EMPREINTES : UN ALGORITHME POUR L'EXTRACTION AUTOMATIQUE DE DESCRIPTEURS AUDIO ET L'ÉVALUATION DE LEUR PERTINENCE POUR L'ANALYSE MUSICALE

Jean-Henri Nothias
STMS Lab - IRCAM, Sorbonne Université,
CNRS (UMR9912)
nothias@ircam.fr

Florian Iochem
Université de Strasbourg,
ACCRA (UR3402)
florian.iochem@etu.unistra.fr

RÉSUMÉ

Empreintes est un algorithme d'Analyse Musicale Assistée par Ordinateur (AMAO) *open source* [1] dont l'objectif est de recommander à l'analyste une collection de descripteurs audio pertinents pour décrire un enregistrement sonore. À partir d'une segmentation temporelle par fenêtrage glissant, le modèle mesure la similarité cosinus de spectrogrammes Mel puis construit une matrice de dissimilarité. Il filtre ensuite cette matrice à l'aide d'un graphe de franchissement, afin d'obtenir une métrique sur l'espace des segments. Une quarantaine de descripteurs audio, pour la plupart issus de la librairie FluCoMa [17], sont extraits et évalués par un score de localisation quantifiant leur cohérence avec la structure métrique dégagée. L'algorithme génère des « topographies timbrales » tridimensionnelles interprétables par l'analyste. Nous l'appliquons ici à trois interprétations de la *Danse Sacrale* du *Sacre du Printemps* (1913) d'Igor Stravinsky (1882-1971) afin d'évaluer la pertinence et la stabilité des descripteurs recommandés.

1. INTRODUCTION

En tant que domaine de recherche en constante expansion et présentant des méthodes infiniment malléables, l'AMAO (Analyse Musicale Assistée par Ordinateur) offre aujourd'hui des possibilités analytiques dans deux directions différentes mais complémentaires : d'une part, pour l'étude de données symboliques, dans la continuité des pratiques « historiques » de la musicologie systématique et, d'autre part, pour l'étude d'informations dites sous-symboliques, cherchant à modéliser les attributs perceptifs de la musique. Cette deuxième approche est fondée sur l'extraction de données à partir d'un signal audio via l'application de fonctions mathématiques – appelés descripteurs audio – réduisant en quelque sorte l'information contenue dans un fichier audio [10]. Malt et Jourdan définissent par ailleurs ces outils comme des « paramètre(s) unis ou multidimensionnels, caractérisant un aspect du signal sonore, ramenant une dimension particulière de ce signal à un (ou plusieurs) paramètre(s) numérique(s) »[11]. Leur implémentation dans des *toolbox* telles que FluCoMa [17] ou Timbre Toolbox [13] permet déjà l'automatisation de l'extraction d'un certain nombre d'entre eux, facilitant, par exemple, l'analyse de larges corpus d'œuvres musicales.

C'est dans ce contexte que s'inscrit *Empreintes*, un algorithme d'AMAO conçu pour automatiser la recommandation de descripteurs audio pertinents à partir d'un enregistrement donné. Son principe repose sur la construction d'une métrique sur l'espace des segments audio, à partir de laquelle un score de localisation évalue la cohérence de chaque descripteur audio avec la structure de l'enregistrement. L'algorithme produit également des visualisations tridimensionnelles — que nous appelons ici des « topographies timbrales » — permettant à l'analyste d'explorer visuellement ces structures. Dans cet article, nous présentons en détail l'architecture d'*Empreintes* (section 2), avant de l'appliquer à trois interprétations de la *Danse Sacrale* du *Sacre du Printemps* (1913) d'Igor Stravinsky (1882-1971) (section 3). Nous discutons enfin des performances actuelles de l'algorithme, de ses limites et de ses perspectives de développement (section 4).

Si l'utilisation des descripteurs audio est de plus en plus commune, le statut conceptuel de cet outil reste problématique et révélateur d'un manque de standardisation. La littérature fait en effet état de recouvrements terminologiques entre concepts distincts ou, à l'inverse, de dénominations divergentes pour des mesures analogues [12]. Cette hétérogénéité conceptuelle engendre, selon Peeters, plusieurs difficultés majeures : elle complique la comparaison des études, leur reproductibilité et freine par là même l'émergence de standards méthodologiques partagés [13]. Par ailleurs, le musicologue, encore rarement formé à l'utilisation de tels outils, n'est pas toujours en mesure de déterminer avec assurance quels descripteurs seraient les mieux adaptés à son projet analytique.

Résoudre ces problèmes d'ordre sémantique reste aujourd'hui largement hors de portée. Néanmoins, la recommandation automatique d'axes d'analyse permise par *Empreintes*, couplée au calcul de scores de localisation, conçus comme des indicateurs quantitatifs de la pertinence des descripteurs dans la caractérisation de l'enregistrement donné, nous semble constituer une approche

1 . Le code d'*Empreintes* est disponible ici : https://github.com/JHNothias/Empreintes.

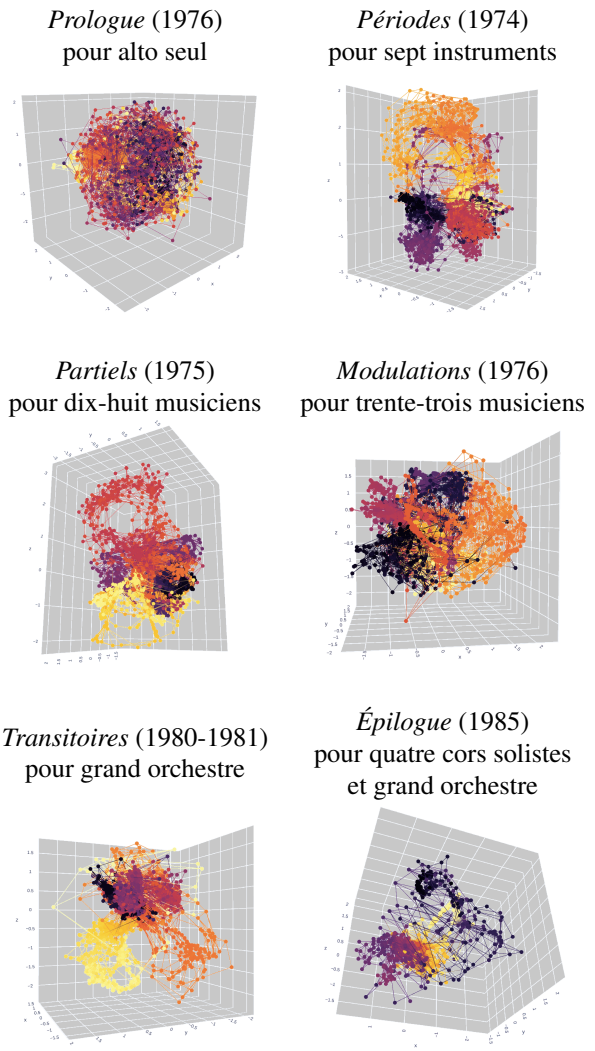

Figure 1 – Topographies timbrales des *Espaces Acoustiques* (1976-1985) de Gérard Grisey (1946-1998).

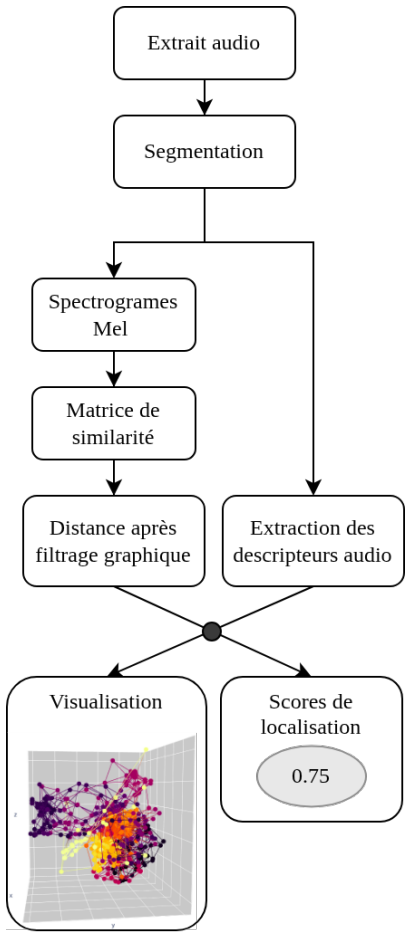

Figure 2 – Diagramme explicatif du worfklow d'*Empreintes*

originale pour réduire cette fragmentation conceptuelle. D'autre part, si les représentations graphiques issues de méthodes d'AMAO doivent être considérées comme des outils au service d'une démarche analytique — et non comme sa finalité —, l'obtention de topographies timbrales dans des espaces tridimensionnels manipulables et paramétrables, en fonction des descripteurs audio jugés les plus pertinents par l'algorithme, permet à l'analyste de penser de potentiels croisements entre les différentes dimensions sémantiques des descripteurs audio et ses propres hypothèses d'écoute. Notre algorithme est implémenté en Python, avec une optimisation GPU [2] via Taichi [7].

La Figure 1 montre les topographies timbrales obtenues pour chacune des pièces des *Espaces Acoustiques* (1976-1985) de Gérard Grisey (1946-1998). Ces représentations tridimensionnelles sont manipulables dans l'espace par l'analyste. Elles sont colorables, pour représenter l'évolution de n'importe quel descripteur audio extrait par l'algorithme, afin de visualiser la manière dont ils caractérisent la topographie. Ici, l'évolution des niveaux de couleur ne représente que l'évolution temporelle des pièces (couleur sombre = début, couleur claire = fin). Chaque point de ces nuages correspond à un segment issu de la segmentation des pièces : la comparaison de leurs positions dans l'espace permet d'évaluer leur similarité sur le plan timbral, notamment. En d'autres termes, plus deux points sont proches l'un de l'autre dans l'espace, plus les spectrogrammes Mel des segments qu'ils représentent sont similaires.

2. MÉTHODE

L'architecture d'*Empreintes* se décompose en sept étapes – parfois menées en parallèle, comme on peut le voir sur la figure 2 – que nous décrivons en détail dans cette section :

(1) Segmentation temporelle de l'enregistrement sonore.

(2) Conversion des segments en spectrogrammes Mel.

2 . Le GPU (*Graphics Processing Unit* ou « processeur graphique » en français) est une unité de calcul permettant de faire des opérations massivement parallèles. Nous l'utilisons ici afin de mener à bien l'extraction du descripteur audio de la rugosité ainsi que pour l'évaluation des scores de localisation, dont le calcul est distribué sur tous les points de l'espace des segments.

(3) Construction d'une matrice de dissimilarité initiale.

(4) Filtrage métrique par graphe.

(5) Calcul des distances graphiques.

(6) Extraction des descripteurs audio.

(7) Calcul des scores de localisation.

2.1. Segmentation

La segmentation des fichiers audio se fait actuellement par fenêtrage glissant : le fichier est découpé en segments d'une durée de 2 secondes chacun, avec un pas d'1 seconde (soit 50% de recouvrement) ou de 0.25 secondes (75% de recouvrement), selon les limites de temps de calcul. Ce choix permet une manipulation plus systématique des données que ce qu'une segmentation basée sur les transitoires d'attaque, sur la détection de tempo ou encore sur la suppression des silences permettrait. Ainsi, les segments obtenus sont de tailles égales, ce qui facilite leur comparaison, et cette segmentation reste indépendante de la nature du fichier audio. Ce choix permet aussi de préserver la continuité de l'analyse en cas de changements soudains des paramètres musicaux.

Chaque segment est ensuite converti en spectrogramme Mel (512 bandes, en amplitude logarithmique ($S_{norm} = \log(1+S)$, pour un spectrogramme S donné). Si le fichier analysé est au format stéréo, les spectrogrammes des deux pistes sont moyennés. Le spectrogramme est calculé en amont sur le fichier audio complet avant d'être segmenté.

2.2. Construction d'une matrice de dissimilarité

La matrice de dissimilarité est un outil très utilisé en AMAO (introduite dans ce contexte par Foote [5]), permettant d'après Couprie de « mettre en évidence les structures sous-jacentes aux données » [3]. Cette matrice se construit sur la base de mesures de dissimilarité entre segments, extraits par exemple d'un fichier audio, et permet notamment la détection de structures répétées [6]. Notre algorithme étend les capacités de cet outil en proposant, pour chaque descripteur audio, une visualisation de sa capacité à rendre compte de l'autosimilarité structurelle de l'enregistrement analysé.

Dans notre algorithme, la dissimilarité entre deux segments i et j est calculée comme une dissimilarité par concordance entre leurs spectrogrammes. Soit $S_i, S_j \in \mathbb{R}^{n_{\text{mels}} \times T}$ les spectrogrammes des segments i et j, où T est leur nombre de trames temporelles et n_{mel} le nombre de bandes mel utilisées. On calcule la similarité cosinus segment-par-segment :

$$\text{sim}(i,j) = \frac{\langle S_i, S_j \rangle}{||S_i||\,||S_j||} \tag{1}$$

où $||S_k|| := \sqrt{\sum_{t=1}^{T} \sum_{m=1}^{n_{\text{mels}}} S_k[t,m]^2}$ est la norme euclidienne du spectrogramme S_k et $\langle S_i, S_j \rangle$ le produit scalaire des spectrogrammes. La dissimilarité par concordance est alors définie comme :

$$d_c(i,j) = 1 - \text{sim}(i,j) \tag{2}$$

Deux segments dont les spectrogrammes sont identiques auront une dissimilarité de 0, tandis que les segments orthogonaux auront une dissimilarité de 1.

2.3. Filtrage graphique de la matrice de dissimilarité

L'information contenue dans l'extrait audio est réduite à un espace métrique. Ce dernier correspond à la donnée (M,d) d'un ensemble M de points et d'une application $d : M^2 \to \mathbb{R}_+$, telle que pour tous points $p,q,r \in M$:

— séparation : $d(p,q) = 0 \iff p = q$,

— symétrie : $d(p,q) = d(q,p)$,

— inégalité triangulaire : $d(p,r) \leqslant d(p,q) + d(q,r)$.

La matrice de dissimilarité d_c satisfait les propriétés de symétrie, de positivité et de séparation, mais pas nécessairement l'inégalité triangulaire. Ce n'est donc pas une métrique au sens strict. Qui plus est, deux phénomènes peuvent nuire à la lisibilité des données :

— Une structure globale trop marquée peut écraser des structures locales plus fines, les rendant difficilement exploitables par l'analyste.

— Un échantillonnage trop grossier peut introduire des structures topologiques non représentatives des données sous-jacentes, ce que nous appellerons ici le « bruit topologique ».

Pour y remédier, on construit un graphe où chaque point de l'espace est relié à d'autres points par des arêtes pondérées par leur dissimilarité. La distance calculée sur ce graphe constitue alors une métrique à proprement parler, et ce filtrage permet de sélectionner les caractéristiques de la métrique à préserver ou à atténuer, limitant ainsi l'impact des deux phénomènes cités plus haut. Nous avons choisi d'utiliser ce que l'on a appelé un « graphe de franchissement », l'union entre un graphe de k-plus proches voisins (ici, $k = 4$) et un graphe G constitué d'arêtes (p,q) décrites par la relation suivante :

$$(p,q) \in G \iff \forall r \in V_q,\ d(p,q) \leqslant d(p,r) \tag{3}$$

où V_q est un voisinage ici constitué des k plus proches voisins de q, et $d = d_c$. Nous appelons q un « point de franchissement » relatif à p, car pour une distance à p croissante, q est atteint par p avant tout voisin de q. Nous ne conservons, pour chaque point p, que les 4 points de franchissement q dont les dissimilarités $d_c(p,q)$ sont les plus faibles. Intuitivement, les points de franchissement permettent de passer d'une région à l'autre de l'espace en effectuant le plus petit pas possible. La figure 3 donne un exemple d'un tel graphe. Sur celle-ci, p et q sont des points de franchissement mutuels tandis que q est un point de franchissement relatif à p', et non l'inverse.

Cette construction particulière réduit les phénomènes cités plus haut, et permet d'obtenir des visualisations plus

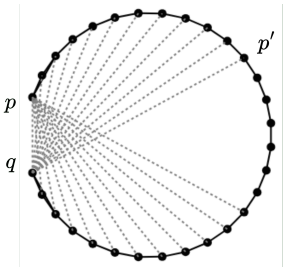

Figure 3 – Un graphe de franchissement sur un échantillonnage d'un arc de cercle. Les deux plus proches points de franchissement sont reliés par des arêtes grises et les deux plus proches voisins par des arêtes noires.

claires des données. Nous appelons d_f la distance obtenue après filtrage de d_c par le graphe de franchissement.

2.4. L'extraction des descripteurs audio

Empreintes extrait automatiquement une collection de descripteurs audio via la bibliothèque FluCoMa [17]. Cette dernière propose différentes catégories de descripteurs audio, présentés dans la liste ci-dessous :

- *Descripteurs spectraux* : centroïde, étalement, *skew*, kurtosis, *rolloff*, platitude, facteur de crête.
- *Descripteurs de hauteur* : hauteur estimée, confiance statistique de l'estimation via la méthode YIN [4].
- *Descripteurs de variations* : nouveauté spectrale.
- *Descripteurs psychoacoustiques* : rugosité (calculée sans FluCoMa).
- *Descripteurs d'intensité* : sonie ou « intensité sonore » et pic.

Nous avons décliné chacun de ces descripteurs en trois « versions » différentes :

1. Le descripteur tel quel, ou sa moyenne sur la durée d'un segment.
2. La variance de ce descripteur sur la durée d'un segment.
3. La moyenne des différences du descripteur entre trames consécutives sur un segment, utilisée comme approximation de sa dérivée temporelle.

En prenant en compte ces déclinaisons et le descripteur temps, notre outil permet actuellement d'évaluer 40 descripteurs différents.

2.5. Scores de localisation

L'objectif du score de localisation est d'évaluer la pertinence d'un descripteur audio pour caractériser un enregistrement donné. Plus il est haut, plus le descripteur varie de manière cohérente avec la structure du fichier audio, et inversement. Par exemple, la constance de ce descripteur dans les voisinages (passages similaires) et sa variabilité entre régions distantes sont prises en compte.

Pour un descripteur h et un ensemble M de segments de distance métrique d_f, le score de localisation $L_\alpha(h)$ est déterminé comme suit :

$$L_\alpha(h) \coloneqq \frac{1}{|M|} \sum_{p \in M} \left| \sum_{q \in M} (w_p(q) - \overline{w_p}(q)) \cdot |h(p) - h(q)| \right|$$

Soit Δh_p le vecteur $[h(p) - h(q)]_{q \in M}$ et Δw_p le vecteur $[w_p(q) - \overline{w_p}(q)]_{q \in M}$. Admettant une valeur absolue s'appliquant composante par composante, on peut réécrire la formule précédente comme :

$$L_\alpha(h) \coloneqq \frac{1}{|M|} \sum_{p \in M} |\langle \Delta w_p, |\Delta h_p| \rangle| \tag{4}$$

où

$$w_p(q) \coloneqq \frac{2}{Z_p \alpha} e^{-\frac{1}{2}\left(\frac{d_f(p,q)}{\sigma_p \alpha}\right)^2} \tag{5}$$

et $\overline{w_p}$ est la distribution complémentaire à w_p, définie par :

$$\overline{w_p}(q) = \frac{w_p(p) - w_p(q)}{\sum_{s \in M} w_p(p) - w_p(s)} \tag{6}$$

Ici :

— $\sigma_p^2 = \frac{1}{|M|} \sum_{q \in M} d_f(p,q)^2$ est la variance des distances à p.

— α est un paramètre contrôlant l'étalement de la gaussienne.

— Z_p est une constante de normalisation telle que $\sum_{q \in M} w_p(q) = 1$.

L'idée de cette formule est la suivante : Δw_p attribue des poids positifs aux points proches de p et des poids négatifs aux points distants, avec un poids nul pour les points à mi-chemin. $|\Delta h_p(q)|$ est élevé lorsque h varie fortement entre p et q, et faible dans le cas contraire. Le produit scalaire $\langle \Delta w_p, |\Delta h_p| \rangle$ mesure donc dans quelle mesure les variations de h sont corrélées à la distance à p. Dans le cas d'un descripteur bruité, dont les variations sont indépendantes de la distance, les contributions positives (points proches) et négatives (points distants) se compensent dans le produit scalaire, donnant un score faible. Si au contraire h est similaire entre points proches et dissimilaire entre points distants, les contributions distantes dominent ; leurs poids étant négatifs, la valeur absolue les ramène dans le positif, produisant un score élevé. Le score est donc élevé précisément quand h varie de façon cohérente avec la structure métrique.

Si le descripteur h est borné entre 0 et 1, $L_\alpha(h)$ le sera aussi, ce qui permet aux scores finaux d'être indiqués en pourcentages. Pour capturer des structures à différentes échelles, on calcule le score avec plusieurs valeurs d'étalement (dans notre cas, $\alpha^{-1} \in [\![1, 5]\!]$) et on retient le maximum.

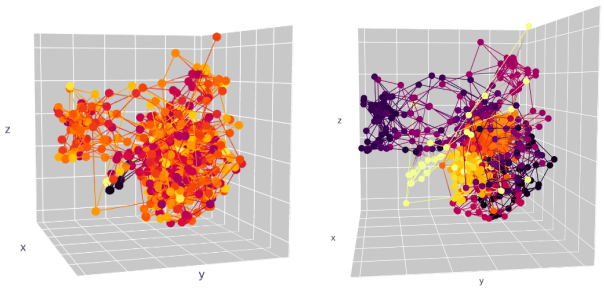

Figure 4 – Un exemple de topographie timbrale, extraite d'un enregistrement de la *Danse sacrale* du *Sacre du printemps* (1913) d'Igor Stravinsky (enregistrement « Wiener » (1975)), coloriée par un descripteur avec un score de localisation faible (gauche) et par un descripteur avec un score élevé (droite).

2.6. Visualisations sous forme de « topographies timbrales »

Notre algorithme génère des « topographies timbrales » (voir Figure 4), permettant de visualiser la métrique d_f via le *multidimentional scaling* [2], plongeant l'ensemble des segments en 2 ou 3 dimensions. Nous générons des représentations en 3 dimensions interactives via plotly [8], représentant un réseau dont les sommets sont les segments audio et dont les arêtes correspondent au graphe de franchissement. Cette visualisation peut être coloriée par chaque descripteur. Passer la souris sur un segment permet de connaître sa position temporelle dans l'enregistrement ainsi que la valeur du descripteur et l'indice du segment. Les axes présents sur les figures pourraient être supprimés, mais permettent de mieux se rendre compte de la profondeur.

3. APPLICATION ANALYTIQUE

Afin d'évaluer notre algorithme, nous l'avons appliqué à trois enregistrements de la *Danse Sacrale*, extraite du *Sacre du Printemps* (1913) d'Igor Stravinsky. Ce choix s'explique notamment par la richesse timbrale de l'œuvre ainsi que par la complexité de son écriture rythmique. La diversité de ces interprétations nous a conduits à comparer les scores de localisation obtenus pour chacune d'entre elles, dans le but d'évaluer la fiabilité de l'algorithme quant à l'identification des descripteurs les plus pertinents pour caractériser la pièce, tout en examinant la corrélation entre les scores obtenus [3] .

Les trois enregistrements analysés, disponibles sur YouTube, sont les suivants :

- *Wiener* [15] : Enregistrement réalisé en 1975 par le Wiener Philharmoniker sous la direction de Lorin Maazel, publié par Decca.

- *Radio France* [16] : Enregistrement réalisé en 2010 par l'Orchestre Philharmonique de Radio France sous la direction de Myung-Whun Chung, publié par Decca à l'occasion du centième anniversaire de l'œuvre.

- *Chicago* [14] : Enregistrement réalisé en 1974 par le Chicago Symphony Orchestra sous la direction de Sir Georg Solti, publié par Decca.

Les scores de localisation obtenus pour chaque descripteur audio sont présentés dans la Table 1.

Descripteur	Wiener	Chicago	Radio Fr.
time	75.4	75.0	71.8
centroid	55.8	63.1	60.7
centroid-var	43.5	47.0	43.3
centroid-dif	28.6	29.6	28.6
spread	64.1	50.7	49.8
spread-var	46.4	38.6	42.9
spread-dif	23.8	23.7	27.0
skew	54.0	37.5	59.2
skew-var	50.7	57.3	55.8
skew-dif	25.1	26.2	31.7
kurtosis	49.8	58.8	58.5
kurtosis-var	48.8	57.9	55.8
kurtosis-dif	24.2	26.7	31.1
rolloff	56.8	56.7	56.2
rolloff-var	52.6	40.7	41.6
rolloff-dif	25.0	24.9	27.9
flatness	74.1	52.7	63.3
flatness-var	51.1	24.4	24.5
flatness-dif	23.9	24.5	27.2
crest	62.1	56.9	57.1
crest-var	46.7	47.6	43.8
crest-dif	23.9	27.6	28.1
pitch measured	54.4	52.7	53.5
pitch measured-var	31.9	32.2	27.4
pitch measured-dif	26.0	28.2	24.9
pitch confidence	59.1	65.3	65.4
pitch confidence-var	51.7	52.8	46.6
pitch confidence-dif	22.3	22.0	23.4
novelty	56.2	54.4	58.1
novelty-var	32.8	50.5	56.5
novelty-dif	17.2	16.8	19.8
roughness	64.2	62.1	62.8
roughness-var	52.2	46.6	53.1
roughness-dif	24.3	24.1	27.4
loudness	72.2	67.9	69.0
loudness-var	53.5	54.7	55.8
loudness-dif	33.6	35.5	38.4
true peak	72.2	67.5	69.2
true peak-var	24.4	39.8	42.4
true peak-dif	28.2	29.5	37.3

Table 1 – Scores de localisation obtenus par chaque descripteur pour chacun des trois enregistrements étudiés.

3 . Bien que la qualité d'enregistrement influe de manière non négligeable sur les résultats obtenus, puisque l'on se place à un niveau d'analyse sous-symbolique, nous avons choisi des interprétations où les artefacts d'enregistrement/mixage ne paraissaient pas rédhibitoires.

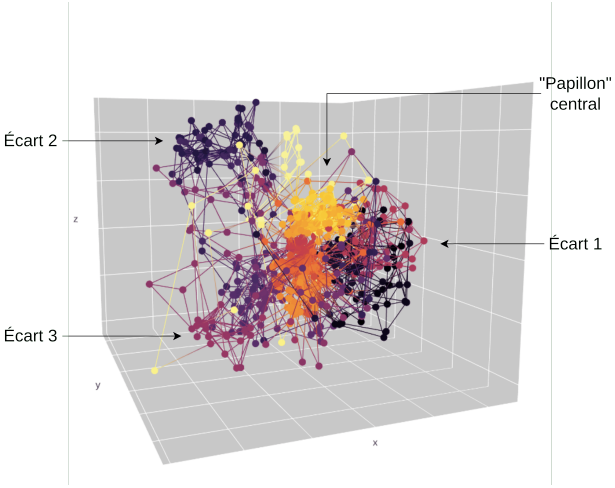

Figure 5 – Topographie timbrale de l'enregistrement « Wiener » avec identification des principales régions timbrales.

3.1. Analyse de l'enregistrement « *Wiener* »

Les scores de localisation obtenus pour l'enregistrement « Wiener » sont légèrement supérieurs, sur les descripteurs les mieux classés par l'algorithme, à ceux des deux autres enregistrements. Les descripteurs les mieux notés pour cet enregistrement sont le temps (75%) la platitude spectrale (74.1%), le pic (« true peak ») [4] et l'intensité sonore (72.2%), la rugosité (64.2%) [5], l'étalement spectral [6] (64.1%) et le facteur de crête [7] (62.1%). Les topographies timbrales correspondantes sont visibles sur la figure 6.

La topographie timbrale de l'enregistrement (figure 5) permet de distinguer quatre régions timbrales majeures :

- La région intitulée « écart 1 » englobe les 30 premières secondes du morceau ainsi que l'extrait de 1'58" à 2'24". Elle est caractérisée par une rugosité moyenne faible, une platitude et un étalement spectral faibles, une forte intensité sonore et un facteur de crête plutôt variable (on peut distinguer une sous région où ce dernier est élevé, et une autre où il est faible). Sur le plan musical, il s'agit d'un dialogue assez complexe, notamment au niveau rythmique, entre les timbales et les autres instruments de l'orchestre, souvent joués *staccato*.
- La région intitulée « écart 2 » englobe les extraits de 0'28" à 0'40", de 0'51" à 0'58" et de 1'32" à 1'40", des parties beaucoup moins dynamiques de la pièce où les discrets *pizzicati* des cordes côtoient les *staccato* des bois et des cuivres. Pour cause, cette zone de la topographie est caractérisée par une platitude et un étalement spectral importants (>70%), une rugosité et une intensité sonore faibles, et par un facteur de crête important.
- La région intitulée « écart 3 » représente le passage de 1'45" à 1'58". Elle est caractérisée par des descripteurs de platitude et d'étalement moyens (autour de 50%), une intensité sonore forte et une faible rugosité. Au niveau de la partition, cette partie est caractérisée notamment par des trilles virtuoses aux cordes et par un quintolet aux trompettes joué *fortissimo*, sur Si♭ et La♭.
- Le « papillon », région centrale de la topographie timbrale, se sépare en deux « ailes » : la première représentant la partie de 2'30" à 3'05" (en orange et la plus foncée des deux sur la figure 5) et la seconde celle de 3'35" à 4'30" (en jaune et plus claire sur la figure). Le papillon se distingue du reste du morceau par la rugosité (très forte à cet endroit) et par sa position temporelle, soit vers la fin de la pièce. Les deux ailes sont distinguées par l'intensité sonore et la rugosité, plus fortes dans la première, par le facteur de crête, bien plus important dans la deuxième, et par le temps. La deuxième aile est caractérisée par un dialogue entre des motifs percussifs et l'orchestre jouant principalement en *staccato*, avec des changements soudains d'intensité, en ligne avec la valeur du facteur de crête. La première aile correspond à une section où la dissonance prévaut dans les cuivres, jouant à forte intensité, sur le fond d'un roulement de timbales à intensité moyenne et d'un accompagnement par le reste de l'orchestre. C'est la partie centrale de la pièce, à la fois en termes de timbre sur la topographie et au niveau structurel puisqu'elle correspond au « refrain » de cette forme rondo, en reprenant le motif du début de la pièce.

3.2. Évaluation de la stabilité de l'algorithme : comparaison d'analyses de différentes interprétations de la *Danse Sacrale*

La matrice de corrélation des scores sur chaque paire d'enregistrements est donnée par la table 2.

Elle permet d'établir en un coup d'œil l'importante similarité entre les enregistrements « Chicago » et « Radio France ». Les descripteurs restent globalement très stables (écart-type maximal de 12%, la majorité étant en dessous des 5%). Cependant, l'ordre des descripteurs selon leur score change : les descripteurs d'étalement spectral, de platitude spectrale et de crête se positionnent bien plus haut pour l'enregistrement « Wiener » que pour les deux autres, au profit des descripteurs de centroïde spectral et de confiance statistique de l'estimation de hauteur, qui distinguent les écarts 1 et 3 de l'écart 2 et de la région en « papillon ». L'intensité sonore et le descripteur de pic demeurent d'excellents descripteurs sur les trois enregistrements, distinguant notamment la région « écart 2 » du reste de la pièce. L'examen des topographies timbrales révèle, pour les enregistrements « Chicago » et « Radio

4 . L'amplitude maximale du signal en dB sur une certaine fenêtre de temps.

5 . Descripteur psychoacoustique participant à notre perception de la dissonance [1].

6 . Étalement du spectre autour de son centroïde, représentant le poids relatif des fréquences aiguës et graves et fortement corrélé à l'impression de brillance du timbre.

7 . Ratio entre le pic et la moyenne quadratique du signal.

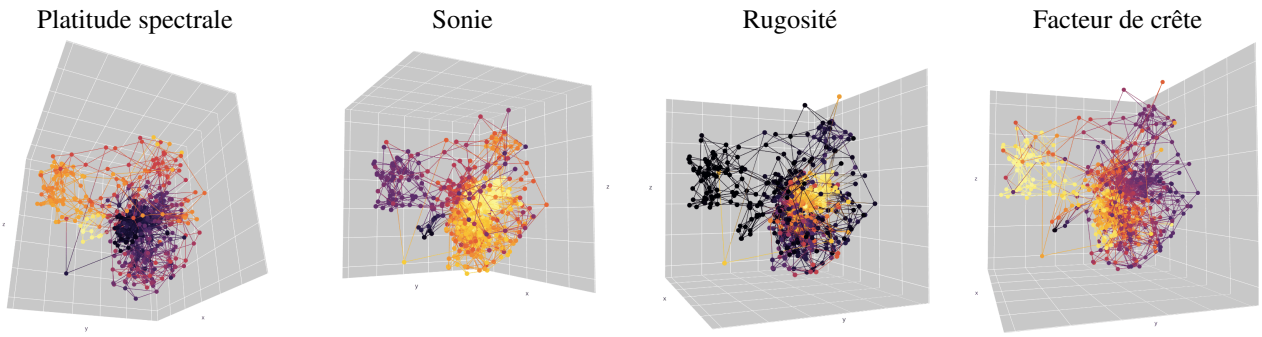

Figure 6 – Topographie timbrale de l'enregistrement « Wiener » coloriées avec différents descripteurs. On observe comment la rugosité identifie le « papillon » et comment le facteur de crête distingue une aile de l'autre.

	Wiener	Radio France	Chicago
Wiener	1	0.88	0.87
Radio France	0.88	1	0.95
Chicago	0.87	0.95	1

Table 2 – Matrice de corrélation des scores entre trois interprétations différentes de la *Danse Sacrale* du *Sacre du Printemps* (1913) d'Igor Stravinsky.

France », une séparation en régions analogues à celles décrites dans la section 3.1. La corrélation relativement forte observée dans cette matrice est, à notre sens, un indicateur de la pertinence et de la fiabilité de notre approche afin de déterminer quels descripteurs audio sont les plus à même de décrire la structure d'un enregistrement sonore donné [8] .

4. CONCLUSION ET PERSPECTIVES

4.1. Évaluation des performances actuelles d'*Empreintes* et développements ultérieurs envisagés

Les premiers résultats obtenus par l'application d'*Empreintes* à un corpus varié d'œuvres musicales sont encourageants. Au fur et à mesure des différentes analyses menées, nous détectons néanmoins quelques contraintes, d'ordre techniques et conceptuelles, auxquelles il conviendra de palier au fur et à mesure de l'avancée du projet.

Premièrement, l'analyse menée par *Empreintes* porte sur le signal audio dans sa globalité, sans distinction des sources sonores qui le composent. L'intégration d'un module de séparation de sources — reposant par exemple sur des architectures de réseaux neuronaux convolutifs (CNN) — ouvrirait des perspectives analytiques considérables : il deviendrait alors possible d'interroger des phénomènes tels que la fusion ou la différenciation de timbres au sein de flux sonores complexes, ou encore d'affiner la caractérisation de modes de jeu instrumentaux spécifiques.

Par ailleurs, la segmentation par fenêtrage glissant, susceptible d'introduire des artefacts de phase et particulièrement granulaire, n'est qu'une solution parmi bien d'autres. Si la solution actuelle assure une résolution temporelle satisfaisante, elle engendre néanmoins des volumes de données importants et, subséquemment, des temps de calcul qui peuvent s'avérer prohibitifs sur des corpus étendus. L'exploration de stratégies de segmentation adaptative constitue, à cet égard, une piste de recherche prioritaire. Nous prévoyons en effet d'ajouter plusieurs méthodes de segmentation alternatives (détection de tempo, des transitoires d'attaque, etc.) à l'algorithme que l'analyste pourrait expérimenter selon ses hypothèses d'écoute.

Par ailleurs, nous avons effectué un choix particulier de métrique pour comparer les segments audio, celle de la dissimilarité cosinus — qui permet de comparer précisément le contenu spectral entre les segments —, mais elle peut être adaptée selon les besoins analytiques [9] .

Enfin, sur un plan plus conceptuel, la formalisation de « descripteurs croisés », notion que nous proposons d'introduire et qui pourrait être définie comme une « sélection de descripteurs extraits d'un flux sonore dont les scores de localisation sont les plus élevés » – étape actuellement « manuelle » dans notre protocole analytique – reste à approfondir. Si elle ne peut palier totalement le problème de l'hétérogénéité conceptuelle des descripteurs audio soulevé en introduction, cette notion de « descripteurs croisés » pourrait néanmoins aider à tisser des liens entre hypothèses d'écoute et données objectivées.

4.2. Limites intrinsèques aux outils utilisés et aux concepts évoqués

Toute démarche analytique fondée sur les descripteurs audio se heurte à une limite fondamentale : ces descrip-

8 . Les descripteurs audio de la librairie FluCoMa ne sont évidemment pas tous significatifs sur le plan de l'analyse perceptive. C'est ici et surtout la pertinence structurelle de ces outils qui nous intéresse. Par ailleurs, n'importe quel descripteur audio d'une autre librairie pourrait tout à fait être intégré à notre algorithme.

9 . Par exemple, on peut utiliser une distance de Wasserstein si l'on s'intéresse à la forme globale des spectres des enregistrements étudiés ou bien encore la métrique induite de l'espace latent d'un VAE (*Variational Auto Encoder* [9]) afin de déterminer quels descripteurs audio expliquent la structure des résultats de l'apprentissage du modèle et de parvenir, ainsi, à une meilleure interprétabilité d'un algorithme d'IA donné.

teurs ne constituent, par définition, qu'une réduction du phénomène musical. Aussi exhaustive que soit l'extraction, elle ne saurait épuiser la complexité intrinsèque à l'œuvre. Il convient d'ailleurs de rappeler qu'un score de localisation faible n'invalide pas la pertinence analytique d'un descripteur : rappelons ici qu'*Empreintes* est avant tout conçu comme un outil de recommandation, dont la vocation est de faire émerger des pistes d'investigation et d'offrir une caractérisation générale de l'enregistrement, et non de se substituer au jugement de l'analyste. Par exemple, certains descripteurs peuvent tout à fait être corrélés à l'environnement acoustique de l'enregistrement ou encore au *mastering*, et représenteront alors une caractéristique constante de l'enregistrement, avec un faible score de localisation, mais pouvant conserver un intérêt analytique indéniable.

4.3. Fondements épistémologiques d'une automatisation de l'analyse musicale

L'AMAO proposant de saisir la complexité du phénomène musical par la manipulation et la modélisation d'informations musicales (qu'elles soient symboliques ou sous-symboliques), son application ne va pas sans soulever des questions quant au risque de réduire l'analyse musicale à ses seules dimensions formalisables. C'est précisément pour cette raison qu'*Empreintes* est conçu dans une perspective résolument interdisciplinaire, articulant informatique et musicologie, dans un dialogue où chaque discipline conserve ses exigences propres. Malgré la complexité inhérente au modèle, l'interprétabilité de chacune des étapes par lesquelles il passe doit rester le fil conducteur du projet. Il s'agit là d'une condition *sine qua non* pour garder un lien intelligible entre hypothèses d'analyse et données extraites des enregistrements sonores. Les méthodes computationnelles ne sauraient constituer une finalité en elles-mêmes : elles n'ont de valeur qu'en tant qu'instruments au service d'une démarche analytique rigoureuse, dont le musicologue reste le garant. Si notre algorithme permet de guider l'analyste, ce dernier doit aussi s'intéresser *a posteriori* aux descripteurs qui lui sont recommandés, en évitant au maximum de vouloir faire rentrer de force ses impressions sensibles dans les potentielles catégories sémantiques que lui évoquent les descripteurs recommandés.

5. REFERENCES

[1] Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America*, 94, 09 1993.

[2] Ingwer Borg and Patrick Groenen. *Modern Multidimensional Scaling : Theory and Applications (Springer Series in Statistics)*. 08 2005.

[3] Pierre Couprie. Quelques propos sur les outils et les méthodes audionumériques en musicologie. L'interdisciplinarité comme rupture épistémologique. *Revue musicale OICRM*, 6 :25–44, 03 2020.

[4] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4) :1917–1930, 2002.

[5] Jonathan Foote. Visualizing music and audio using self-similarity. In *MULTIMEDIA '99*, 1999.

[6] Jonathan Foote and Matthew L. Cooper. Visualizing musical structure and rhythm via self-similarity. In *International Conference on Mathematics and Computing*, 2001.

[7] Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi : A language for high-performance computation on spatially sparse data structures. In *ACM Transactions on Graphics (SIGGRAPH Asia)*, volume 38, pages 201 :1–201 :16, 2019.

[8] Plotly Technologies Inc. Collaborative data science, 2015.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 2022.

[10] Philippe Lalitte. Du son au sens : vers une approche sub-symbolique de l'analyse musicale assistée par ordinateur. *Musurgia*, 18(1–2) :99–116, 2011.

[11] Mikhail Malt and Emmanuel Jourdan. Le « BSTD » – une représentation graphique de la brillance et de l'écart type spectral, comme possible représentation de l'évolution du timbre sonore. In Xavier Hascher, Mondher Ayari, and Jean-Michel Bardez, editors, *L'analyse musicale aujourd'hui*, pages 107–131. Delatour, Sampzon, 2015.

[12] Geoffroy Peeters. Descripteurs audio : de la simple représentation aux modèles de connaissances. In Pierre Couprie and Alain Bonardi, editors, *Geste sonore et paramètres. L'analyse musicale à l'heure des outils multimédia. Séminaire IReMus*. Paris, France, 2015. Consulté le 26 juin 2025.

[13] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The Timbre Toolbox : Extracting audio descriptors from musical signals. volume 130, pages 2902–2916, 2011.

[14] Igor Stravinsky. The rite of spring, 1974. Originally recorded 1974. Re-released 2017. From *Stravinsky : The Rite of Spring*. Producer : Ray Minshull ; Recording Engineer : Kenneth Wilkinson ; Second Engineer : James Lock.

[15] Igor Stravinsky. Le sacre du printemps, 1975. Originally recorded 1975. Re-released 2019-05-10. Catalogue No. 4840180. Barcode : 00028948401802. Producer : Michael Woolcock ; Balance Engineer : Gordon Parry ; Engineer : Jack Law. From *Stravinsky, Bartók : Ballet Music*.

[16] Igor Stravinsky. Le sacre du printemps, 2010. Originally recorded 2010. Released 2012-01-01. From

Stravinsky : Le Sacre Du Printemps – 100th Anniversary Collector's Edition (20 CD set). Catalogue No. 4783729. Barcode : 0028947837299.

[17] Pierre Alexandre Tremblay, Owen Green, Gerard Roma, James Bradbury, Ted Moore, Jacob Hart, and Alex Harker. The fluid corpus manipulation toolbox, July 2022.