en évènement satellite de la conférence internationale ISMIR (*International Symposium on Music Information Retrieval*).

Le principe de plongement de mots (*word embedding*) qui est à l'origine du modèle *word2vec* [10] a été appliqué sur des séquences d'accords [11], de tranches de notes [12] et expérimenté dans le cadre de substitutions musicales [13].

Des architectures de réseaux de neurones profonds dédiés au traitement de données séquentielles et initialement élaborés pour le TALN, comme les réseaux *transformeurs* [14], ont largement été adaptées ces dernières années pour la modélisation de musique. Ces travaux ont entre autres été appliqués à l'analyse harmonique [15] et à la génération de musique [6, 16, 17, 18]. La performance de ces modèles, permettant souvent de dépasser les résultats de l'état de l'art, attire cependant généralement plus l'attention que leur fonctionnement interne et leur aptitude à s'adapter spécifiquement au langage musical. Certaines initiatives semblent cependant montrer un intérêt pour l'ouverture de ces modèles complexes, souvent comparés à des "boites noires", permettant de visualiser la manière dont ils s'adaptent au langage musical [19]. Le présent projet de recherche s'inscrit dans cette démarche.

3. EXPÉRIENCE SUR PLONGEMENTS STATIQUES

3.1. Representations vectorielles des mots

Effectuer des tâches de TALN nécessite généralement de passer par des représentations numériques spécifiques des mots afin de faciliter leur manipulation par des algorithmes informatiques. Une première représentation primitive, appelée *sac-de-mots* (*bag-of-words*), consiste à considérer un espace multi-dimensionnel dont les dimensions sont respectivement associées à chacun des mots du vocabulaire V considéré), et à voir le contenu lexical d'un texte comme un vecteur dans cet espace. Le désavantage de cette représentation naïve est que la distance euclidienne entre deux mots quelconques y est toujours la même : contre toute intuition *mer* et *océan* sont à la même distance que *mer* et *radis*.

Une hypothèse datant des années 50, dite hypothèse distributionnelle [20], suppose que des mots dans des contextes similaires devraient avoir des “sens” similaires, et donc des représentations avec des distances moindres que des mots apparaissant dans des contextes différents. Il existe diverses façons étant donné un corpus $\mathcal{C}$ de textes utilisant un vocabulaire V de construire des representations dite *distributionnelles* des mots. Parmi les plus récentes certaines, comme *word2vec* [10], font appel à l'entraîne-

1 . https://sites.google.com/view/nlp4musa

2 . Il est à noter toutefois que la majeure partie des travaux présentés lors de ce workshop rendent comptent de l'utilisation de méthodes de TALN pour l'extraction de connaissances sur des textes portant sur des sujets liés à la musique et non sur des représentations séquentielles musicales comme c'est l'objet du présent projet.

ment d'un réseau de neurones pour modéliser le langage utilisé dans un corpus. Cet entraînement ne nécessite pas des données annotées mais s'apparente plutôt à l'apprentissage d'un modèle de langue, c'est à dire que la supervision provient des occurrences de langage observées. Une fois le modèle entrainé, la sortie d'une des couches intermédiaires du réseau sera utilisée comme vecteur de représentation du mot passé en entrée, avec l'idée que la représentation du mot *plongé* (*embedded*) dans cet espace sera similaire aux représentations des mots apparaissant dans des contextes semblables dans le corpus d'apprentissage.

D'autres approches [21, 22] accumulent les fréquences de co-apparition des mots et construisent des représentations parfois appelées *sac-de-contextes* (*bag-of-contexts*) ([23]). Dans ces approches, un vocabulaire représentatif $V_c \subset V$ est sélectionné, et tout mot w est représenté par un vecteur de taille $|V_c|$ dont les composantes rendent compte de la fréquence de co-occurrence de w avec les mots de V_c. Une co-occurence de deux mots a lieu chaque fois que l'un apparait dans le contexte de l'autre. Le contexte d'un mot w peut désigner de multiples choses, le plus souvent il est définit comme étant les k mots précédents et succédants w (c.à.d. $2k$ mots, sans l'information de leur distance à w) dans la séquence de mots. On obtient ainsi une matrice S de dimensions $|V| \times |V_c|$, où S_{ij} représente la fréquence de co-occurrence des mots $w_i \in V$ et $c_j \in V_c$. Il y a différentes façons de caractériser ce degré d'association. Parmi celles-là, nous nous sommes intéressés à l'information mutuelle ponctuelle positive (PPMI : *positive pointwise mutual information*) [23] $S_{ij} = \text{PPMI}(w_i, c_j)$ définie par :

$$\text{PPMI}(w,c) = \begin{cases} \text{PMI}(w,c) & \text{si PMI}(w,c) \geq 0 \\ 0 & \text{sinon} \end{cases}$$

et

$$\text{PMI}(w,c) = \log \frac{N_{wc} N}{N_w N_c}$$

où N est le nombre total d'occurrences dans le corpus, N_{wc} est le nombre de fois que le mot c apparait dans le contexte du mot w, N_w et respectivement N_c sont le nombre de fois que les mots w et c apparaissent dans le corpus.

Que ce soit en collectant des statistiques sur un corpus comme on vient de le voir où en entraînant un réseau de neurones comme expliqué plus haut, les représentations obtenues sont des représentation dites *statiques* des mots. Lors de l'étude d'une phrase, chaque mot possède une représentation unique calculée à partir du corpus d'apprentissage et indépendante de son contexte dans la phrase étudiée. On s'intéressera dans la section 4 à des représentations dites contextuelles qui, au contraire, varient en fonction du contexte dans lequel apparaissent les mots.

3.2. Plongement de tranches musicales

L'approche des plongements lexicaux en musique renouvelle la question de similarité musicale en l'abordant du point de vue du contexte des objets musicaux plutôt que des objets eux-mêmes. Pour être appliquée de manière analogue au TALN, cette approche fait toutefois l'hypothèse d'une représentation exclusivement séquentielle de la musique. Cela peut se justifier dans le cas d'une mélodie ou d'une séquence de symboles d'accords, mais nécessite des approximations dans le cas de la musique polyphonique.

L'expérience décrite dans cette section a pour motivation d'évaluer le principe de plongement d'accords, plus spécifiquement le plongement de *tranches de notes*, qui s'extraient automatiquement des partitions numériques, et permettent donc une étude systématique sur de larges corpus. Nous appliquons cette approche à travers une expérience de transformation de partition par substitution de tranches.

La représentation sous forme de tranches musicales (parfois appelée *salami slicing* [24]) consiste à réduire un extrait polyphonique sous la forme d'une séquence d'ensembles de hauteurs simultanées, chacun associé à une durée. Une nouvelle tranche commence chaque fois qu'une nouvelle hauteur apparait ou disparait. Si cette représentation permet de manipuler de manière systématique un extrait musical, même polyphonique, sous la forme d'une séquence, il ne permet que difficilement de conserver l'information des notes tenues entre tranches successives [3]. L'utilisation de cette représentation pour des tâches de génération ou de substitution a ainsi tendance à produire un excès indésirable de notes répétées.

3.3. Co-occurrence de tranches et hypothèses de substitutions

La représentation d'un corpus de partitions sous la forme de séquences de tranches de notes permet d'appliquer la méthode présentée dans la section 3.1 et de calculer systématiquement la fréquence de co-occurrence de toute paire de tranches de notes apparaissant dans ce corpus. Cette approche est ici utilisée pour effectuer des transformations de partition par substitution de tranche.

À partir d'un corpus de référence $\mathcal{C}$, on calcule en premier lieu la matrice $S_\mathcal{C}$ qui rend compte de la co-occurrence de toute paire de tranches de notes dans ce corpus. On choisit par ailleurs une séquence de tranches T dans laquelle on désire effectuer une substitution. À chaque position i de T, il est possible de calculer un plongement rendant compte des tranches constituant le contexte de la position i dans la séquence T. La matrice $S_\mathcal{C}$ permet ensuite d'identifier les tranches de $\mathcal{C}$ qui apparaissent dans un contexte similaire au contexte de la position i dans T. Ces tranches constituent des candidats pour substituer la tranche à la position i dans T. Deux types de substitutions apparaissent alors : les substitutions que nous appelleront *conventionnalisantes* consistent à remplacer dans T

3 . Une méthode consiste à ajouter dans chaque tranche un ensemble des notes tenues. Cette approche est notamment utilisée pour l'encodage vectoriel de corpus afin de faciliter leur utilisation par des algorithmes d'apprentissage automatique [25]. Cette méthode a toutefois l'inconvénient d'élargir l'espace des valeurs prises par les tranches et donc de compliquer leur identification.

la tranche à la position i par une tranche dont le contexte moyen, calculé à partir du corpus, est d'avantage similaire au contexte de la position i dans T. À l'inverse, les substitutions *originalisantes* consistent à remplacer dans T la tranche à la position i par une tranche dont le contexte moyen est moins similaire au contexte de la position i dans T.

3.4. Expériences de substitution

Cette section présente une expérience [4] visant à évaluer la méthode de substitution présentée dans la section 3.3. Un ensemble de 355 chorals de J.-S. Bach [5] a été utilisé comme corpus de références. Chaque choral a été représenté sous la forme d'une séquence de tranches de notes comme indiqué dans la section 3.2 afin de produire la matrice de co-occurrence S. Un total de 20 tranches issues de séquences du corpus ont été substituées par des tranches de contexte similaire. La figure 1 illustre un exemple de substitution sur le choral BWV 353 de J.-S. Bach. La partition de gauche fait apparaitre l'extrait avec la tranche originale (Sib3, Ré4, Sol4, Sib4), la partie droite l'extrait avec la tranche substituée (Ré4, Ré4, Fa#4, La4). Cet exemple illustre la tendance de la réduction en tranches à briser les notes tenues évoquée dans la section 3.2.

Une étude regroupant 54 participants a été menée pour étudier les effets perceptifs de ces substitutions. 12 participants déclarent être familiers avec les chorals de Bach. Pour chaque substitution, un extrait audio de quelques secondes a été créé afin de faire entendre la tranche dans son contexte. Les extraits ont une durée moyenne de 10 secondes. La tranche substituée est généralement située au milieu de l'extrait, sauf dans les cas où la substitution a lieu au début ou à la fin d'une phrase. Les extraits sont présentés aux participant dans leur version originale et dans leur version transformée. Pour chaque paire, il est demandé au participant d'indiquer dans un premier temps si il a perçu une différence. Si oui, il lui est demandé d'indiquer si un des deux extraits semble *sonner* le mieux, ou si il n'a pas d'avis sur la question.

Les 20 substitutions ayant été présentées aux 54 participants, un total de 1080 comparaisons ont été effectuées. Dans 950 cas (88%), la différence entre l'extrait original et l'extrait transformé a été perçue. Parmi ces cas, 906 ont fait l'objet d'un classement des deux extraits et 34% de ces classements privilégient l'extrait transformé. Il n'est pas surprenant que le choix des participants se porte majoritairement sur les extraits originaux. Toutefois, le fait que la proportion d'extraits transformés dépasse un tiers des comparaisons semble confirmer la capacité du modèle à effectuer des substitutions raisonnables en ne se basant que sur la similarité de leur contexte.

4. https://enquetes.univ-lille.fr/index.php/178616?lang=fr

5. téléchargé au format MIDI sur le site http://kern.ccarh.org/

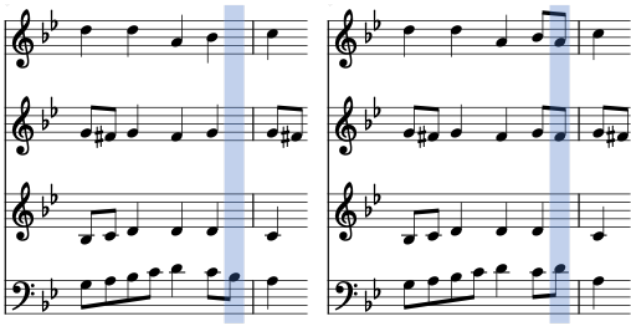

Figure 1. Exemple de substitution présentée aux participants, extraite du choral BWV 353 de J.-S. Bach (1ère mesure).

4. EXPÉRIENCE SUR PLONGEMENTS CONTEXTUELS

Cette section présente une série d'expériences visant à analyser le fonctionnement d'un réseau de neurones *transformeur* sur des données musicales. Les transformeurs constituent un type de réseaux de neurones profonds utilisant le principe d'*attention mutuelle* [14]. Ils sont souvent considérés comme une alternative aux réseaux de neurones récurrents, permettant une prise en compte plus fine des relations liant les éléments constituant une séquence. L'objet de cette expérience est d'"ouvrir la boîte noire" d'un transformeur entraîné sur des données musicales, et d'apporter des interprétations musicales aux valeurs de ses coefficients internes.

4.1. Le mécanisme d'*attention*

Dans un phrase textuelle, les relations des mots ne se réduisent pas à leurs contiguïtés, les mots qui se suivent ne sont pas nécessairement ceux qui ont le plus d'influence les uns sur les autres). Par exemple dans la phrase :

"Il allait cueillir cette **pomme** *à la belle couleur rouge quand il a vu qu'***elle** *avait un ver"*

le pronom (elle) et l'entité (la pomme) à laquelle il se rapporte se trouvent à distance dans la séquence.

Certains mots ont plus d'influence sur le sens global de la phrase que d'autres. C'est le cas par exemple de la négation : le simple mot "pas" inverse à lui seul le sens de la phrase.

La transposition de ce principe dans le domaine musical consiste à considérer que les notes et accords constituant une séquence musicale contribuent différemment au *sens* que l'on en perçoit. L'attribution d'un *sens* à une phrase musicale relève cependant d'un processus plus subjectif que pour une phrase textuelle. Même si ce sens n'est pas clairement définit, il semble toutefois raisonnable de considérer que les notes peuvent jouer un rôle différent, et même avoir une importance variable au sein d'une phrase musicale. Il n'est pas rare de qualifier les notes de musique d'après leurs relations mutuelles ou d'après une fonction qu'on leur assigne. C'est le cas pas exemple des *notes*

fondamentales et réelles des accords qui sont le signe de progressions harmoniques sous-jacentes, ou encore à une échelle plus fine des *notes modulantes* qui annoncent une transition vers une nouvelle tonalité.

Parmi les multiples types de réseaux de neurones utilisés en TALN, les transformeurs [14] ont fourni des résultats particulièrement prometteurs en modélisation du langage, par exemple pour la conception de systèmes de questions-réponses [26] ou pour la traduction automatique [14]. Le fonctionnement des transformeurs repose sur le principe d'*attention mutuelle*, qui incite un modèle, lors de son entraînement, à évaluer l'influence mutuelle des termes successifs d'une séquence, d'où le terme original de *self-attention*. Le modèle de transformeur décrit dans [14] est entraîné comme un modèle de langue : le début d'une phrase est donné au réseau de neurone qui doit prédire la suite de la séquence (ou la séquence dans une autre langue dans le cas de la traduction automatique). Le mécanisme d'attention est utilisé une première fois pour "encoder" les relations entre les éléments de la séquence, puis de nouveau pour "décoder" les relations avec les élements à prédire.

Les réseaux de neurones transformeurs encodent l'attention mutuelle sous la forme de matrices de coefficients qui sont ajustés itérativement par rétro-propagation au cours d'une phase d'entrainement sur un ensemble de séquences.

Transposé dans le domaine des représentations musicales, le mécanisme d'attention offre une approche originale pour comparer la contribution de chacune des notes pour la modélisation statistique d'une séquence musicale.

Malgré leur performance, les réseaux de neurones profonds sont souvent critiqués pour leur opacité qui limite l'identification des abstractions qu'a re-constitué le modèle au cours de son entraînement et qui lui permettent de prendre des décisions correctes. Le travail présenté dans cette section vise à examiner les représentations internes d'un transformeur entraîné sur un corpus de musique classique pour piano. L'objectif est d'analyser les représentations internes au modèle, apparaissant à travers les coefficients des *unités d'attention* regroupées en matrices désignées sous le terme de *têtes d'attention*, elles-même regroupées dans les différentes couches du réseau de neurones. L'observation de ces valeurs a pour but d'identifier les éléments du langage musical sur lesquels a tendance à se focaliser le mécanisme d'attention, puis de les comparer avec des règles issues de la théorie musicale.

4.2. Expériences

L'expérience décrite dans cette section se base sur le modèle *Music Transformer* [6] élaboré par l'équipe Magenta de Google Brain. Une contribution majeure de ce système est l'élaboration d'un mécanisme d'*attention relative* privilégiant la position relatives de deux éléments d'une séquence musicale plutôt que leurs positions absolues. Ce modèle a été élaboré dans le but de générer de la musique faisant apparaitre une structure long terme cohérente. L'expérience présenté dans cette section a pour objet de ré-utiliser le code du modèle *Music Transformer* non pas dans un but de génération, mais pour l'étude du mécanisme d'attention via l'observation des représentations internes du modèle. Le modèle a été entrainé à l'aide de l'ensemble de données *Maestro* [17] qui inclue près de 200 heures de piano au format MIDI dans un répertoire majoritairement classique incluant des compositeurs du XVII ème siècle au XX ème siècle. Les messages MIDI constituant les fichiers de l'ensemble de données sont converties sous la forme d'entiers $n \in [0, 387]$ à leur tour représentés sous la forme de vecteurs *one-hot* communément utilisés pour l'entrainement des réseaux de neurones.

Le transformeur utilisé pour cette expérience est constitué de 6 couches comprenant chacune 4 têtes d'attention. Il prend en entrée des séquences de L (≤ 2048) événements MIDI représentés par des vecteurs one-hot en 388 dimensions. Chaque évenement MIDI subis un plongement de position en 256 dimensions qui sont ensuite passées en parallèle par paquets de 64 à chacune des 4 têtes d'attention de la première couche. Le réseau a été entraîné par descente de gradient stochastique avec des minibatch de 2 séquences sur un GPU du service de Google colab.

Lorsqu'une séquence de taille L est présentée au modèle, chacune des 24 têtes d'attention calcule une matrice d'attention de taille $L \times L$ rendant chacune compte d'un aspect de l'attention mutuelle des éléments de la séquence. L'analyse du mécanisme d'attention passe donc par une observation conjointe de l'ensemble de ces matrices. Une matrice d'attention fait apparaitre la séquence d'éléments fourni en entrée sur l'axe des abscisses et sur l'axe des ordonnées. L'élément d'abscisse x et d'ordonnée y indique l'attention que l'élément y porte à l'élément x, qu'il soit dans son passé ou dans son futur. Par conséquent, le triangle inférieur gauche (resp. supérieur droit) correspond à des attentions portées vers le passé (resp. vers le futur).

4.2.1. Gamme chromatique

Afin d'observer le comportement des têtes d'attention dans un cadre simple, une séquence de valeurs correspondant à une progression chromatique sur plusieurs octaves consécutives est fournie en entrée du modèle entrainé. Afin de faciliter la visualisation de l'influence mutuelle des différentes hauteurs de la gamme, seuls les éléments correspondant à des messages MIDI Note On sont conservés dans la séquence d'entrée.

La figure 2 illustre l'attention mutuelle entre ces éléments (du bleu foncé pour une faible attention, au jaune clair pour une forte attention), calculée par une des quatre têtes d'attention de la troisième couche. Cette figure fait apparaitre d'importantes valeurs d'attention entre les notes ayant une relation d'octave. Par exemple, le $do2$ porte une attention élevée aux notes $do1$ et $do3$. Cette observation indique la capacité du modèle à apprendre l'importance de la relation d'octave en musique. En effet, la représentation utilisée en entrée du modèle attribue un entier à chaque hauteur sans apporter aucune connaissance musicale supplémentaire à priori.

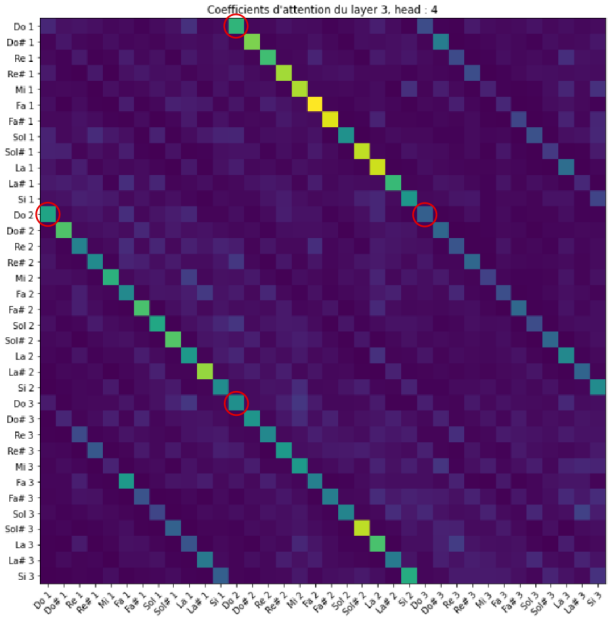

Figure 2. Attentions calculées par la tête 4 de la couche 3 pour une séquence chromatique ascendante. Les axes des abscisses et des ordonnées font apparaitre les messages MIDI Note On de 36 demi-tons constituant 3 octaves contiguës (de do 1 à si 3). Cette tête d'attention fait apparaître une attention importante entre les hauteurs ayant une relation d'octave.

La figure 3 illustre la matrice d'attention produite par une tête d'attention de la quatrième couche. Cette matrice fait apparaitre une importante attention entre une note et les 2 ou 3 notes qui la précèdent. La séquence consistant en une gamme chromatique, les notes situées les unes après les autres s'avèrent aussi être des notes séparées par de petits intervalles (de un à quatre demi-tons). La séquence d'entrée utilisée dans cette expérience ne permet toutefois pas de distinguer si l'attention capturée par cette tête résulte d'une proximité dans le temps et/ou dans les hauteurs des notes.

La figure 4 illustre les matrices d'attentions produites par les 4 têtes d'attention de la sixième couche du transformeur. Il est connu dans le domaine de l'apprentissage profond que les couches les plus élevées ont tendance à capturer les notions les plus abstraites. Il n'est donc pas surprenant d'observer des valeurs d'attention réparties de manière plus uniforme dans cette couche. Malgré cette montée en abstraction, des comportements distincts semblent se dessiner pour chaque tête d'attention. Dans la tête 0, les notes portent leur attention sur les notes qui les précèdent. Au contraire, la tête 3 porte l'attention des notes sur les notes qui lui succèdent. On remarque dans cette dernière matrice une attention forte portée par les dernières notes sur les premières notes de la séquence. La complexité du réseau de neurones rend naturellement difficile d'interpréter ce phénomène.

Cette première expérience illustre la faculté des transformeurs à assigner des rôles distincts aux différentes têtes

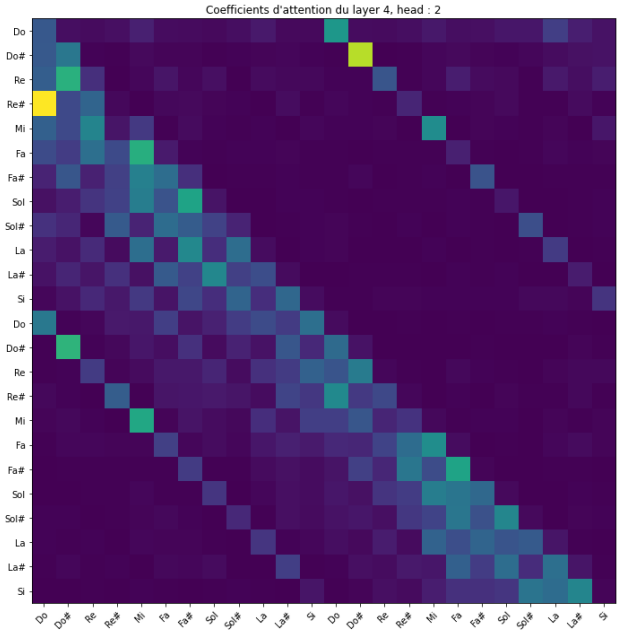

Figure 3. Attentions calculées par la tête 2 de la couche 4 pour une séquence chromatique ascendante sur deux octaves contiguës. Cette tête d'attention fait apparaitre une attention importante entre notes consécutives.

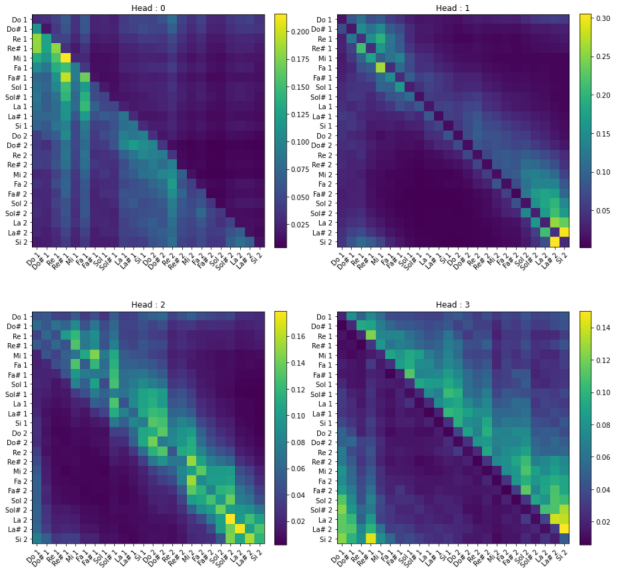

Figure 4. Attentions calculées par les têtes de la couche 6 pour une séquence chromatique ascendante sur deux octaves contiguës. Les têtes d'attention des couches élevées font apparaitre des attention réparties de manière plus uniforme.

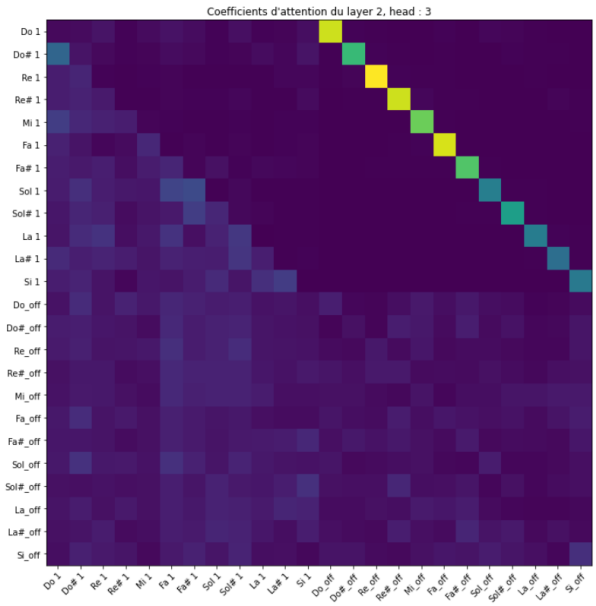

Figure 5. Attentions calculées par la tête 3 de la couche 2 pour une séquence chromatique ascendante constituée de 12 messages Note On et 12 messages Note Off. La matrice d'attention fait apparaitre une attention importante entre les messages Note Off et les messages Note On de la même hauteur.

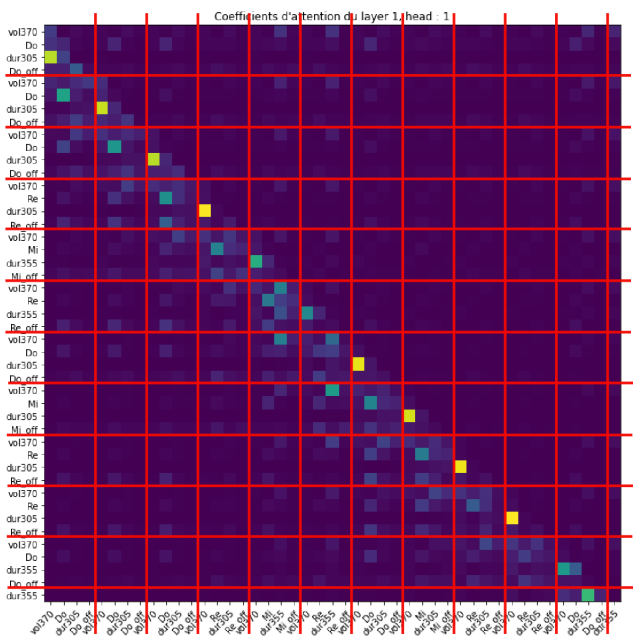

Figure 6. Attentions calculées par une tête de la première couche pour les 11 premières notes de *Au Clair de la Lune*. 4 messages MIDI sont associés à chaque note : Volume, Note On, Time shift (reflétant la durée de la note) et Note Off. Cette tête d'attention fait apparaitre un lien fort entre les informations de volume et de durée d'une même note.

d'attention. Ces têtes d'attention semblent par ailleurs traduire des types de relations entre notes qui ont un sens un musique. Dans les deux exemples ci-dessus, la relation d'octave et la relation de proximité (dans le temps ou dans les hauteurs) sont traduites dans deux têtes d'attention distinctes.

4.2.2. Relations Note On et Note Off

On concatène à la séquence précédente les messages Note Off qui indiquent originalement dans le fichier MIDI qu'une hauteur particulière cesse d'être jouée. La figure 5 montre une tête d'attention de la troisième couche à l'intérieur de laquelle chaque message Note On porte son attention sur l'élément Note Off futur correspondant à sa hauteur. Par exemple, le message Note On sur la hauteur Do porte son attention sur le message Note Off de cette même hauteur, qui survient 12 éléments plus tard dans la séquence. Cette constatation montre la capacité des transformeurs à assimiler non seulement les relations musicales qui lient les notes, mais aussi les artefacts résultants de nos outils de représentations informatiques de la musique.

4.2.3. Durées et volumes

La figure 6 illustre l'attention mutuelle calculée par une tête de la première couche, à partir des onze premières notes de la mélodie *Au Clair de la Lune*. Aux messages Note On et Note Off ont été ajoutés les messages de durée et de volumes des notes. Les éléments de la séquence sont ainsi groupés par quatre. Cette matrice fait apparaitre un lien fort entre les messages de volumes [6] et de durées. Ce phénomène peut s'expliquer par le fait que les notes longues dans le répertoire classique ont en moyenne tendance à être jouée avec un volume plus élevé que les notes courtes. Ce type d'hypothèses peut se vérifier de manière calculatoire sur le corpus, ce qui permettrait de valider la capacité du transformeur à capturer ce phénomène. Cela fait partie des perspectives de ce travail.

4.2.4. Accords et notes étrangères

La table 1 illustre l'attention totale portée à chaque note de l'ensemble {Do4, Mi4, Fa#4, Sol4}. L'objectif de cette expérience est de visualiser l'effet en terme d'attention d'une note étrangère dans un accord parfait (ici la note Fa# dans l'accord de Do majeur). Afin de lever toute ambiguïté sur l'influence de la position des notes dans la séquence de valeurs passée en entrée, l'expérience est effectuée en ordonnant de manière différente les notes de l'ensemble. L'attention plus faible portée à la note Fa#, et ce quel que soit l'ordre, traduit le fait que cette note est moins associée aux autres. La sur-représentation des accords parfaits dans ce corpus de musique de style majoritairement classique

6 . Les informations de volume constituent une composante essentielles du dataset MAESTRO étant donné que les pièces qui le constituent ont été enregistrées lors de performance pianistiques professionnelles. Ce dataset se distingue ainsi de ceux constitués à partir de partitions qui regroupent des notes dont le volume ne résulte que des annotations de la partition et fait donc apparaitre beaucoup moins de variété qu'à l'issue d'une performance réelle.

Attention reçue	Do4	Mi4	Fa#4	Sol4
Do4 - Mi4 - Fa#4 - Sol4	2.7	2.9	1.26	2.3
Mi4 - Sol4 - Do4 - Fa#4	3.89	2.97	1.17	1.69

Table 1. Attention cumulée reçue par chacune des notes constituant l'ensemble de notes {Do4, Mi4, Fa#4, Sol4}. Les notes sont successivement présentées au modèle suivant deux ordonnancements différents.

explique probablement cette attention mutuelle prédominante entre les notes de l'accord de do majeur dans cet exemple.

4.3. Bilan et perspectives

Ces expérience nous aident à visualiser la manière dont les transformeurs, initialement imaginés pour la modélisation du langage naturel, apprennent le concept d'attention mutuelle entre les éléments d'une séquence de notes. Ces expériences préliminaires ouvrent la voie à de nombreuses perspectives pour améliorer notre compréhension de ces algorithmes souvent comparés à des boites noires. La confirmation des tendances observées dans les matrices d'attentions par des tests de corrélation sur les données d'apprentissage constitue une des perspectives majeures de ce projet.

Les capacités d'attention d'un transformeur prennent la forme d'un ensemble complexe de matrices réparties en différentes couches. Ces expériences nous offrent des intuitions sur le potentiel de ces matrices, mais leur étude systématique requiert l'élaboration d'approches formelles qui sont centrales dans les perspectives de ce projet. Nous nous pencherons en particulier sur la notion de *vecteur d'attention* [27] qui consiste à résumer l'attention mutuelle entre deux éléments arbitraires d'une séquence à travers un vecteur prenant en compte l'ensemble des têtes d'attention. Les vecteurs d'attentions permettent d'évaluer de manière commode la capacité d'un transformeur à identifier des relations entres des couples d'éléments spécifiés à priori. Une perspective de ce projet consiste à évaluer cette approche pour l'identification d'éléments déterminants dans le déroulement d'une cadence, en particulier le point de préparation et le point de cadence [28].

5. CONCLUSIONS ET PERSPECTIVES

À travers les plongements de mots et le mécanisme d'attention, ce projet de recherche exploratoire vise à améliorer notre compréhension du potentiel des techniques de TALN pour la modélisation de données musicales. Ces expériences ouvrent de nombreuses questions sur la pertinence de l'analogie entre le langage naturel et le langage musical. La question du *sens* d'une phrase musicale apparait à la fois commune limite et un défi majeur pour l'étude de cette analogie. Des concepts abstraits propres au langage naturel, tels que l'humour et l'ironie nous interrogent particulièrement lorsqu'on tente de les interpréter dans le domaine musical. Les représentations internes d'un modèle mettant en œuvre le mécanisme d'attention ouvrent la voie à des approches originales pour la classification stylistique musicale et pour notre compréhension d'éléments du langage tonal tels que les cadences et les notes modulantes. Ces techniques du TALN font actuellement l'objet de nombreuses recherches et améliorations. Leur utilisation en musique est amenée à renouveler ces questions dans les années à venir. Les outils élaborés en TALN visent à manipuler des séquences d'éléments. La musique est cependant généralement structurée de manière plus complexe. L'adaptation de ces outils à des représentation *multi-sequences* s'annonce prometteuse pour la modélisation de musique.

6. REFERENCES

[1] W. Piston, M. DeVoto, and A. Jannery, *Harmony*. WW Norton New York, 1978.

[2] G. Lample, S. Subramanian, E. Smith, L. Denoyer, M. Ranzato, and Y.-L. Boureau, "Multiple-attribute text rewriting," in *International Conference on Learning Representations*, 2018.

[3] S. Dai, Z. Zhang, and G. G. Xia, "Music style transfer : A position paper," *arXiv preprint arXiv :1803.06841*, 2018.

[4] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "Midi-vae : Modeling dynamics and instrumentation of music with applications to style transfer," *arXiv preprint arXiv :1809.07600*, 2018.

[5] J. Sakellariou, F. Tria, V. Loreto, and F. Pachet, "Maximum entropy models capture melodic styles," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[6] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," *arXiv preprint arXiv :1809.04281*, 2018.

[7] M. Mongeau and D. Sankoff, "Comparison of musical sequences," *Computers and the Humanities*, vol. 24, no. 3, pp. 161–175, 1990.

[8] B. Cornelissen, W. Zuidema, and J. A. Burgoyne, "Mode classification and natural units in plainchant," in *Proceedings of the 21th International Conference on Music Information Retrieval (ISMIR 2020). Montréal, Canada*, 2020.

[9] T. Nuttall, M. García Casado, V. Núñez Tarifa, R. Caro Repetto, and X. Serra, "Contributing to new musicological theories with computational methods : the case of centonization in arab-andalusian music," in *20th Conference of the International Society for Music Information Retrieval (ISMIR 2019) : 2019 Nov 4-8 ; Delft, The Netherlands.[Canada] : ISMIR ; 2019. p. 223-8.* International Society for Music Information Retrieval (ISMIR), 2019.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv :1310.4546*, 2013.

[11] S. Madjiheurem, L. Qu, and C. Walder, "Chord2vec : Learning musical chord embeddings," in *Proceedings of the constructive machine learning workshop at 30th conference on neural information processing systems (NIPS2016), Barcelona, Spain*, 2016.

[12] D. Herremans and C.-H. Chuan, "Modeling musical context with word2vec," *arXiv preprint arXiv :1706.09088*, 2017.

[13] C.-H. Chuan, K. Agres, and D. Herremans, "From context to concept : exploring semantic relationships in music with word2vec," *Neural Computing and Applications*, vol. 32, no. 4, pp. 1023–1036, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv :1706.03762*, 2017.

[15] T.-P. Chen and L. Su, "Attend to chords : Improving harmonic analysis of symbolic music using transformer-based models," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, 2021.

[16] Y.-S. Huang and Y.-H. Yang, "Pop music transformer : Generating music with rhythm and harmony," *arXiv preprint arXiv :2002.00212*, 2020.

[17] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the maestro dataset," *arXiv preprint arXiv :1810.12247*, 2018.

[18] Y.-H. Chen, Y.-H. Huang, W.-Y. Hsiao, and Y.-H. Yang, "Automatic composition of guitar tabs by transformers and groove modeling," *arXiv preprint arXiv :2008.01431*, 2020.

[19] A. Huang, M. Dinculescu, A. Vaswani, and D. Eck, "Visualizing music self-attention," 2018.

[20] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove : Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available : http://www.aclweb.org/anthology/D14-1162

[22] P. D. Turney and P. Pantel, "From frequency to meaning : Vector space models of semantics," *J. Artif. Int. Res.*, vol. 37, no. 1, p. 141–188, Jan. 2010.

[23] O. Levy and Y. Goldberg, "Linguistic regularities in sparse and explicit word representations," in *Proceedings of the eighteenth conference on computational natural language learning*, 2014, pp. 171–180.

[24] C. W. White and I. Quinn, "The yale-classical archives corpus," *Empirical Musicology Review*, vol. 11, no. 1, 2016.

[25] G. Hadjeres, F. Pachet, and F. Nielsen, "Deepbach : a steerable model for bach chorales generation," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1362–1371.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert : Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv :1810.04805*, 2018.

[27] A. Coenen, E. Reif, A. Yuan, B. Kim, A. Pearce, F. Viégas, and M. Wattenberg, "Visualizing and measuring the geometry of bert," *arXiv preprint arXiv :1906.02715*, 2019.

[28] L. Bigo, L. Feisthauer, M. Giraud, and F. Levé, "Relevance of musical features for cadence detection," in *International Society for Music Information Retrieval Conference (ISMIR 2018)*, 2018.