Apprentissage automatique de caractéristiques audio : application à la génération de listes de lecture thématiques

Yann Bayle

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Yann Bayle**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Apprentissage automatique de caractéristiques audio : application à la génération de listes de lecture thématiques

---

Date de soutenance : Mardi 19 Juin 2018

Devant la commission d'examen composée de :

Myriam Desainte-Catherine	Professeure des universités, Univ. Bordeaux	Présidente du jury
Frédéric Bimbot	Directeur de recherche, Univ. Rennes	Rapporteur
Julien Pinquier	Maître de conférences, Univ. Toulouse	Rapporteur
Olivier Adam	Professeur des universités, Univ. Paris Sud	Examinateur
Pierre Hanna	Maître de conférences, Univ. Bordeaux	Directeur
Matthias Robine	Maître de conférences, Univ. Bordeaux	Directeur

– 2018 –

Résumé Ce mémoire de thèse de doctorat présente, discute et propose des outils de fouille automatique de mégadonnées dans un contexte de classification supervisée musicale. L'application principale concerne la classification automatique des thèmes musicaux afin de générer des listes de lecture thématiques homogènes.

Le premier chapitre introduit les différents contextes et concepts autour des mégadonnées musicales et de leur consommation. Le deuxième chapitre s'attelle à la description des problématiques concernant la variété et les proportions inégales des thèmes contenus dans les bases de données musicales existantes dans le cadre d'expériences académiques d'analyse audio. Le troisième chapitre détaille le développement et l'extraction de caractéristiques audio afin de décrire le contenu des morceaux. Ce chapitre explique plusieurs phénomènes psychoacoustiques et utilise des techniques de traitement du signal afin de calculer des caractéristiques musicales. De nouvelles méthodes d'agrégation de caractéristiques audio locales sont proposées afin d'améliorer la classification des morceaux. Le quatrième chapitre décrit l'utilisation des caractéristiques musicales extraites afin de trier les morceaux par thèmes et donc de permettre les recommandations musicales et la génération automatique de listes de lecture thématiques homogènes. Cette partie implique l'utilisation d'algorithmes d'apprentissage automatique afin de réaliser des tâches de classification musicale. Les contributions de ce mémoire sont résumées dans le cinquième chapitre qui propose également des perspectives de recherche dans l'apprentissage automatique et l'extraction de caractéristiques audio.

Mots-clés Annotations musicales automatiques, Apprentissage automatique et profond, Classification supervisée, Fouille de Mégadonnées, Psychoacoustique, Traitement du signal audio numérique

Laboratoire d'accueil Laboratoire Bordelais de Recherche en Informatique (LaBRI), Univ. Bordeaux, CNRS, UMR 5800, F-33400 Talence, France

Title Machine learning algorithms applied to audio features analysis: Application in the automatic generation of thematic musical playlists

Abstract This doctoral dissertation presents, discusses and proposes tools for the automatic information retrieval in big musical databases. The main application is the supervised classification of musical themes to generate thematic playlists.

The first chapter introduces the different contexts and concepts around big musical databases and their consumption. The second chapter focuses on issues concerning the variety and unequal proportions of themes contained in existing music databases as part of academic experiments in audio analysis. The third chapter explains the importance of extracting and developing relevant audio features in order to better describe the content of music tracks in these databases. This chapter explains several psychoacoustic phenomena and uses sound signal processing techniques to compute audio features. New methods for aggregating local audio features are proposed to improve song classification. The fourth chapter describes the use of the extracted audio features in order to sort the songs by themes and thus to allow the musical recommendations and the automatic generation of homogeneous thematic playlists. This part involves the use of machine learning algorithms to perform music classification tasks. The contributions of this dissertation are summarized in the fifth chapter which also proposes research perspectives in machine learning and audio features extraction.

Keywords Big data mining, Machine and Deep learning, Digital audio signal processing, Music information retrieval, Psychoacoustics, Supervised classification

Remerciements

« No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude. »

– Alfred North Whitehead

Je remercie vivement Laurent Demany qui dès 2014 m'a fait découvrir la recherche fondamentale et qui m'a aidé à apprécier le milieu rigoureux mais captivant de la recherche.

Je suis très reconnaissant à Myriam Desainte-Catherine qui m'a donné l'opportunité d'obtenir cette thèse et qui s'est montrée disponible et compréhensive dès mon entrée en école d'ingénieur et jusqu'à la fin de mes trois années de doctorat. Je te remercie pour avoir présidé mon jury de soutenance ainsi que pour l'émotion avec laquelle tu as annoncé que j'étais Docteur.

Je tiens à exprimer toute ma gratitude à Pierre Hanna et Matthias Robine pour leur encadrement et pour les conditions de travail uniques qu'ils m'ont fournies. Je les remercie pour leur perspicacité en traitement du signal, pour leurs points de vue différents mais constructifs ainsi que pour m'avoir fait découvrir les coulisses de la recherche appliquée. Pendant ces trois années vous m'avez soutenu et m'avez aidé à m'épanouir en tant que chercheur autonome en me montrant la voie de la réflexion scientifique.

J'exprime toute ma gratitude à mes deux rapporteurs, Frédéric Bimbot et Julien Pinquier pour leur « gourmandise scientifique » à l'égard de mon mémoire ainsi que pour leurs commentaires constructifs et remarques, suggestions et améliorations pertinentes et de qualité. Je les remercie avec Olivier Adam pour avoir participé à mon jury et pour leur bienveillance concernant mes travaux ainsi que pour les discussions scientifiquement et musicalement intéressantes.

Je remercie l'entreprise NVIDIA qui a accepté l'octroi d'une carte graphique pour que je mène à bien mes expériences. Je remercie le soutien de l'Université de Charles ainsi que du MCIA qui m'ont permis d'utiliser leurs supercalculateurs pour mes expériences. Je remercie les douze participants à mes deux expériences psychoacoustiques pour leur implication, leur compréhension ainsi que les nombreuses heures passées dans les cabines insonorisées. Je remercie également les membres de l'INCIA, Marie-Caroline, Catherine et Max, qui m'ont accompagné durant les expériences psychoacoustiques ainsi que les membres du SCRIME György Kurtág Jr. et Jean-Luc Rouas. J'ai-

merais remercier Florian Iragne de Simbals pour son aide technique et pour sa disponibilité. Je remercie également Thibault Langlois, Fabien Gouyon et Loreto Parisi pour leur aide et pour les explications qu'ils m'ont fournies. Je remercie Jean-Manuel Moussalam de Deezer pour les discussions que nous avons eues ainsi que pour le partage de son point de vue sur l'industrie musicale mondiale actuelle. Je remercie Jean-Baptiste Défossez et Michel Vermeulin pour leur confiance ainsi que pour notre partenariat fructueux. Je remercie Bob L. Sturm pour les échanges et discussions autour des *Horses* et pour son invitation à présenter mes travaux dans son laboratoire à Londres. Je remercie Maria Panteli et Jordi Pons pour leur aide sur l'apprentissage profond. Laci, je te remercie de ne pas avoir compté les heures ni regardé les horaires de travail pour notre projet et Martin pour la collaboration et l'aide que tu nous a fournies et qui a permis d'aboutir non pas juste à un article mais à un projet commun et à une expérience humaine sans frontières. Je remercie Kimberley Malcolm pour son aide et ses relectures en anglais de mes articles et pour sa compréhension envers un domaine qui ne lui est pas familier. Je remercie toutes les personnes de mon laboratoire et plus généralement dans ma vie qui ont participé de près ou de loin au déroulement de ma thèse et je vous prie de m'excuser si tous vos noms ne figurent pas ici mais sachez que je pense également à vous.

Je remercie ma famille (Marie-Anne, Henri, Thomas, Karen, Justine et Guy), ma belle-famille (Nathalie, Anoé et Keijo) et Jadou d'avoir tenté pendant ces trois ans de retenir la signification du sigle de mon laboratoire et non, vous n'aurez pas à m'appeler Docteur Bayle.

« *They say behind every great man there's a woman. While I'm not a great man, there's a great woman behind me.* »[1]. Je remercie Fidji, la personne la plus précieuse dans ma vie pour être à mes côtés et pour m'avoir aidé tout au long de cette thèse et même avant son commencement. Merci pour ta présence, ta patience, ton honnêteté, ton soutien sans faille, tu as su m'aider à canaliser mon énergie « omnidirectionnellement » débordante. « Ainsi que la corde d'une harpe, fortement pincée, peut réveiller ses sœurs, ou qu'une note pure émise par une voix aiguë peut faire frémir le cristal, ainsi tu as suscité la vérité en moi. »[2]

1. *Most courageous athlete of 1945* du quotidien *The Port Arthur News*, February 1946, Meryll Frost
2. *Les Aventuriers de la mer*, 1998, Robin Hobb

Table des matières

Avant-propos

> *« [...] it appears probable that the progenitors of man, either the males or females or both sexes, before acquiring the power of expressing their mutual love in articulate language, endeavoured to charm each other with musical notes and rhythm. »*
>
> – *The Descent of Man*, 1871, Charles Darwin

La lecture de mon mémoire de thèse de doctorat peut s'effectuer en écoutant une liste de lecture musicale conçue pour l'occasion. La liste des morceaux est présentée en Annexe A et il est possible de les écouter sur Deezer [1] ou Spotify [2]. Les morceaux ont été choisis pour illustrer des concepts décrits dans cette thèse ou pour des raisons purement personnelles.

La version numérique de ce mémoire est interactive et les mots en gris renvoient au glossaire ou à la liste des acronymes. Les définitions proposées sont issues du dictionnaire Larousse [3], du Centre National de Ressources Textuelles et Lexicales [4], de la littérature citée ou bien ont été proposées par moi-même.

1. http://www.deezer.com/playlist/4046702022, consulté le 19 Avril 2018.
2. https://open.spotify.com/user/zikbone/playlist/4ytJ3BrPf3iTeJE16hIrEj, consulté le 19 Avril 2018.
3. http://larousse.fr/dictionnaires/francais/, consulté le 19 Avril 2018.
4. http://www.cnrtl.fr, consulté le 19 Avril 2018.

1

Introduction

> *« I was born with music inside me. Music was one of my parts. Like my ribs, my kidneys, my liver, my heart. Like my blood. It was a force already within me when I arrived on the scene. It was a necessity for me like food or water. »*
>
> – *Ray*, 2004, Ray Charles

Afin de comprendre le contexte dans lequel s'inscrit cette thèse, la section 1.1 détaille les enjeux actuels qui entourent la consommation de musique. La section 1.2 définit ensuite la problématique de recherche et le cadre d'insertion des travaux proposés dans ce contexte. Les chapitres de ce mémoire détaillent les méthodes existantes et proposées afin d'atteindre les objectifs définis dans la section 1.2. La section 1.3 décrit enfin les outils de validation des méthodes proposées.

1.1 Enjeux et motivations

> *« Whenever humans come together for any reason, music is there: weddings, funerals, graduation from college, men marching off to war, stadium sporting events, a night on the town, prayer, a romantic dinner, mothers rocking their infants to sleep ... music is a part of the fabric of everyday life. »*
>
> – *This Is Your Brain on Music*, 2007, Daniel J. Levitin

Afin d'appréhender les apports scientifiques de cette thèse, les sections 1.1.1, 1.1.2 et 1.1.3 présentent respectivement les contextes musical, sociétal et académique qui s'y rapportent.

1.1.1 Contexte musical

« La limite entre la musique que vous possédez et celle que vous voulez écouter s'estompe. »
– Contextualize your listening: The playlist as recommendation engine, 2011, Benjamin Fields

Les premières traces d'instruments de musique ont été datées de 35 000 ans. Il apparaît néanmoins que l'existence de la musique produite par des êtres humains est antérieure à cette date [Conard *et al.*, 2009]. La musique constitue en effet chez l'être humain un vecteur de transmission culturelle et présente des rôles social, médical et psychologique [Huron, 2001; Paiva, 2013]. Le plus ancien chant connu est parvenu jusqu'à notre époque parce qu'il a été gravé sur une tablette d'argile [1]. Ce support physique a permis la redécouverte de la culture mésopotamienne et son analyse montre une première pérennisation matérielle de la musique, en rupture avec les représentations musicales vivantes pratiquées jusqu'alors.
Depuis le XIX^{ème} siècle, la possibilité d'enregistrer de la musique et de la propager grâce à de nouvelles formes de diffusion a révolutionné la consommation musicale moderne. Les supports matériels tels que le vinyle et le *Compact Disc* (CD) ont de plus largement démocratisé l'écoute de la musique hors du cadre de la représentation en temps réel. Plus récemment encore, la révolution apportée par la mise en place d'internet a fait basculer la consommation musicale sur des supports dématérialisés. Depuis 2017, les revenus mondiaux du secteur musical proviennent en effet majoritairement de la consommation sur supports dématérialisés, ce qui souligne le délaissement de l'achat et de la possession de supports physiques, comme l'indiquent les rapports de l'International Federation of the Phonographic Industry (IFPI) [2], d'Ipsos [3], de BusinessInsider [4], du RIAA [5] et celui de Nielsen [6].

La liste de lecture musicale, ou *playlist*, qu'il est possible d'écouter en ligne grâce aux sites web de streaming audio, constitue aujourd'hui le canal le plus utilisé pour la transmission musicale [Fields, 2011]. L'un des intérêts principaux de ces plateformes réside notamment dans le fait que les auditeurs peuvent choisir les morceaux qu'ils écoutent, sans être contraints par l'ordre de lecture imposé par un vinyle, par exemple.

1. https://lejournal.cnrs.fr/articles/sur-un-air-de-musique-antique, consulté le 19 Avril 2018.
2. http://www.ifpi.org/downloads/GMR2018.pdf, consulté le 19 Avril 2018.
3. http://www.snepmusique.com/?p=16023&preview=true, consulté le 19 Avril 2018.
4. http://www.businessinsider.com/inside-spotify-and-the-future-of-music-streaming?IR=T, consulté le 19 Avril 2018.
5. http://riaa.com/media/D1F4E3E8-D3E0-FCEE-BB55-FD8B35BC8785.pdf, consulté le 19 Avril 2018.
6. http://www.melodicrock.com/sites/default/files/Nielsen%20Music%202015%20mid-year%20report.pdf, consulté le 19 Avril 2018.

De plus, il est aujourd'hui possible d'écouter davantage de musique variée grâce à l'explosion des données mises à disposition sur internet. Plus de 4 000 nouveaux CDs étaient en effet ajoutés tous les mois sur les plateformes de streaming en 1999 [Pachet et Roy, 1999] alors que ce chiffre a au moins quadruplé en 2015 selon le rapport du RIAJ [1]. Plusieurs questions découlent de ces constats. En tant qu'auditeur, comment découvrir et écouter dans cette profusion de données de la musique susceptible de me plaire ? En tant que musicien, comment rendre ma musique accessible et me démarquer des autres morceaux ? En tant qu'industriel, comment traiter les nouveaux morceaux qui arrivent en temps réel et comment gérer le catalogue existant ?

La section suivante décrit par conséquent les outils mis en place pour accompagner les auditeurs et les musiciens dans cette révolution numérique.

1.1.2 Contexte sociétal

> *« Now, the making of a good compilation tape is a very subtle art. Many do's and don'ts. First of all you're using someone else's poetry to express how you feel. This is a delicate thing. »*
>
> – *High Fidelity*, 1995, Nick Hornby

Les émotions, sentiments, genres et activités sont les thèmes les plus utilisés pour décrire un morceau [Inskip *et al.*, 2012; Tintarev *et al.*, 2017] mais également lors de la recherche de listes de lecture musicale. Pour chacune de ces catégories, le tableau 1.1 illustre des exemples de thèmes musicaux, également appelés étiquettes, ou *tags*.

Tableau 1.1 – Exemple de thèmes musicaux couramment utilisés.

Catégorie	Thème
Émotion	Joie, Tristesse, Surprise
Sentiment	Épanouissement, Fascination, Anxiété
Genre	Rock, Jazz, Pop
Activité	Sport, Danse, Cuisine

L'utilisation de ces catégories musicales répond à des contraintes sociales et psychologiques [Huron, 2000] mais également à un aspect pratique d'utilisation [Aljanaki, 2016]. Cependant, les titres des morceaux ne contiennent généralement pas d'informations relatives à ces catégories, ce qui rend ces morceaux difficilement accessibles sur les plateformes de streaming. La chanson *Trains* de Porcupine Tree ne contient par exemple aucun son produit par un train. De plus, quatre des vingt millions de morceaux proposés à l'écoute par le site de

1. http://www.riaj.or.jp/riaj/pdf/issue/industry/RIAJ2015E.pdf, consulté le 19 Avril 2018.

streaming Spotify [1] en 2013 n'avaient jamais été écoutés [2]. Les sites de streaming Spotify et Deezer [3] emploient des éditeurs de listes de lecture musicale thématiques afin de suggérer de nouveaux morceaux à leurs utilisateurs. Ces éditeurs sont au nombre de 32 [4] pour Spotify [5] et de 50 pour Deezer [6] en 2018. Ils associent des thèmes à chaque morceau qu'ils écoutent et les rassemblent au sein de listes de lecture afin de les rendre accessibles aux utilisateurs de leur plateforme.

Cependant, face à l'augmentation constante de la quantité de morceaux disponibles et produits, les éditeurs de listes de lecture ne parviennent pas à tous les annoter manuellement. Spotify propose donc à ses éditeurs le *Truffle Pig*, qui constitue son outil automatique privé d'assistance à la création de listes de lecture musicale [7]. Cet outil est un algorithme analysant automatiquement tous les morceaux contenus dans une base de données musicale afin de leur associer un thème. Un éditeur peut alors utiliser le *Truffle Pig* pour compléter une liste de lecture musicale thématique de morceaux auxquels il n'aurait autrement pas eu accès. Malgré l'utilisation de cet outil, tous les morceaux analysés par le *Truffle Pig* ne deviennent pas accessibles grâce à une recherche par mots-clés. Les méthodes d'analyse musicale du *Truffle Pig* ne sont en effet pas parfaites et produisent des erreurs d'annotation qui doivent être corrigées manuellement. La section suivante décrit et commente donc les études qui tentent d'améliorer les performances des algorithmes d'analyse musicale automatique.

1.1.3 Contexte académique

> *« We are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline. »*
>
> – *Conjectures and Refutations: The Growth of Scientific Knowledge*, 1963, Karl R. Popper

Le champ de recherche entourant l'analyse musicale automatique est pluridisciplinaire et a émergé de la nécessité de gérer la quantité croissante des données musicales digitales [Futrelle et Downie, 2002, 2003]. Ces disciplines

1. https://www.spotify.com, consulté le 19 Avril 2018.
2. https://news.spotify.com/us/2013/10/07/the-spotify-story-so-far/, consulté le 19 Avril 2018.
3. http://www.deezer.com, consulté le 19 Avril 2018.
4. http://www.businessinsider.com/inside-spotify-and-the-future-of-music-streaming?IR=T, consulté le 19 Avril 2018.
5. https://en.wikipedia.org/wiki/The_Echo_Nest, consulté le 19 Avril 2018.
6. https://www.deezer.com/fr/company/about, consulté le 19 Avril 2018.
7. https://techcrunch.com/2014/10/19/the-sonic-mad-scientists/, consulté le 19 Avril 2018.

incluent notamment l'intelligence artificielle, le traitement du signal, la psychologie, l'épistémologie et la psychoacoustique et font également appel à des musicologues, des musiciens et des libraires. De l'interaction complexe de ces champs de recherche découlent des travaux d'analyse musicale qui doivent faire face à deux défis principaux dans la création automatique de listes de lecture.

Le premier défi consiste à accéder à des données musicales utilisables dans un but scientifique. Plus de 150 millions de morceaux sont actuellement disponibles sur les plateformes de streaming [1]. Cependant, la base de données musicales de recherche la plus utilisée contient 1 000 morceaux [Tzanetakis et Cook, 2002; Sturm, 2014b]. Le nombre de morceaux disponibles à l'écoute diffère donc largement de celui utilisé par les chercheurs pour leurs expériences d'analyse musicale. Il paraît alors complexe d'obtenir des conclusions significatives quant aux algorithmes d'analyse musicale et par conséquent d'améliorer efficacement les méthodes utilisées. Or, il est difficile de mettre en place une base de données musicales académique, principalement en raison de la gestion des droits d'auteur musicaux. Le manque de moyens financiers et logistiques académiques ainsi que la faible propension des chercheurs aux partenariats avec des entreprises réduit également la mise en place et la gestion d'une telle quantité de données musicales.

Le second défi provient du fossé sémantique [Wiggins, 2009] qui existe entre les chercheurs, les auditeurs de musique et les entrepreneurs. Ce fossé sémantique sépare la description algorithmique d'un son produit dans le domaine acoustique d'une description musicale perceptive pertinente pour l'auditeur. Un fossé sémantique sépare par exemple la présence de certaines fréquences dans un morceau et la perception d'une émotion suite à l'écoute de ce morceau.

Méthodes de génération automatique de listes de lecture

Il n'existe actuellement pas de consensus concernant la définition d'une liste de lecture musicale [Fields, 2011]. La notion de liste de lecture retenue et utilisée dans les chapitres qui suivent est celle proposée par Fields [2011] et Bonnin et Jannach [2014] et qui désigne des morceaux qui sont regroupés afin d'être écoutés ensemble. Ce regroupement est effectué pour répondre à une thématique musicale. À partir de cette définition, Fields [2011] a réalisé un état de l'art sur la création de listes de lecture et d'autres études ont été proposées depuis 2011. Les principaux travaux dans la génération automatique de listes de lecture musicale sont décrits ci-après.

La méthode proposée par Chen *et al.* [2012] consiste à générer de nouvelles listes de lecture musicale suivant les mêmes proportions thématiques qu'un ensemble de modèles de listes existant. L'avantage majeur de leur méthode provient du fait qu'il n'est pas nécessaire d'extraire des caractéristiques audio

1. http://www.ifpi.org/downloads/GMR2017.pdf, consulté le 19 Avril 2018.

des morceaux. L'inconvénient principal de cette méthode provient toutefois de la récupération de modèles de listes de lecture sur les plateformes de streaming. En effet, les listes de lecture qui y sont présentes ne garantissent pas forcément de cohérence thématique des morceaux qu'elles contiennent.

Une autre méthode mise en place par Aizenberg *et al.* [2012] utilise l'historique d'écoute des utilisateurs afin de rassembler les morceaux similaires dans une même liste de lecture. Toutefois, Hidasi *et al.* [2016] posent comme hypothèse que l'utilisation de l'ensemble de l'historique d'écoute ne permet pas des recommandations prenant en compte l'évolution des goûts musicaux d'un auditeur. L'une des manifestations de ce phénomène est notamment visible pour les morceaux présents dans les *Tops*[1], qui sont régulièrement renouvelés. Hidasi *et al.* [2016] proposent alors de restreindre l'historique d'écoute à une session de connexion afin de prendre en compte une évolution temporelle des préférences musicales. Cette méthode affiche l'inconvénient de ne pas pouvoir recommander de nouveaux morceaux qui n'ont reçu aucune interaction mais elle ne nécessite pas le calcul, parfois chronophage, de caractéristiques audio. Ikeda *et al.* [2017] et Tintarev *et al.* [2017] argumentent toutefois que le calcul des caractéristiques audio est pertinent puisqu'il permet de créer des listes de lecture qui améliorent les transitions musicales grâce à la minimisation des différences de certaines caractéristiques audio entre deux morceaux. Les auteurs valident leurs méthodes algorithmiquement mais montrent, après une expérience utilisateur, que du point de vue des auditeurs, l'ordre des morceaux a peu d'importance sur la qualité perçue d'une liste de lecture musicale [Tintarev *et al.*, 2017]. Ces auteurs montrent notamment que l'un des aspects importants attendus par les auditeurs concerne la présence de nouveaux morceaux dans les listes de lecture générées automatiquement.

Afin d'augmenter le taux de découverte musicale des auditeurs, Zhang *et al.* [2012] ont donc proposé d'introduire de la sérendipité dans la génération automatique des listes de lecture, ce qui consiste en l'ajout de morceaux de manière aléatoire. La sérendipité se définit en effet comme l'art de faire une découverte par hasard et se traduit chez l'auditeur par l'écoute d'une liste de lecture dont certains morceaux semblent avoir été introduits aléatoirement. L'ajout de ces morceaux dans cette liste n'est en fait pas entièrement dû au hasard. De tels morceaux ont en effet été ajoutés dans cette liste de lecture thématique car ils possédaient des similarités avec le thème affiché. Il est par exemple possible de suggérer l'ajout d'un morceau de Pop joyeux dans une liste de lecture initialement définie comme contenant du Rock joyeux. Le réglage du taux de sérendipité est toutefois biaisé par les performances imparfaites des méthodes d'analyse musicale automatique. Une sérendipité nulle ne garantit pas, en effet, de liste de lecture ne contenant que des morceaux appartenant à un thème donné. Or, la présence de quelques morceaux « hors thème » suffit à diminuer

1. https://www.deezer.com/fr/channels/charts, consulté le 19 Avril 2018.

la qualité de service perçue d'un système de recommandation [Chau *et al.*, 2013] puisque les utilisateurs sont plus sensibles aux messages négatifs qu'aux messages positifs [Yin *et al.*, 2010]. L'étude proposée par Pohle *et al.* [2007] a montré que la similarité entre plusieurs morceaux est une exigence de la part des auditeurs pour juger de la qualité d'une liste de lecture musicale. Il apparaît alors pertinent de concevoir des algorithmes de génération de listes de lecture musicale répondant parfaitement à un thème avant d'y introduire de la sérendipité. Les recherches proposées dans ce mémoire ont donc été orientées suivant cette optique.

La proposition et l'évaluation d'une nouvelle méthode de reconnaissance des thèmes musicaux dans une base passe par la mesure des résultats objectifs fournis par ces méthodes. L'interprétation des résultats des méthodes algorithmiques est par ailleurs sujette à discussion, comme cela est illustré dans la section suivante.

1.1.4 *Horses*

> *« Horse: A system appearing capable of a remarkable human feat, e.g., music genre recognition, but actually working by using irrelevant characteristics (confounds). »*
>
> – *A simple method to determine if a music information retrieval system is a "horse"*, 2014, Bob L. Sturm

Afin de valider un fait scientifique ou une méthode algorithmique, il est préférable de mettre en perspective l'observation des résultats obtenus vis-à-vis du protocole utilisé pour obtenir de tels résultats. L'anecdote qui suit permet d'illustrer ce propos.

En 1891, le professeur de mathématiques Wilhelm von Osten a déclaré son cheval Hans comme étant intelligent puisque ce dernier était capable de résoudre des problèmes mathématiques [Pfungst, 1911]. Ce cheval était capable de donner les facteurs communs du nombre 28 en frappant le sol de son sabot. Après de multiples expériences, le cheval fut déclaré astucieux mais il apparut par la suite qu'il ne résolvait pas les problèmes qui lui étaient posés. Cette conclusion est survenue suite à une série d'expériences au cours desquelles le cheval était interrogé et devait alternativement répondre à une question en ayant les yeux bandés ou non. Or, le cheval répondait correctement dans 86% des cas lorsqu'il n'avait pas les yeux bandés alors que ses performances chutaient à 6% dans la condition opposée. Au cours des expériences, Pfungst [1911] a découvert que ce cheval utilisait des indices visuels correspondant aux attitudes de l'interrogateur afin de répondre correctement à la question posée. La capacité de ce cheval à résoudre des problèmes dépendait donc de l'aide des humains [Lesimple *et al.*, 2012]. Le fait que l'interrogateur lui fournissait involontairement des indices est maintenant connu sous le nom du *Clever Hans*

Effect. Une méthode qui aujourd'hui semble résoudre un problème alors que ce n'est pas le cas est donc désignée comme étant un Horse.

L'un des exemples les plus récents de la présence d'un Horse concerne une méthode de prédiction de rendement de cultures agricoles. Peng *et al.* [2018] ont en effet tenté de proposer un modèle plus précis à une échelle temporelle inférieure à celle de l'état de l'art. En analysant les résultats de leur modèle, Peng *et al.* [2018] ont obtenu un rendement final identique à celui que produisaient les méthodes existantes. L'équipe de recherche s'est cependant rendue compte que les modèles existants fournissaient des résultats corrects mais pour des mauvaises raisons. Les modèles existants sous-estimaient en effet la biomasse et surestimaient l'index de récolte. Cet exemple souligne bien la différence à effectuer entre deux méthodes qui fournissent une bonne réponse mais dont l'une modélise incorrectement le phénomène sous-jacent.

Il existe également et évidemment des Horses au sein des méthodes d'analyse audio et une présentation sur le sujet a été effectuée au Queen Mary University of London (QMUL)[1]. Des outils sont détaillés et proposés dans chacun des chapitres afin de réduire la probabilité qu'une méthode d'analyse musicale ne soit un Horse.

1.2 Problématiques

> *« Pour un esprit scientifique toute connaissance est une réponse à une question. S'il n'y a pas eu de question il ne peut pas y avoir connaissance scientifique. Rien ne va de soi. Rien n'est donné. Tout est construit. »*
>
> – *La formation de l'esprit scientifique*, 1938, Gaston Bachelard

Les sections précédentes ont dépeint la quantité grandissante de morceaux proposés à l'écoute et qui se heurte au manque de moyens qui pourraient permettre de trier les mégadonnées, ou *big data*, musicales. Les mégadonnées désignent notamment des données dont le volume, la variété et la croissance rapide [McAfee *et al.*, 2012] complexifient leur traitement par des êtres humains ou des outils informatiques classiques de gestion de bases de données et de l'information. Dans ce contexte, l'analyse audio automatisée utilisée afin de générer des listes de lecture musicale constitue une solution au tri de ces mégadonnées musicales. La question de recherche qui découle de ce constat vise par conséquent à étudier les conditions dans lesquelles il est possible de générer une liste de lecture musicale thématique parfaite, c'est-à-dire ne contenant aucun morceau hors thème. Afin de répondre à cette question, trois étapes principales sont identifiées et représentées dans la figure 1.1.

1. http://c4dm.eecs.qmul.ac.uk/horse2017/, consulté le 19 Avril 2018.

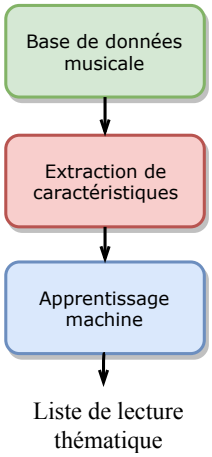


FIGURE 1.1 – Chaîne de traitement permettant de constituer des listes de lecture musicale thématiques.

La figure 1.1 indique que la génération d'une liste thématique requiert tout d'abord une base de données musicales. La qualité des résultats des méthodes de génération de listes de lecture dépend de l'exactitude des informations référencées dans cette base. La deuxième étape consiste à extraire des caractéristiques audio pertinentes des morceaux de la base. Cette étape nécessite la mise en place de mesures quantifiant les capacités des caractéristiques audio extraites à identifier correctement les thèmes des morceaux étudiés dans le cadre d'une classification automatique. La troisième étape utilise des méthodes d'apprentissage machine afin d'associer un thème à chaque morceau. Il est par la suite possible de constituer une liste de lecture thématique en regroupant les morceaux affichant le thème souhaité. La notion de thème musical est définie dans la section 2.2. L'importance de cette dernière étape réside dans l'utilisation et l'amélioration des méthodes d'apprentissage machine afin de s'approcher d'une liste de lecture thématique parfaite, c'est-à-dire contenant des morceaux thématiquement homogènes. La notion de perfection est discutée dans la section 4.1.2. Cependant, si une telle méthode est proposée, comment est-il possible de garantir que celle-ci n'est pas un Horse ?

Ces questions constituent le fil conducteur de ce mémoire et la méthodologie utilisée pour y répondre se rapporte à la figure 1.1. L'évaluation des méthodes proposées nécessite de définir au préalable des métriques objectives qui sont présentées dans la section suivante.

1.3 Métriques

« Over the piano was printed a notice: Please do not shoot the pianist. He is doing his best. »
– Personal Impressions of America, 1883, Oscar Wilde

Plusieurs métriques ont été proposées afin de mesurer la qualité d'une méthode d'analyse audio *per se* mais également afin de comparer objectivement deux méthodes entre elles. Cette section détaille les métriques d'intérêt pour la génération automatique de listes de lecture musicale. Afin de comprendre la définition des métriques ainsi que leur intérêt, un exemple fictif de visualisation de données est proposé dans la figure 1.2.

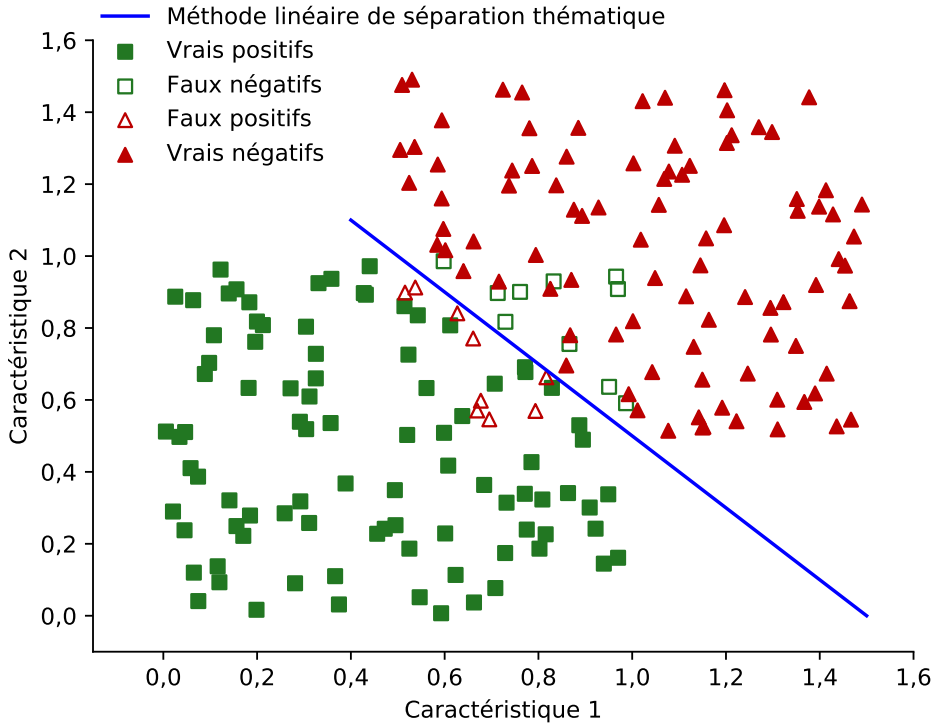


FIGURE 1.2 – Les morceaux du thème 1 correspondant à du Rock sont représentés par les carrés. Les triangles représentent quant à eux les morceaux qui ne sont pas du Rock. Chaque morceau possède deux caractéristiques représentées par chacun des axes. La droite bleue représente le seuil de décision d'une méthode linéaire de séparation thématique.

La droite représente le seuil de décision d'une méthode linéaire qui sépare les morceaux détectés comme étant du Rock en bas à gauche et le reste des

morceaux en haut à droite. Cette méthode apparaît imparfaite puisque certains morceaux ont été attribués au mauvais thème. Afin de comptabiliser les morceaux qui ont été correctement identifiés ainsi que ceux qui ne l'ont pas été, quatre métriques sont couramment utilisées. Ces quatre métriques sont résumées dans le tableau 1.2.

Tableau 1.2 – Nomenclature des prédictions associées par un algorithme à chaque instance en fonction du thème original.

	Morceau du thème 1	Morceau n'appartenant pas au thème 1
Morceau détecté comme appartenant au thème 1	Vrai positif (VP)	Faux positif (FP)
Morceau détecté comme n'appartenant pas au thème 1	Faux négatif (FN)	Vrai négatif (VN)

Les carrés pleins représentent des morceaux de Rock qui ont correctement été détectés et qu'on dénomme vrais positifs (VP). Les carrés vides représentent des morceaux de Rock qui n'ont pas été détectés comme étant du Rock et dénotés faux négatifs (FN). Les triangles pleins représentent des morceaux qui ont correctement été détectés comme n'étant pas du Rock, ils sont donc appelés vrais négatifs (VN). Les triangles vides représentent des morceaux identifiés comme étant du Rock mais qui n'en sont pas, ils sont donc considérés comme étant des faux positifs (FP).

Un système d'analyse thématique idéal n'afficherait aucun faux négatif ni faux positif, ce qui n'est pas le cas du système dont les résultats sont présentés dans la figure 1.2. Afin d'améliorer l'exemple de système d'identification de thème représenté dans la figure 1.2, il faudrait que ce système utilise une méthode de prise de décision non linéaire ou bien que les caractéristiques audio utilisées soient plus distinctives. Le chapitre 3 détaille des méthodes d'extraction de caractéristiques audio pour la description musicale. Le chapitre 4 présente ensuite plus en détails les méthodes d'apprentissage machine utilisées en analyse musicale.

Par ailleurs, le dénombrement de vrais positifs et négatifs peut se révéler identique pour deux bases de données de tailles distinctes alors que la proportion d'éléments à trouver dans ces bases est très différente. En présence d'un tel déséquilibre, il s'avère primordial d'utiliser une mesure qui prenne en compte ce rapport. Le seul dénombrement de vrais positifs et négatifs ne suffit alors pas à comparer deux méthodes d'analyse entre elles et il faut par conséquent rapporter les résultats obtenus à la taille totale des bases de données musicales considérées. Afin de prendre en compte la taille de la base de données dans une telle comparaison, la précision P_t et le rappel R_t sont les métriques les plus souvent utilisées. La précision P_t pour le thème t est définie comme étant le nombre de morceaux correctement associés au thème t par rapport à tous les morceaux associés à ce thème, comme indiqué dans l'équation 1.1.

$$P_t = \frac{VP_t}{VP_t + FP_t} \tag{1.1}$$

où

- VP_t est le nombre de vrais positifs pour le thème t,
- FP_t est le nombre de faux positifs pour le thème t.

La précision est comprise dans l'intervalle [0 ;1] et les meilleures méthodes ont une précision proche de 1.

Le rappel R_t pour le thème t correspond au nombre de morceaux correctement associés au thème t par rapport au nombre total de morceaux de ce thème, comme indiqué dans l'équation 1.2.

$$R_t = \frac{VP_t}{VP_t + FN_t} \tag{1.2}$$

où

- VP_t est le nombre de vrais positifs pour le thème t,
- FN_t est le nombre de faux négatifs pour le thème t.

Le rappel est compris dans l'intervalle [0 ;1]. Une valeur proche de 1 indique que la méthode a correctement identifié la majorité des morceaux d'un thème qui étaient présents dans la base de données musicales.

La moyenne harmonique de la précision et du rappel, dénotée *f-score*, fait également partie des mesures les plus couramment utilisées et est définie par l'équation 1.3.

$$\text{F-score} = \frac{1}{T}\sum_{t=1}^{T} 2\frac{P_t R_t}{P_t + R_t} \tag{1.3}$$

où

- T est le nombre de thèmes,
- P_t est la précision obtenue pour le thème t,
- R_t est le rappel obtenu pour le thème t.

Lors de la comparaison de deux méthodes, la plus performante est celle qui a la valeur de *f-score* la plus proche de 1.

Dans le cas où l'on souhaite mesurer les performances globales d'une méthode pour plusieurs thèmes, la moyenne arithmétique des vrais positifs pour tous les thèmes $\overline{VP}$, plus généralement appelée *accuracy*, est utilisée et est définie dans l'équation 1.4.

$$Accuracy = \frac{1}{N} \sum_{t=1}^{T} VP_t \tag{1.4}$$

où

- N est le nombre de morceaux dans la base de données musicales,
- T est le nombre de thèmes,
- VP_t est le nombre de vrais positifs obtenus pour le thème t.

L'*accuracy* est comprise dans l'intervalle [0 ;1] et la performance d'une méthode est d'autant plus importante que la valeur de son *accuracy* est proche de 1.

De plus, certaines méthodes indiquent le thème détecté pour un morceau mais fournissent également une probabilité que ce morceau appartienne à ce thème. La prise en considération de ces probabilités permet une comparaison plus fine entre deux méthodes. Dans certains cas, il est préférable d'utiliser une méthode qui détecte de manière sûre un nombre réduit de vrais positifs plutôt qu'une méthode qui trouve davantage de vrais positifs mais avec un intervalle de confiance moindre. La moyenne logarithmique de la fonction de coût, ou *loss*, est une métrique qui prend en compte ces probabilités. Le *loss* quantifie la somme des différences entre la probabilité qu'un morceau appartienne à un thème et l'annotation de ce morceau. Comme indiqué dans l'équation 1.5, avec une méthode de détection parfaite cette métrique tend vers zéro.

$$Loss = -\frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T} y_{mt} \ln(p_{mt}) \tag{1.5}$$

où

- M est le nombre de morceaux évalués,
- T est le nombre de thèmes,
- y_{mt}, compris dans $[0, 1]$, indique si le morceau m appartient au thème t,
- p_{mt} est la probabilité que le morceau m appartienne au thème t.

Par ailleurs, une méthode attribue par défaut un thème à un morceau si la probabilité d'appartenance de ce morceau au thème est supérieure au seuil de 50%. Il peut donc être intéressant de faire varier ce seuil afin d'obtenir différents résultats pour une méthode donnée en fonction de l'application souhaitée. Certaines applications critiques (tel qu'une reconnaissance de panne industrielle) ou affichant des contraintes en temps réel (tel que les voitures autonomes) peuvent en effet nécessiter une garantie de précision plus importante. Il est

également possible de comparer deux méthodes en évaluant leurs performances respectives pour l'ensemble de ces seuils. La représentation couramment utilisée est la fonction d'efficacité du récepteur ou courbe sensibilité/spécificité qui représente le taux de vrais positifs ($\frac{VP}{VP+FN}$) en fonction du taux de faux positifs ($\frac{FP}{FP+VN}$). Un exemple de fonction d'efficacité du récepteur est présenté dans la figure 1.3.

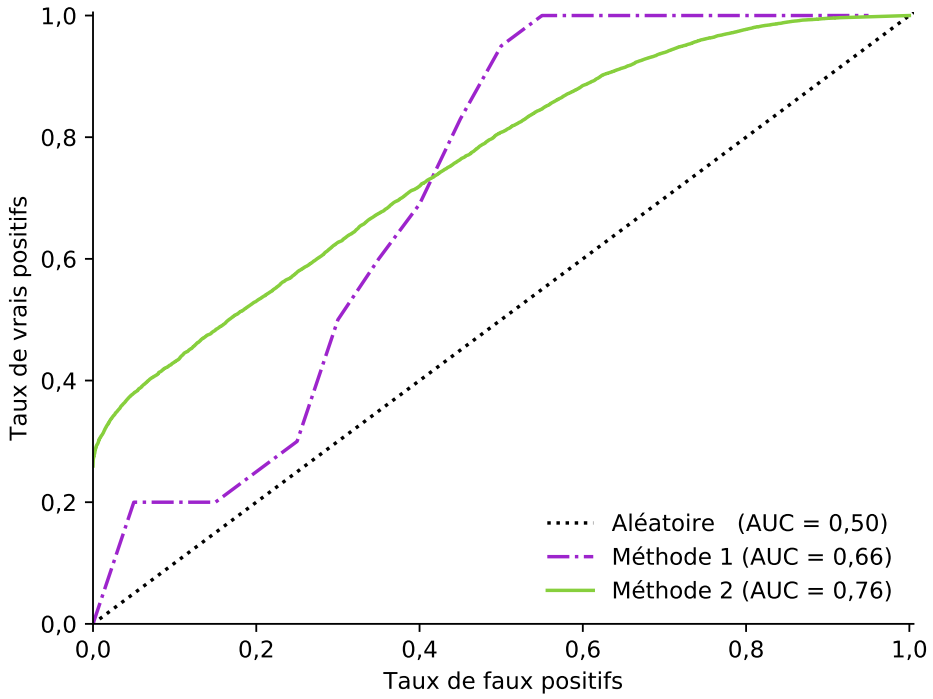


FIGURE 1.3 – Comparaison des courbes de sensibilité/spécificité pour deux méthodes par rapport à une méthode aléatoire. La valeur de l'Area Under the Curve (AUC) est donnée entre parenthèses pour chaque méthode.

La figure 1.3 compare les courbes de sensibilité/spécificité pour plusieurs méthodes et seuils de probabilité. Outre l'aspect visuel de ces courbes, la valeur de l'aire sous ces courbes, en anglais Area Under the Curve (AUC), constitue un indicateur de comparaison de ces différentes méthodes. L'AUC est comprise dans l'intervalle $[0; 1]$ et la méthode de classification utilisée est d'autant plus performante que l'AUC est importante. L'AUC de la méthode 2 est supérieure à celle de la méthode 1 et dans la majorité des cas il est préférable d'utiliser la méthode 2. Toutefois, en fonction de l'application considérée l'AUC n'est pas toujours la meilleure métrique à utiliser. Dans le diagnostic médical par exemple, il est en effet préférable d'utiliser une méthode qui affiche, pour un taux de faux positifs constant, un taux de vrais positifs supérieur à

celui d'autres méthodes. Dans ce cas particulier, la méthode 1 est préférable puisque, comme indiqué sur la figure 1.3, son taux de vrais positifs est supérieur à celui de la méthode 2 à partir d'un taux de faux positifs de 45%. Néanmoins, une courbe de sensibilité/spécificité utilise les taux de vrais et de faux positifs qui ne renseignent pas sur la disproportion d'éléments dans un ou plusieurs thèmes.
Dans le cas où la proportion entre les thèmes est déséquilibrée, il est donc préférable d'utiliser la courbe de précision/rappel dont deux exemples sont présentés dans la figure 1.4. Sur une telle courbe, la meilleure méthode est

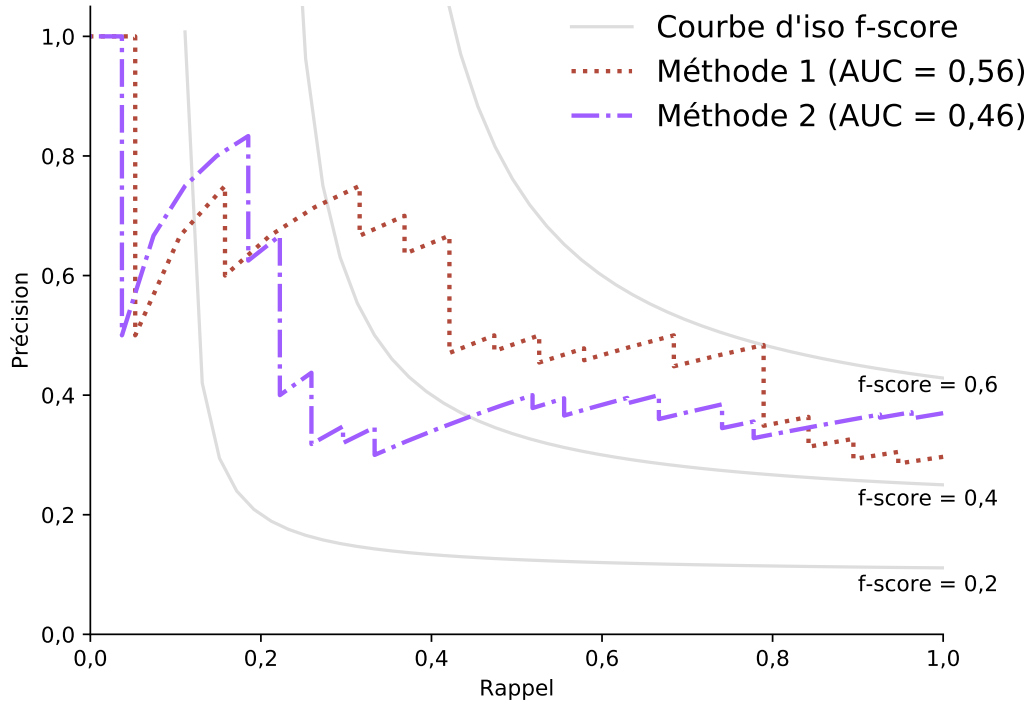


FIGURE 1.4 – Comparaison des courbes de précision/rappel pour deux méthodes. La valeur de l'Area Under the Curve (AUC) est donnée entre parenthèses pour chaque méthode.

celle qui affiche des points dans le cadran supérieur droit, qui correspond à une précision et un rappel proches de 1. La figure 1.4 indique que la méthode 1 offre globalement de meilleures performances que la méthode 2 puisqu'elle affiche une aire sous la courbe supérieure. De même que pour la courbe de sensibilité/spécificité, la valeur de l'aire sous la courbe de précision/rappel fournit une métrique synthétique de comparaison de différentes méthodes. Toutefois, en fonction de l'application souhaitée il peut être intéressant de regarder des points particuliers de chaque courbe plutôt que l'aire correspondante. Pour un rappel de 1 par exemple, la méthode 2 est préférable à la méthode 1 puisqu'elle

affiche une précision de 0,38 alors que la méthode 1 affiche une précision de 0,30.

Les métriques introduites dans cette section seront utilisées tout au long du mémoire afin d'évaluer les méthodes existantes et celles proposées. Cependant, avant de pouvoir utiliser de telles méthodes, il est d'abord nécessaire d'avoir accès à une base de données de morceaux, comme souligné par la figure 1.1. Le chapitre qui suit décrit donc les bases de données musicales existantes et qui permettent l'accès à ces morceaux.

2

Bases de données musicales

> *« Données, adj. pris subst. terme de Mathématique, qui signifie certaines choses ou quantités, qu'on suppose être données ou connues, & dont on se sert pour en trouver d'autres qui sont inconnues, & que l'on cherche. Un problème ou une question renferme en général deux sortes de grandeurs, les données & les cherchées, data & quæsita. V. Problème, &c. »*
>
> – *L'Encyclopédie*, 1759, Denis Diderot et Jean le Rond d'Alembert

L'un des objectifs principaux de cette thèse est de détecter et de différencier automatiquement les thèmes des morceaux d'une base de données afin de générer des listes de lecture musicale thématiques. La notion de base de données musicales qui est celle utilisée dans ce mémoire est à différencier de la notion plus générale de base de données structurées qui utilise des outils de mise en forme, de stockage et de structuration et qui est généralement intégrée dans un système de gestion de bases de données. Ainsi, le développement de méthodes d'extraction d'informations nécessite préférablement des bases de données musicales normalisées et en libre accès afin de garantir un certain degré de réplicabilité des expériences afférentes. Dans les contextes académique et industriel, la qualité de ces bases de données est par ailleurs d'autant plus cruciale qu'elle détermine les conclusions effectuées quant aux méthodes d'extraction d'information.

La communauté de recherche dans le domaine de la vision par ordinateur a par exemple mis en place trois bases de données qui sont celle du Modified National Institute of Standards and Technology (MNIST) [LeCun *et al.*, 1998], celle du Canadian Institute For Advanced Research (CIFAR) [Krizhevsky et Hinton, 2009] et ImageNet [Deng *et al.*, 2009]. Celles-ci se sont révélées être essentielles dans le développement de nouvelles méthodes d'extraction d'informations, plus performantes que celles de l'état de l'art. L'existence de ces bases de données permet donc de constater l'évolution des performances et des

progrès des méthodes de reconnaissance d'images [1] sur des bases de référence [2]. Les trois bases de données mentionnées précédemment contiennent un nombre conséquent d'images libres de droit qui sont annotées en fonction des objets qu'elles contiennent. Il existe notamment 60 000 images dans la base de données du CIFAR, 70 000 images dans celle du MNIST et 10 000 000 d'images dans ImageNet. Le nombre d'images contenues dans ces bases de données de recherche représente une faible quantité comparé aux 350 millions d'images qui ont été ajoutées en moyenne chaque jour sur Facebook en 2013 [3]. En comparaison, environ 40 000 nouveaux morceaux ont été ajoutés en moyenne chaque jour [4] sur la plateforme de streaming audio SoundCloud [5] entre 2016 et 2017. Au total, la plateforme de streaming audio qui propose le plus de morceaux est l'entreprise SoundCloud puisqu'elle référence plus de 150 millions [6] de titres. Il faut toutefois nuancer la quantité de morceaux annoncée par les sites de streaming par les nombreux doublons et rééditions de morceaux qui existent sur ces sites. En résumé, il existe un écart considérable entre le nombre d'images existantes et produites quotidiennement et le nombre de morceaux également déjà disponibles et créés chaque jour. Cet écart se reflète dans les bases de données académiques et explique en partie la faible quantité de morceaux disponibles pour la recherche.

Les méthodes d'extraction d'informations musicales n'ont pas pu être testées et comparées à la même échelle que celles relatives au domaine de l'image sur de grandes bases de données. Cette différence en terme de base provient principalement de la facilité à produire et à diffuser une image, grâce à un smartphone par exemple, comparée à la complexité de production d'un morceau.

Le deuxième frein au développement de bases de données musicales concerne la création et la vérification des annotations, également appelées *groundtruths*. Comme mentionné précédemment, il est primordial de constituer une base d'annotations relative aux éléments présents et absents dans la base. La différence majeure entre des annotations visuelles et auditives provient du type de média puisque la consommation d'une image est différente de celle d'un morceau. Le temps passé à regarder une image est en effet généralement inférieur à celui passé à écouter un morceau et le processus d'annotation requiert davantage de temps lorsqu'il concerne un morceau que lorsqu'il concerne une

1. https://github.com/RedditSota/state-of-the-art-result-for-machine-learning-problems, consulté le 19 Avril 2018.
2. Peter Eckersley et Yomna Nasser, 2017. *EFF AI Progress Measurement Project* sur https://www.eff.org/ai/metrics, consulté le 19 Avril 2018.
3. http://www.businessinsider.fr/us/facebook-350-million-photos-each-day-2013-9/, consulté le 19 Avril 2018.
4. http://www.ifpi.org/downloads/GMR2017.pdf, consulté le 19 Avril 2018.
5. http://press.soundcloud.com/134631-soundcloud-now-has-more-than-135-million-tracks, consulté le 19 Avril 2018.
6. http://www.ifpi.org/downloads/GMR2017.pdf, consulté le 19 Avril 2018.

image. Il est, en effet, plus simple et rapide d'annoter une image avec le mot « voiture » si celle-ci contient une voiture que d'annoter un morceau avec le mot « guitare » si celui-ci contient des sons produits par une guitare.

Un troisième élément limitant le développement des bases de données musicales et leur annotation provient de la capacité humaine à analyser un morceau. Pour la majorité des humains il est, en effet, plus facile de confirmer la présence ou l'absence d'une voiture dans une image que de confirmer la présence ou l'absence d'une guitare dans un morceau. Par exemple, une personne peut affirmer qu'un morceau ne contient pas de guitare alors qu'elle n'a en fait pas reconnu le son provenant d'une guitare électrique saturée. En raison du manque de connaissances musicales, le nombre de personnes pouvant donc contribuer aux annotations des bases de données musicales est plus faible que celui pouvant contribuer aux annotations des bases de données d'images.

Malgré ces limitations, plusieurs bases de données musicales ont été proposées et utilisées. Les principales sont détaillées dans ce chapitre, ainsi que leurs bases d'annotations correspondantes. Néanmoins, l'objectif n'est pas de décrire de manière exhaustive les bases de données musicales existantes en recherche. Les bases de données musicales les plus remarquables sont en effet décrites afin de situer les travaux existants à propos de la classification des thèmes musicaux.

Les sections 2.1 et 2.2 décrivent respectivement des bases de données musicales remarquables et leurs bases d'annotations correspondantes. Ces sections détaillent les avantages et inconvénients de leur mise en place, de leur utilisation ainsi que de leur archivage. La section 2.3 propose ensuite deux bases de données musicales mises en place durant cette thèse grâce à des partenariats industriels et qui tentent de pallier les inconvénients majeurs des bases de données musicales existantes. Chacune des deux bases de données musicales proposées est dédiée à des tâches différentes. La première tâche présentée dans la section 2.3.1 consiste à distinguer les chansons des instrumentaux, c'est-à-dire des morceaux qui ne contiennent pas de chant. L'identification des reprises musicales est la seconde tâche considérée et est exposée dans la section 2.3.2. Pour chacune de ces deux tâches, l'état de l'art des bases de données musicales et des annotations correspondantes est proposé avant d'introduire l'apport des nouvelles bases proposées. La section 2.4 présente des solutions permettant d'utiliser une base qui contient peu de données ou bien des thèmes inégalement répartis. La section 2.5 conclut finalement ce chapitre.

2.1 Introduction sur les bases de données musicales

« Mieux vaut n'être rien que d'exister dans l'ignorance. »
– La main sur l'échiquier, 26 Avril 2017, Ken Troop

Une liste des bases de données musicales est maintenue par Lerch [2012], qui en référence 200 en Avril 2018 [1]. L'objectif de cette section n'est pas de décrire toutes les bases de données existantes mais uniquement les plus remarquables ainsi que les bases d'intérêt pour cette thèse.

Les premières études sur l'extraction d'informations musicales ont été effectuées à partir des bases de données musicales personnelles des chercheurs [Von Schroeter *et al.*, 2000]. Dans ce cadre, l'impact des conclusions obtenues était limité puisque les tailles de bases de données utilisées étaient faibles, que la représentativité musicale pouvait être biaisée par les goûts des chercheurs et que la réplicabilité des expériences était impossible. Afin de permettre la réplicabilité des expériences, Goto *et al.* [2002] ont proposé la base *RWC* [2] qui a permis de mettre les morceaux de cette base à la disposition des chercheurs en échange d'une contribution financière [3]. Malgré le frein financier, plusieurs centaines d'expériences ont pu être menées sur la base de données *RWC* [4].

GTzan [5] est la première base de données musicales accessible gratuitement et a été proposée par Tzanetakis et Cook [2002]. *GTzan* contient 1 000 morceaux répartis équitablement suivant dix thèmes correspondant à des genres différents [6]. *GTzan* constitue la base de données musicales la plus utilisée [Sturm, 2014b] malgré les différents problèmes qui la caractérisent, telles que des annotations incorrectes et la réutilisation d'artistes [Sturm, 2013]. On peut donc s'interroger quant à l'inexactitude d'une partie des annotations et remettre en cause la validité de certains résultats des méthodes de classification proposées utilisant *GTzan*. La réutilisation d'artistes est une raison moins évidente de la faiblesse d'une base de données musicales et s'explique comme suit. Si une base de données musicales contient un thème pour lequel les enregistrements proviennent du même artiste, il n'est alors pas évident de savoir si la méthode de classification détecte le thème ou bien l'artiste. Si l'on consi-

1. http://www.audiocontentanalysis.org/data-sets/, consulté le 19 Avril 2018.
2. https://staff.aist.go.jp/m.goto/RWC-MDB/, consulté le 19 Avril 2018.
3. https://staff.aist.go.jp/m.goto/RWC-MDB/#how_to_use, consulté le 19 Avril 2018.
4. https://scholar.google.fr/scholar?cites=11216225272868032997&as_sdt=2005&sciodt=0,5&hl=fr, consulté le 19 Avril 2018.
5. http://marsyas.info/downloads/datasets.html, consulté le 19 Avril 2018.
6. Blues, Classique, Country, Disco, Hip-Hop, Jazz, Métal, Pop, Reggae et Rock

dère, par exemple, que le genre Blues est représenté par des morceaux de John Lee Hooker, comme c'est le cas dans *GTzan*, on introduit alors un biais de représentativité. Or, ce biais n'est pas pris en compte par la méthode de classification qui risque alors d'associer au thème Blues des caractéristiques propres aux enregistrements de John Lee Hooker (prééminence de la guitare, timbre et tessiture de la voix, ...) sans néanmoins généraliser ces mêmes caractéristiques à tous les morceaux de Blues. Si l'on présente un morceau de Janis Joplin à la même méthode de classification, elle risque donc de ne pas l'identifier comme étant du Blues.
L'ensemble des études qui ont proposé des méthodes d'extraction d'informations musicales utilisant *GTzan* seraient donc des Horses, comme défini dans la section 1.1.4, puisqu'elles ne détectent pas les informations musicales (ici le genre) qu'elles prétendent extraire. L'une des solutions permettant de pallier ce problème consiste à filtrer les morceaux en fonction des artistes [Pampalk *et al.*, 2005; Flexer, 2007; Flexer et Schnitzer, 2009, 2010], c'est-à-dire à limiter la réutilisation d'un même artiste au sein d'un thème. Ce filtre est également applicable à d'autres métadonnées telles que les albums, les studios d'enregistrement et les maisons de disques et ce même pour des bases de données musicales contenant davantage de morceaux que *GTzan* [Flexer et Schnitzer, 2010]. De plus, le biais provenant des albums est plus important que celui provenant des artistes [Flexer et Schnitzer, 2010] puisque les morceaux correspondants peuvent cumuler des caractéristiques des enregistrements par le même artiste, dans un même studio d'enregistrement ainsi que par une même maison de disques.

Afin de filtrer une base de données musicales par rapport à différentes métadonnées, le projet AcousticBrainz [Porter *et al.*, 2015] tente depuis 2014 [1] de rassembler des métadonnées provenant de différentes sources. À ce jour, AcousticBrainz propose la plus grande base de données musicales avec des caractéristiques audio en libre accès [Defferrard *et al.*, 2017] puisqu'elle référence plus de 3 millions de morceaux uniques [2]. AcouticBrainz ne propose cependant pas les 3 millions de morceaux en téléchargement, les chercheurs doivent donc trouver eux-mêmes les fichiers audio correspondant à ces références. Pour pallier ce problème, AcousticBrainz met directement à disposition plusieurs caractéristiques audio pour chaque morceau. Bien que ces caractéristiques audio constituent une source inégalée en nombre, elles sont pré-calculées et il est donc impossible de développer et d'extraire de nouvelles caractéristiques audio en tirant profit des 3 millions de morceaux existants. De nouvelles caractéristiques audio ne seraient, en effet, extraites qu'à partir de nouveaux morceaux. Ces caractéristiques audio existantes restreignent par conséquent les choix possibles d'extraction d'informations musicales et limitent donc les

1. https://beta.acousticbrainz.org/statistics-graph, consulté le 19 Avril 2018.
2. https://acousticbrainz.org/, consulté le 19 Avril 2018.

types d'analyses concevables. De plus, les caractéristiques audio proposées sont calculées pour l'ensemble d'un morceau et non à une échelle plus petite telle qu'un ensemble d'échantillons audio. Il n'est donc pas possible, par exemple, d'étudier la présence d'un instrument à la seconde près. La mise à disposition de caractéristiques audio à une échelle plus réduite est par ailleurs restreinte puisque certaines caractéristiques permettent de reconstruire le signal audio [Sturmel et Daudet, 2011; Průša et Rajmic, 2017; Oyamada *et al.*, 2018] [1]. La reconstruction du signal audio constitue évidemment une fuite de données potentielle enfreignant directement les droits des auteurs, comme c'est le cas pour le téléchargement illégal. Les caractéristiques audio actuellement proposées par AcousticBrainz permettent donc uniquement de comparer les différences de performances des méthodes de classification et non les caractéristiques audio qu'elles utilisent.

L'accès à une base de données musicales proposant les fichiers audio des enregistrements permet à des chercheurs de développer de nouvelles caractéristiques audio. Defferrard *et al.* [2017] ont proposé la Free Music Archive (FMA), qui est à ce jour la base de données musicales offrant le plus grand nombre d'enregistrements. FMA propose l'intégralité des 106 574 morceaux qu'elle contient contrairement à *GTzan* qui ne propose que les 30 premières secondes de ses 1 000 morceaux. FMA propose 28 fois moins de morceaux qu'AcousticBrainz, cependant sa taille indique l'ordre de grandeur des bases de données musicales de recherche actuellement disponibles. Cette taille restreinte s'explique par la difficulté à collecter des morceaux librement utilisables dans un cadre de recherche.

Les bases de données musicales présentées précédemment référencent des enregistrements et leurs genres musicaux associés. L'identification automatique des genres musicaux est en effet le thème le plus étudié dans l'extraction d'informations musicales [Fu *et al.*, 2011; Sturm, 2014b]. Toutefois, l'analyse des genres musicaux est complexe de par la nature subjective de ceux-ci [Craft *et al.*, 2007] et les résultats des méthodes d'analyse du genre musical doivent être considérés avec précaution [Sturm, 2013]. Un exemple récent de méthode proche de l'état de l'art identifiant le genre musical en temps réel a été proposé par DeepSound [2] et indique par exemple que le titre *I Want To Break Free* de Queen peut être caractérisé par des genres variant du Classique au Country en passant par le Reggae. Le genre musical perçu varie en fonction du niveau des connaissances musicales, des goûts, de l'humeur de la personne qui annote le morceau mais également de son âge [Berenzweig *et al.*, 2004]. La

1. http://anclab.org/software/phaserecon/, consulté le 19 Avril 2018.
2. http://deepsound.io/genres/play.html?utm_content=buffer9c949&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer#break-free, consulté le 19 Avril 2018.

présence de chant est ici considérée comme une annotation plus objective que la reconnaissance d'un genre ou d'une émotion. Dans ce cadre, la section 2.3.1 décrit une base de données musicales dédiée à la détection des chansons et des instrumentaux. Avant de présenter cette base, la section qui suit détaille les avantages et inconvénients de la mise en place, de l'utilisation, du partage et de l'archivage des bases d'annotations qui permettent de développer des méthodes d'extraction d'informations musicales.

2.2 Mise en place et caractéristiques des annotations

> *« [...] for every feeling there's a song. »*
> – *Love Is Like A Violin*, 1977, Barclay James Harvest

Afin de constituer une liste de lecture musicale relative à un thème, il est nécessaire d'y inclure des morceaux qui ont été correctement annotés comme appartenant à ce thème [Jäschke *et al.*, 2007]. Ces annotations permettent aux méthodes de classification d'apprendre quelles sont les caractéristiques sous-jacentes d'un thème donné et de pouvoir identifier si un élément inconnu appartient à ce thème. Le principe des annotations musicales est présenté dans la section suivante.

2.2.1 Annotations musicales

> *« Je ne joue pas du Jazz. Le Jazz, c'est la musique que les Noirs jouent pour faire plaisir aux Blancs et gagner leur vie. Ma musique mérite plus de respect que ça. Elle vaut bien mieux que cette étiquette. C'est de la grande musique, pour tout le monde. »*
> – Interview à l'Hôtel Méridien à Paris, Printemps 1987, Miles Davis

Lors de la recherche de listes de lecture musicale, les émotions, sentiments, genres et activités constituent les catégories les plus utilisées [Inskip *et al.*, 2012; Tintarev *et al.*, 2017] car elles contiennent les thèmes (*e.g.* Joyeux, Rock, Sport) les plus courants et intuitifs à exprimer [Aljanaki, 2016]. Les thèmes correspondants constituent donc les annotations à rechercher et à privilégier dans la description des morceaux. De plus, certaines émotions musicales semblent être universelles puisque des auditeurs occidentaux peuvent détecter la joie, la tristesse et la colère dans des Ragas hindoustanis [Balkwill et Thompson, 1999]. De même, plusieurs études ont montré que de telles émotions étaient perçues par des Japonais lors de l'écoute de morceaux japonais, occidentaux et hindoustanis [Balkwill *et al.*, 2004; Fritz *et al.*, 2009]. Ces derniers soulignent par exemple que la tristesse est principalement exprimée par une diminution

du volume et du tempo alors que la colère l'est par leur augmentation. La musique transmet de plus davantage de nuances dans les émotions dont les perceptions sont spécifiques à chaque culture. Ces émotions comprennent notamment la tendresse, la solennité, le triomphe et l'humour, pour lesquels les relations acoustiques sont plus complexes à identifier [Gabrielsson et Lindström, 2001; Meyers, 2007; Aljanaki, 2016]. Ce fossé sémantique provient d'un premier décalage entre l'émotion induite par un morceau et celle perçue par l'auditeur. Un second décalage provient de la différence entre une émotion, qui est éphémère, et un sentiment, qui persiste plus longuement dans le temps. Les émotions et les sentiments ressentis lors de l'écoute d'un morceau dépendent de plus du passé de chacun et évoluent au cours du temps.

La génération automatique de listes de lecture musicale relatives à des thématiques émotionnelles est complexe puisque les émotions musicales sont subjectives [Li et Ogihara, 2003]. Deux tiers des annotateurs des bases de données musicales relatives à l'identification des émotions n'ont par exemple pas trouvé de compromis pour plusieurs morceaux [Hu *et al.*, 2008]. L'annotation d'un morceau avec une émotion induite ou perçue est donc plus complexe, plus subjective et moins bien définie que pour le genre [Aljanaki, 2016]. Une solution consiste alors à représenter une émotion musicale selon des modèles catégoriels, dimensionnels ou spécifiques à un domaine plutôt que par sa seule présence. Les annotations de la présence d'une émotion peuvent, par ailleurs, être réparties sur plusieurs millisecondes à plusieurs minutes [Lu *et al.*, 2006; Panda et Paiva, 2011; Aljanaki, 2016] sur un même morceau. Un morceau peut enfin susciter une ou plusieurs émotions simultanément qu'il est possible de quantifier de manière binaire ou continue [Sanden et Zhang, 2011].

Lors du traitement de thèmes perçus comme plus objectifs qu'une émotion donnée, tels que la présence d'un instrument, l'annotation d'un morceau avec les instants exacts auxquels un instrument est joué nécessite l'utilisation d'outils à haute précision temporelle. Le logiciel Tony [Mauch *et al.*, 2015] permet par exemple d'indiquer la présence d'un instrument donné à la trame près, c'est-à-dire pour quelques échantillons audio. Il est toutefois à noter que dans le cas des thèmes objectifs, des cas particuliers peuvent également laisser un doute quant à l'annotation adéquate d'un morceau. Il est par exemple possible de définir une chanson comme étant un morceau contenant des paroles. Cette définition exclut néanmoins les fredonnements et onomatopées que contient le morceau *Crowd Chant* de Joe Satriani par exemple. Il est alors possible de définir une chanson comme contenant des sons produits par la bouche. Dans un tel cas, comment traiter un morceau contenant l'effet appelé *talking box* qui vocalise un son de guitare en le modulant par le volume buccal, comme cela est le cas dans *Keep Talking* de Pink Floyd [1] ? De même, on peut se poser la question de la représentativité d'un mot chanté pendant moins d'une seconde dans

1. Vidéo de l'instant où une *talking box* est utilisée : https://youtu.be/4DmsvdCPD2Y?t=2m11s, consulté le 19 Avril 2018.

un morceau de musique instrumentale de plusieurs minutes et de l'annotation à attribuer à ce morceau. Par ailleurs, les annotations humaines de la présence du chant à l'échelle de la trame sont chronophages [Skowronek *et al.*, 2006] et sujettes aux erreurs [Sturm, 2013, 2015] mais également au manque d'exhaustivité. La base *AudioSet* [Gemmeke *et al.*, 2017] contient par exemple l'un de ces défauts puisque, pour un morceau donné, elle affiche comme seule annotation la présence du chant d'une femme [1] alors qu'un homme chante également. Le manque d'annotations exhaustives complexifie donc le développement d'une méthode utilisant *AudioSet* afin de distinguer les chants féminin et masculin.

Les exemples présentés ci-dessus soulignent que les difficultés inhérentes à la constitution d'une base de données musicales sont également dues à la constitution des annotations. Aux difficultés d'accès à de nouveaux morceaux s'ajoute donc leur complexité d'annotation. La section suivante décrit les méthodes mises en place afin de réaliser de telles annotations musicales dans un cadre académique.

2.2.2 Processus d'annotation audio

> *« [...] mais il faut distinguer entre les chansons Folk et les chansons de* type *Folk. Comme le Jazz, et les musiques qui utilisent le langage du Jazz... »*
>
> – *Pacific Park*, 1987, Philip K. Dick

Turnbull *et al.* [2008] distinguent cinq méthodes permettant de collecter des annotations et qui sont listées ci-dessous :

- Les sites web sociaux musicaux [Shardanand et Maes, 1995; Breese *et al.*, 1998; Levy et Sandler, 2007; Shepitsen *et al.*, 2008] représentés notamment par Last.fm [2],
- Les jeux d'annotation musicaux [Law *et al.*, 2007; Turnbull *et al.*, 2007; Mandel et Ellis, 2008],
- Les sondages en ligne [Turnbull *et al.*, 2008],
- L'extraction de métadonnées présentes sur internet [Whitman et Ellis, 2004; Knees *et al.*, 2007],
- L'analyse du signal audio [Tzanetakis et Cook, 2002; Bertin-Mahieux *et al.*, 2010; Prockup *et al.*, 2015].

Il est également possible de combiner ces méthodes [Bu *et al.*, 2010] afin d'annoter des morceaux. Toutefois, les performances de cette combinaison sont limitées par les performances individuelles de chacune des méthodes sources.

1. https://www.youtube.com/watch?v=3w-AEo6rgI8, consulté le 19 Avril 2018.
2. https://www.last.fm, consulté le 19 Avril 2018.

L'extraction de métadonnées présentes sur internet est néanmoins fastidieuse et faillible puisqu'elle implique de trier une large quantité de données redondantes, contradictoires et sémantiques provenant de sources disparates. Les jeux d'annotation musicaux ne passent par ailleurs pas à l'échelle et, de même que les sites web sociaux musicaux, sont enclins aux annotations contradictoires pour deux raisons principales. Tout d'abord, un utilisateur peut appliquer plusieurs annotations, même incorrectes, à un même morceau afin que celui-ci remonte dans les résultats de recherche. Une seconde raison provient de la différence perceptive de l'écoute d'un morceau par différentes personnes de par leurs différences, générationnelle ou démographique par exemples.

La recherche d'un thème en utilisant l'analyse du signal audio ne présente pas les principaux défauts des quatre autres méthodes précédentes puisqu'elle peut notamment se révéler plus rapide et mieux supporter le passage à l'échelle [Logan, 2002; Hoashi *et al.*, 2003; Celma *et al.*, 2005; Eck *et al.*, 2007; Sordo *et al.*, 2007; Turnbull *et al.*, 2007; Mandel et Ellis, 2008; Tingle *et al.*, 2010]. Le chapitre 3 détaille plus largement la méthode de recherche d'un thème à partir de l'analyse du signal audio. La section suivante décrit quant à elle les bases de données musicales proposées afin de répondre aux problèmes des bases existantes.

2.3 Bases de données musicales proposées

« De la musique avant toute chose [...] »
– Art poétique, 1874, Paul Verlaine

La faible quantité de morceaux et le manque de représentativité des bases de données musicales académiques actuelles ne favorisent pas un contexte de développement de méthodes d'extraction d'informations performantes. Afin de pallier ce manque, des partenariats ont été mis en place avec des entreprises afin d'avoir accès à davantage de données adéquates pour cette tâche. Cette section décrit les partenariats industriels mis en place au cours de cette thèse afin de proposer deux nouvelles bases de données musicales ainsi que les annotations correspondantes.

2.3.1 SATIN et SOFT1

« The recording industry is a talent-driven, creative industry, and as such it is totally dependent on copyright protection. Copyright is the essential building block of the music business, allowing artists, song writers and record companies to invest their revenues and their livelihoods in the creative process, secure in the knowledge that they, and no one else, will own the result. Copyright is the incentive to be creative. It protects artists from piracy of their works and it nurtures new talent. »
– Rapport de l'IFPI, 2001

Travaux afférents

Les sections 2.1 et 2.2 décrivent les difficultés inhérentes à la constitution d'une base de données musicales. Afin de réduire l'impact de ces inconvénients, une base est proposée et l'application considérée implique la distinction entre les chansons et les instrumentaux. L'annotation chanson/instrumental a été choisie en particulier car elle fournit pour chaque morceau une indication relativement objective, mutuellement exclusive et toujours pertinente [Gouyon *et al.*, 2014] puisqu'un morceau est soit une chanson, soit un instrumental. Hormis les bases de données musicales personnelles et non détaillées par leurs auteurs [Ghosal *et al.*, 2013], huit bases de données musicales ont déjà été utilisées pour le développement de méthodes de détection des chansons et des instrumentaux. Ces huit bases de données musicales sont détaillées ci-après.

- *CAL500* [Turnbull *et al.*, 2008] contient 502 morceaux de 502 artistes uniques. Les morceaux ont été annotés grâce à un sondage. Chaque morceau a reçu des annotations d'au moins trois personnes différentes.
- *ccMixter* [Liutkus *et al.*, 2014] contient 50 chansons librement utilisables à des fins de recherche et qui ont été collectées sur le site de l'entreprise ccMixter [1]. Pour chaque morceau, la piste instrumentale est fournie indépendamment de la piste vocale.
- *Jamendo* [2] [Ramona *et al.*, 2008] contient 93 chansons ainsi que les annotations de la présence de chant à l'échelle de la trame. Cette base est proposée en libre accès et les chansons proviennent de l'entreprise Jamendo Music [3].
- *MagTag5k* [Marques *et al.*, 2011] est une version restreinte de la base de données musicales *Magnatagatune* de Law *et al.* [2009]. *MagTag5k* contient 2 349 extraits de morceaux de 230 artistes différents.
- *MedleyDB* [4] contient toutes les pistes enregistrées pour chacun des 163 morceaux proposés par Bittner *et al.* [2014] en libre accès.
- *MSD24k* est une version restreinte de la base de données musicales Million Song Dataset (MSD) [Bertin-Mahieux *et al.*, 2011] ne proposant pas les morceaux mais les caractéristiques audio extraites. *MSD24k* contient 1 677 morceaux.
- *QUASI* [5] contient 11 morceaux enregistrés et annotés par l'équipe INRIA/IRISA Metiss. *QUASI* contient des annotations de présence d'instruments à l'échelle de la trame.
- *RWC POP* contient 100 chansons rassemblées et annotées par Goto *et al.* [2002] et est accessible de manière payante.

Le tableau 2.1 détaille la répartition des chansons et des instrumentaux de chacune de ces huit bases de données musicales. En considérant le déséquilibre qui existe entre le nombre de chansons et d'instrumentaux présenté dans le tableau 2.1, il est possible d'améliorer les résultats issus des méthodes de classification automatique grâce à des outils qui sont introduits dans la section 2.4.2. Par ailleurs, les huit bases de données musicales présentées dans le tableau 2.1 totalisent 5 003 morceaux, ce qui représente un nombre limité d'éléments et soulève des questions quant à la pertinence des résultats des

1. http://www.ccmixter.org/, consulté le 19 Avril 2018.
2. http://www.mathieuramona.com/wp/data/jamendo, consulté le 19 Avril 2018.
3. https://www.jamendo.com, consulté le 19 Avril 2018.
4. http://medleydb.weebly.com, consulté le 19 Avril 2018.
5. http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/, consulté le 19 Avril 2018.

Tableau 2.1 – Répartition du nombre de chansons (#C) et d'instrumentaux (#I) dans huit bases de données musicales utilisées pour la détection de présence de chant dans les morceaux.

Nom	#C	#I	Total	Utilisation par
CAL500	444	58	502	[Hespanhol, 2013; Gouyon *et al.*, 2014]
CCMixter	50	50	100	[Bayle *et al.*, 2016]
Jamendo	0	93	93	[Bayle *et al.*, 2016]
MagTag5k	1 626	723	2 349	[Hespanhol, 2013; Gouyon *et al.*, 2014]
MeydleyDB	63	100	163	[Bayle *et al.*, 2016]
MSD24k	1 146	531	1 677	[Hespanhol, 2013; Gouyon *et al.*, 2014]
QUASI	9	10	19	[Bayle *et al.*, 2016]
RWCPOP	0	100	100	[Bayle *et al.*, 2016]
Total	3 338	1 665	5 003	

méthodes appliquées à ces morceaux. Il faut également questionner la représentativité musicale de ces 5 003 morceaux face aux 43 millions de morceaux disponibles sur Deezer [1] en 2018. Ce constat et ces questionnements ont amené à la constitution d'une base de données musicales plus conséquente, pour laquelle il a fallu mettre en place des partenariats industriels.

Concept clé de la base de données musicales proposée

Afin de proposer un nouveau cadre de développement et de comparaison d'algorithmes d'analyse musicale, la base de données musicales Set of Audio Tags and Identifiers Normalized (SATIN) [Bayle *et al.*, 2017a] a été proposée et assure également la cohérence et la persistance des morceaux référencés. La mise en place de SATIN a mis à profit les outils utilisés par l'industrie musicale afin de pérenniser le référencement des morceaux contenus dans cette base. L'idée principale consiste à utiliser un identifiant unique et propre à chaque enregistrement audio qui est l'International Standard Recording Code (ISRC) [2]. Celui-ci est fourni par l'International Federation of the Phonographic Industry (IFPI) et permet d'identifier chaque morceau mais surtout de distinguer les différentes versions et enregistrements d'une même œuvre musicale. Dans ce cadre, la version originale du titre *Hotel California* du groupe The Eagles a un ISRC différent –USEE19900323 [3]– de celui de la version remastérisée inclue dans une compilation des morceaux du groupe –USEE19900349 [4]–, de même que la version enregistrée dans un concert donné a son propre ISRC. Le mécanisme d'ISRC permet en effet depuis 1986, date à laquelle a été éta-

1. https://www.deezer.com/fr/company/about, consulté le 19 Avril 2018.
2. http://isrc.ifpi.org/en, consulté le 19 Avril 2018.
3. https://musicbrainz.org/isrc/USEE19900323, consulté le 19 Avril 2018.
4. https://musicbrainz.org/isrc/USEE19900349, consulté le 19 Avril 2018.

blie la norme ISO 3901 correspondante, d'identifier les morceaux diffusés et de rémunérer [1] toutes les personnes concernées par l'élaboration d'une œuvre donnée, depuis les compositeurs jusqu'aux diffuseurs. En fournissant l'ISRC de chaque morceau utilisé dans une expérience scientifique, il devient donc possible d'améliorer la réplicabilité des expériences et la comparaison entre différentes méthodes d'analyse musicale. En effet, une méthode peut fournir pour chaque morceau la probabilité que ce dernier présente un thème donné et comparer les probabilités indiquées par différentes méthodes.

Afin d'obtenir un nombre conséquent de chansons et d'instrumentaux ainsi que leurs annotations, un partenariat a été instauré avec l'entreprise Musixmatch [2]. En 2018, cette entreprise possède un catalogue de 50 millions de morceaux. Elle propose les paroles pour 14 millions de ces chansons [3] et référence plusieurs milliers d'instrumentaux. L'annotation d'un morceau en tant que chanson ou instrumental est disponible sur l'Application Programming Interface (API) [4] de Musixmatch. Les millions [5] de morceaux accessibles grâce à l'API de Musixmatch disposent par ailleurs d'un identifiant interne qui n'a pas de correspondance avec le code ISRC. Toutefois, Musixmatch propose le MusicBrainz IDentifier (MBID) [6] pour certains morceaux, qui est un identifiant utilisé par MusicBrainz [7] et qui lui-même lie certains MBID avec des ISRCs [8]. Après cette collecte d'une base d'annotations chanson/instrumental liées à des ISRCs, deux partenariats ont été instaurés avec Deezer [9] et Simbals [10], qui sont deux entreprises disposant de fichiers audio référencés par leurs ISRCs. Ces partenariats avaient pour objectif de permettre l'extraction de caractéristiques audio.

Les problèmes de correspondance entre les différents identifiants musicaux ont limité la quantité de morceaux référencés. Malgré le nombre de morceaux initialement proposés par Musixmatch, AcousticBrainz, Deezer et Simbals – qui dépasse plusieurs dizaines de millions–, il n'a finalement été possible de référencer correctement que 408 380 morceaux, dont 106 466 sont annotés comme chansons et 13 944 comme instrumentaux. Néanmoins, la base ainsi constituée est la plus conséquente de l'état de l'art dans l'analyse des chansons et des instrumentaux puisqu'elle possède 81 fois plus de morceaux que toutes

1. https://www.theglobeandmail.com/report-on-business/ai-technology-may-help-recording-music-industry-recoup-millions-in-royalties-for-musicians/article37470203/, consulté le 19 Avril 2018.
2. https://www.musixmatch.com/fr, consulté le 19 Avril 2018.
3. https://about.musixmatch.com/press/, consulté le 19 Avril 2018.
4. https://developer.musixmatch.com/, consulté le 19 Avril 2018.
5. https://about.musixmatch.com/, consulté le 19 Avril 2018.
6. https://musicbrainz.org/doc/MBID, consulté le 19 Avril 2018.
7. https://musicbrainz.org, consulté le 19 Avril 2018.
8. https://musicbrainz.org/doc/ISRC, consulté le 19 Avril 2018.
9. http://www.deezer.com, consulté le 19 Avril 2018.
10. http://www.simbals.com/, consulté le 19 Avril 2018.

les bases réunies et précédemment citées dans le tableau 2.1.

La base de données musicales ainsi constituée réunit plusieurs identifiants normalisés mondiaux et industriels mais également des annotations, c'est pourquoi le nom choisi pour se référer aux données est Set of Audio Tags and Identifiers Normalized (SATIN) et que celui choisi pour les annotations correspondantes est first Set Of FeaTures (SOFT1). Les résultats des méthodes utilisant SATIN sont référencés par l'Electronic Frontier Foundation (EFF) afin de mesurer les progrès des méthodes d'intelligence artificielle [1].

SATIN est proposée sous la forme d'un fichier Comma Separated Value (CSV) et référence 408 380 morceaux. SATIN contient plusieurs identifiants pour chaque morceau, qui proviennent de différentes sources et qui comprennent :

- L'ISRC de l'IFPI, utilisé notamment par Deezer, Spotify, Musixmatch et MusicBrainz,
- Le MBID de MusicBrainz,
- L'identifiant de Musixmatch,
- L'identifiant de SoundCloud,
- L'identifiant de Spotify,
- L'identifiant de XBoxMusic,
- L'identifiant de Commontrack.

Annotations fournies avec SATIN

SATIN renseigne l'année d'enregistrement des morceaux. Davantage d'informations sont également accessibles sur les API web de Deezer, de Spotify et de MusicBrainz grâce aux codes ISRC fournis.

SATIN propose également des liens vers les paroles des chansons et indique si ces paroles contiennent du vocabulaire explicite.

SATIN fournit de plus un ou plusieurs genres en tant qu'annotations pour certains des morceaux.

Parmi les 120 410 morceaux de SATIN annotés quant à la présence de chant, 106 466 sont des chansons et 13 944, soit 11,58%, sont des instrumentaux. L'existence de ce déséquilibre aurait pu provenir de l'utilisation de l'API de Musixmatch. L'objectif de Musixmatch est de vendre des paroles, cette entreprise n'a donc pas besoin d'informations concernant les instrumentaux. Il s'est néanmoins avéré que cette proportion observée d'instrumentaux était proche de celle de Deezer [2] et ne constituait donc pas une erreur de représentativité de la base. Malgré ce déséquilibre, la quantité d'annotations de SATIN

1. https://www.eff.org/ai/metrics#Instrumentals-recognition, consulté le 19 Avril 2018.
2. Communication personnelle de Manuel Moussallam, scientifique principal de l'équipe de recherche et développement de Deezer

en ce qui concerne les chansons et instrumentaux en fait la plus grande base de données musicales pour l'étude de ces annotations.

Description de SOFT1

Le premier jeu de caractéristiques extraites en utilisant les identifiants contenus dans SATIN est proposé dans SOFT1. Les caractéristiques audio ont été extraites par trois logiciels différents qui sont Essentia [Bogdanov *et al.*, 2013], Marsyas [Tzanetakis et Cook, 2000] et YAAFE [Mathieu *et al.*, 2010]. Toutes les caractéristiques audio proposées à l'extraction par le logiciel Essentia ont été calculées. Marsyas et YAAFE ont permis d'extraire les caractéristiques audio définies par Gouyon *et al.* [2014].

Caractéristiques des morceaux référencés dans SATIN

Le nuage de mots des genres relatifs aux morceaux contenus dans SATIN est visible dans la figure 2.1. Les genres incluent principalement de la musique Pop, du Rock, du Rap et du Hip Hop, ce qui est en accord avec les genres les plus écoutés [1] sur les plateformes de streaming.

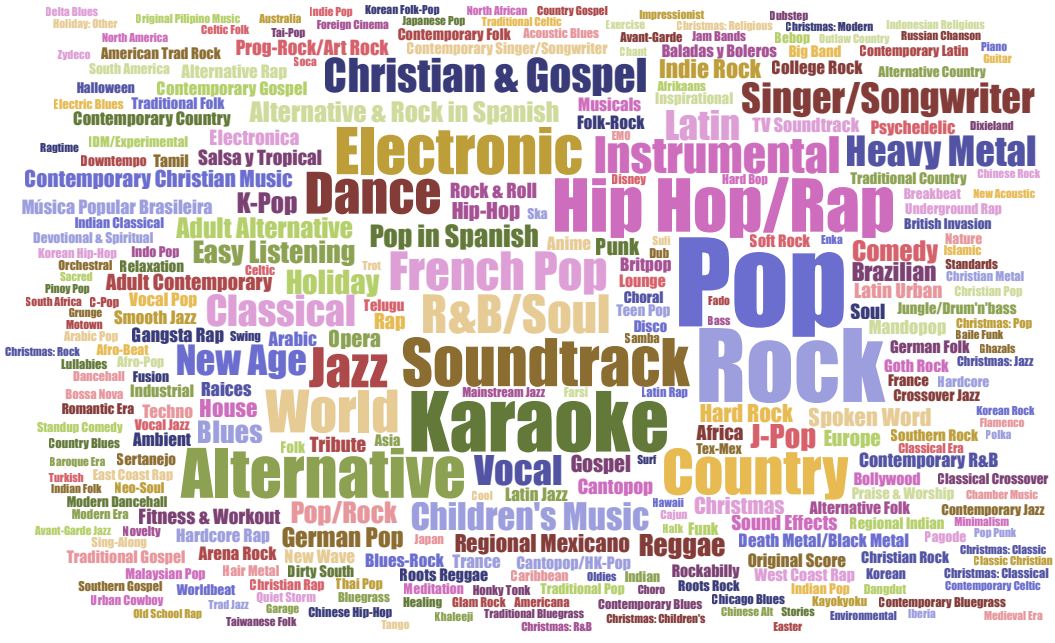


FIGURE 2.1 – Nuage de mots représentant les genres relatifs aux morceaux contenus dans SATIN.

1. https://www.digitalmusicnews.com/2016/03/01/the-most-popular-music-genre-in-every-u-s-state/, consulté le 19 Avril 2018.

La figure 2.2 représente le nombre de morceaux contenus dans SATIN en fonction de leur année d'enregistrement. La date d'enregistrement a pu être récupérée pour 311 141 morceaux en utilisant l'API de Deezer [1].

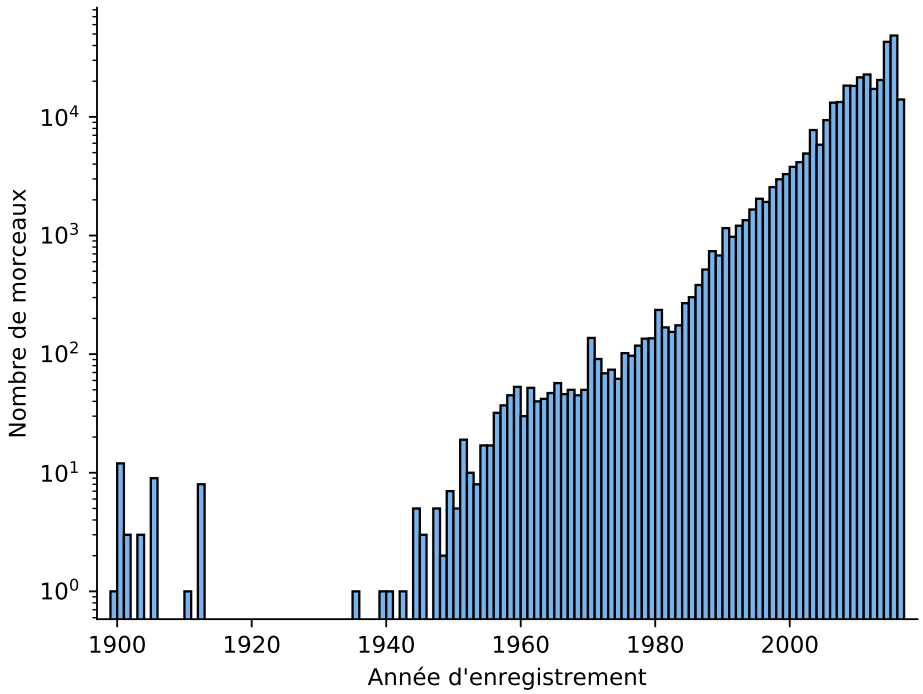


FIGURE 2.2 – Nombre de morceaux de SATIN par année d'enregistrement au registre de l'IFPI en fonction des données fournies par Deezer.

La répartition géographique –par pays– des morceaux contenus dans SATIN en fonction de leur ISRC est visible dans la figure 2.3. La plupart des morceaux ont été enregistrés aux États-Unis d'Amérique, en France, au Royaume-Uni, en Allemagne et en Inde. Bien que le pays d'enregistrement fournit une indication quant au lieu de déclaration de l'existence du morceau à l'IFPI, l'artiste correspondant peut être originaire d'un autre pays.

1. https://developers.deezer.com/api, consulté le 19 Avril 2018.

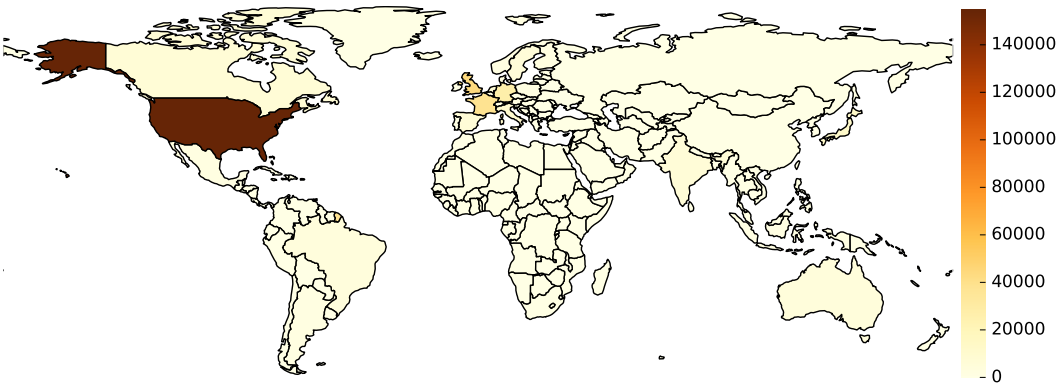


FIGURE 2.3 – Répartition géographique par pays des morceaux contenus dans SATIN, en fonction de leur ISRC.

Code source

Afin de manipuler les données contenues dans SATIN, une API en Python hébergée sur GitHub [1] a été proposée. Les fonctions actuellement disponibles permettent :

- De générer des graphiques, sur la base de données ou d'une sous-partie de celle-ci, tel que représenté dans les figures 2.1, 2.2 et 2.3,
- De comparer les résultats de classification de deux algorithmes,
- D'accéder aux paroles d'une chanson,
- D'accéder aux métadonnées industrielles d'un morceau, telles que le nom d'artiste et d'album grâce à l'API web de Deezer [2], à celle de Spotify [3] et à celle de MusicBrainz [4].

Conclusion

L'aspect principal du projet autour de cette base de données musicales est donc la fourniture d'une composante permettant de faciliter la réplicabilité des expériences et le passage à l'échelle d'analyses musicales. La contribution majeure consiste, en effet, en l'utilisation d'ISRCs en tant qu'identifiants pérennes des morceaux et en la proposition d'une base de données conséquente d'ISRCs ainsi que de la base d'annotations correspondante pour plusieurs thèmes musicaux. La réplicabilité des expériences scientifiques proposées sur cette base est de plus possible grâce à la présence de la liste d'ISRCs et du code source de l'API en Python sur le répertoire GitHub précédemment mentionné.

1. https://github.com/ybayle/SATIN, consulté le 19 Avril 2018.
2. https://developers.deezer.com/api, consulté le 19 Avril 2018.
3. https://developer.spotify.com/web-api/, consulté le 19 Avril 2018.
4. https://musicbrainz.org/doc/Development/XML_Web_Service/Version_2, consulté le 19 Avril 2018.

Perspectives

Afin d'obtenir une base d'annotations plus conséquente, des outils permettant de lier davantage d'ISRCs à d'autres identifiants pourraient être développés. De tels outils permettraient d'ajouter à SATIN l'intégralité des sept millions d'annotations de chansons référencées par Musixmatch. Toutefois, cette augmentation n'est actuellement pas réalisable puisque Musixmatch ne dispose pas des ISRCs pour ces morceaux. Il serait donc possible pour cela de faire correspondre les titres, albums et artistes de chaque morceau de Musixmatch avec ceux des ISRCs. La réalisation de cette correspondance se heurte néanmoins à des problèmes de consistance liés à l'écriture des noms de morceaux (Hide Away (feat. Holly), Hide Away feat. Holly ou Hide Away featuring Holly), d'albums (Reistu þig við, sólin er komin á loft... ou Reistu Big Vio Solin Er Komin A Loft...) et d'artistes (AC/DC, AC-DC ou AC DC). Ce manque de consistance pose problème lors de la tentative de mise en correspondance de plusieurs identifiants musicaux et la méthode de Barone *et al.* [2015, 2016] propose une première approche pour pallier ce problème. Leur méthode consiste en effet à unifier des identifiants différents pour des noms de morceaux, d'albums et d'artistes possédant des écritures différentes ou bien des versions d'enregistrement différentes (Album, Concert, Acoustique, ...). Il reste toutefois à évaluer la qualité des correspondances trouvées par leur méthode.

Une autre solution permettant de lier des identifiants de morceaux provenant de différents services consisterait à utiliser le système de reconnaissance par empreinte acoustique, ou *fingerprinting* [Kim *et al.*, 2008]. Cette méthode n'utilise pas de métadonnées puisqu'elle analyse le signal audio afin d'en extraire une empreinte acoustique qui permet de l'identifier de manière unique. Cette méthode est utilisée par Shazam [1] afin d'identifier un morceau à partir de quelques secondes d'enregistrement sur un téléphone portable. L'un des avantages de cette méthode réside dans le fait qu'elle n'utilise pas de métadonnées, qui peuvent se révéler erronées. L'inconvénient principal de cette méthode provient en revanche du temps qui lui est nécessaire pour calculer une empreinte acoustique pour chaque morceau. Ce temps de calcul est en effet supérieur au temps passé à comparer des chaînes de caractères présentes dans les métadonnées. Il paraît toutefois pertinent d'évaluer la précision des liens obtenus par un système de reconnaissance par empreinte acoustique sur deux grandes bases de données musicales.

1. https://www.shazam.com, consulté le 19 Avril 2018.

2.3.2 Kara1K

> *« Does asking oneself these questions in an attempt to see how the machine works spoil the enjoyment ? It hasn't for me. Music isn't fragile. »*
>
> – *How Music Works*, 2012, David Byrne

Cette section décrit la seconde base de données musicales constituée afin de résoudre certains problèmes des bases existantes. Les morceaux proposés dans cette base sont des reprises, c'est-à-dire des enregistrements de versions alternatives de morceaux par un interprète différent de celui de la version originale. Cette base de données musicales permet de développer des méthodes d'identification de ces reprises. L'identification des reprises permet par exemple à l'industrie musicale de détecter les plagiats, de comptabiliser les écoutes et de collecter les droits d'auteur mais permet également aux auditeurs de trouver des variations des morceaux qu'ils apprécient.

Travaux afférents

En 2006, le Music Information Retrieval Evaluation eXchange (MIREX) [1] a mis en place une campagne annuelle d'évaluation et de comparaison des méthodes d'identification des reprises. La base de données musicales alors utilisée par le MIREX n'était pas disponible en libre accès. Les meilleurs résultats obtenus par la campagne d'évaluation du MIREX ont été atteints par Serrà *et al.* [2009] qui ont correctement identifié 2 426 reprises sur 3 300, soit 73,5% d'entre elles. Depuis 2009, aucun progrès n'a été réalisé dans les résultats affichés par les méthodes d'identification des reprises.

La première base librement téléchargeable a été proposée par Ellis et Cotton [2007] et s'intitule *covers80*. Cette base contient 80 morceaux de Pop américains qui ont chacun été enregistrés par deux artistes différents avec une fréquence d'échantillonnage de 16 000 Hz sur une piste monophonique.

Second Hand Songs (SHS), proposée par Bertin-Mahieux *et al.* [2011], est la plus grande base dédiée à l'identification des reprises puisqu'elle contient 18 196 morceaux enregistrés par des professionnels. Cette base référence des reprises dans plusieurs genres, provenant de plusieurs pays et dont la durée est variable. Toutefois, SHS ne propose pas les morceaux en libre accès mais des caractéristiques audio qui en sont extraites. Il n'est donc pas possible de développer de nouvelles méthodes d'extraction de caractéristiques audio à partir de SHS.

Trois autres bases de données musicales ont été utilisées pour l'identification des reprises mais n'ont pas été initialement conçues pour cette tâche.

1. http://www.music-ir.org/mirex/wiki/2006:Audio_Cover_Song, consulté le 19 Avril 2018.

La base de données musicales MIR-1K a été proposée par Hsu et Jang [2010] et contient 1 000 extraits de 4 à 13 secondes de reprises chinoises populaires différentes chantées par des amateurs. La base de données musicales iKala, proposée par Chan *et al.* [2015], contient 352 extraits de 30 secondes de reprises chinoises populaires différentes chantées par des professionnels. La base de données musicales Digital Archive of Mobile Performances (DAMP)[1] proposée par Smith [2013] contient 34 000 chansons enregistrées par des amateurs sur leur téléphone portable grâce à l'application *The Sing!*[2] de Smule[3]. La partie instrumentale est la même pour toutes les chansons proposées par Smule et la qualité des enregistrements des fichiers audio amateurs est limitée.

Depuis sept ans, les études concernant l'identification des reprises se sont concentrées sur le passage à l'échelle plutôt que sur l'amélioration de la proportion d'identifications correctes [Bertin-Mahieux *et al.*, 2011; Osmalskyj *et al.*, 2013].

Jusqu'à présent, les études concernant l'identification des reprises ne sont parvenues à aucun consensus quant à la définition claire d'une reprise et des différentes modalités pouvant être modifiées. Pourtant, une meilleure définition des types de reprises permet de développer des méthodes d'identification appropriées à chaque modalité. Une reprise peut, en effet, se différencier du morceau original suivant plusieurs aspects tels que :

- La modification des paroles (*e.g.* la chanson populaire *House Of The Rising Sun* et la reprise *Le Pénitencier* par Johnny Hallyday),
- L'allongement de la durée du morceau et la suppression de paroles (*e.g.* *Little Wing* de Jimi Hendrix et la reprise par Stevie Ray Vaughan),
- Le changement de mode, de mineur à majeur (*e.g.* *Smells Like Teen Spirits* de Nirvana et la reprise *Teen Sprite* par Nirvirna[4]),
- Le changement de genre de Pop à Métal (*e.g.* *Face À La Mer* de Calogero et Passi et la reprise *Calojira* par Ultra Vomit),
- La diminution de tonalité d'une octave (*e.g.* *The Sound Of Silence* de Simon and Garfunkel et la reprise par Disturbed).

Ces différents aspects peuvent être également cumulés dans une même reprise, telle que celle de *Time* de Pink Floyd par Easy Star All-Stars, qui en change le genre pour du Reggae-Dub, y ajoute des paroles et en modifie le rythme et la durée. En identifiant les différents types de reprises et en constituant des bases de données musicales correspondantes, il devient alors plus simple de créer des méthodes d'identification adéquates et d'isoler les caractéristiques audio propres à chaque type de reprise. Une méthode peut, en

1. https://ccrma.stanford.edu/damp/, consulté le 19 Avril 2018.
2. https://www.smule.com/apps, consulté le 19 Avril 2018.
3. https://www.smule.com/, consulté le 19 Avril 2018.
4. https://vimeo.com/249694026, consulté le 19 Avril 2018.

effet, être induite en erreur lors de la recherche du morceau Électro-Pop *Sweet Dreams (Are Made Of This)* d'Eurythmics et ne pas trouver la reprise de Marilyn Manson car cette méthode ne prend pas en compte une variation de genre ni de dynamique audio trop importante.

Afin d'améliorer la qualité des méthodes d'identification de reprises, il semble donc pertinent de proposer des bases de données musicales indiquant les types de modification apportées aux morceaux originaux. Ainsi, il devient possible de comparer des méthodes par rapport aux types de reprises qu'elles sont capables d'identifier. Une base de données musicales, qui identifie les types de reprises qu'elle contient [Bayle *et al.*, 2017b], est décrite dans la section suivante.

Description de la base de données proposée

Tout comme pour la base MIR-1K [Hsu et Jang, 2010], la proposition de reprises qui se rapprochent le plus des versions originales a été l'objectif principal de cette nouvelle base de données musicales. Ce choix permet de développer des méthodes d'identification de reprises qui identifient uniquement les variations propres aux différences de timbre des instruments et de la voix, tout en étant robustes aux changements de genre, de mode, de tonalité, de paroles ou d'effets audio. Afin de développer cette base de données musicales ainsi que des méthodes d'identification des reprises, un partenariat a été mis en place avec une équipe de recherche tchèque spécialisée dans la mesure de distance musicale entre deux morceaux. Ce partenariat a permis d'utiliser la base de données constituée afin de proposer une étude de comparaison des caractéristiques audio dans le cadre de l'identification des reprises.

Le rassemblement des morceaux et la constitution des annotations correspondantes ont pu être réalisés grâce à un partenariat instauré avec l'entreprise française Recisio [1]. Recisio propose notamment l'application de karaoké Karafun [2] qui permet aux utilisateurs de chanter des morceaux et de régler le volume de l'accompagnement instrumental ainsi que du chant original. Afin de proposer aux utilisateurs finaux un tel réglage du volume indépendamment des pistes instrumentales et vocales, Recisio possède individuellement chacune de ces pistes qui ont été enregistrées par des professionnels. Les morceaux proposés par Recisio constituent donc une source de reprises professionnelles qui permettent de mettre en place une base de données musicales de qualité dédiée à la recherche. Parmi les 26 872 morceaux proposés par Recisio, les 1 000 les plus chantés en Janvier 2016 ont été utilisés afin de constituer une base qui a été nommée Karaoke database of 1,000 tracks (Kara1K). Il est possible d'explorer ces morceaux sur la page web du projet correspondant afin

1. www.recisio.fr, consulté le 19 Avril 2018.
2. http://www.karafun.fr, consulté le 19 Avril 2018.

d'avoir un aperçu de l'étendue des données et métadonnées proposées [1]. Chacune des reprises proposées ainsi que le morceau original correspondant sont référencés par leur ISRC, de même que dans le cas de la base SATIN présentée précédemment. Bien qu'une reprise ressemble au morceau original au sein de Kara1K, il existe des différences en ce qui concerne les chanteurs, le timbre des instruments, les légères variations de tempo et les conditions d'enregistrement. L'écoute de chacun des morceaux a permis de constituer une base d'annotations liée à Kara1K.

Il est de plus possible d'écouter les morceaux présents dans Kara1K sur l'application Karafun, bien que Recisio se réserve le droit de retirer ou d'ajouter des morceaux à tout moment. Les fichiers audio utilisés pour constituer Kara1K ont par ailleurs été normalisés afin de présenter une unique piste monophonique échantillonnée à 44 100 Hz et encodée en 16 bits non signés au format WAV. Les chansons demeurent toutefois protégées par des droits d'auteurs qui rendent impossible le partage des fichiers audio. Afin de pallier ce problème, la base de données musicales Kara1K est fournie avec de multiples caractéristiques audio et il est possible d'en extraire de nouvelles sur demande. Les logiciels suivants ont été utilisés pour extraire des caractéristiques audio des chansons contenues dans Kara1K :

- *Essentia* 2.1 [Bogdanov *et al.*, 2013] [2],
- *harmony-analyser* 1.2 [Maršík, 2016] [3],
- *Marsyas* 0.5 [Tzanetakis et Cook, 2000] [4],
- *Vamp plugins* [Mauch et Dixon, 2010] [5],
- *YAAFE* 0.64 [Mathieu *et al.*, 2010] [6].

Les caractéristiques audio extraites sont mises à disposition librement sur le site web d'accompagnement de Kara1K.

Métadonnées et base d'annotations

Kara1K contient les noms des artistes, des chansons, les identifiants internes à Recisio ainsi que les ISRCs correspondants. Kara1K propose également l'année d'enregistrement des reprises et une annotation concernant la présence de paroles explicites dans chaque morceau. Comme indiqué dans la figure 2.4, Kara1K fournit des annotations en ce qui concerne le langage utilisé dans les

1. http://yannbayle.fr/karamir/, consulté le 19 Avril 2018.
2. http://essentia.upf.edu, consulté le 19 Avril 2018.
3. http://harmony-analyser.org, consulté le 19 Avril 2018.
4. http://marsyas.info, consulté le 19 Avril 2018.
5. http://www.vamp-plugins.org, consulté le 19 Avril 2018.
6. http://yaafe.sourceforge.net, consulté le 19 Avril 2018.

paroles de chaque chanson. Il est donc possible de concevoir des listes de lecture restreintes à une seule langue.

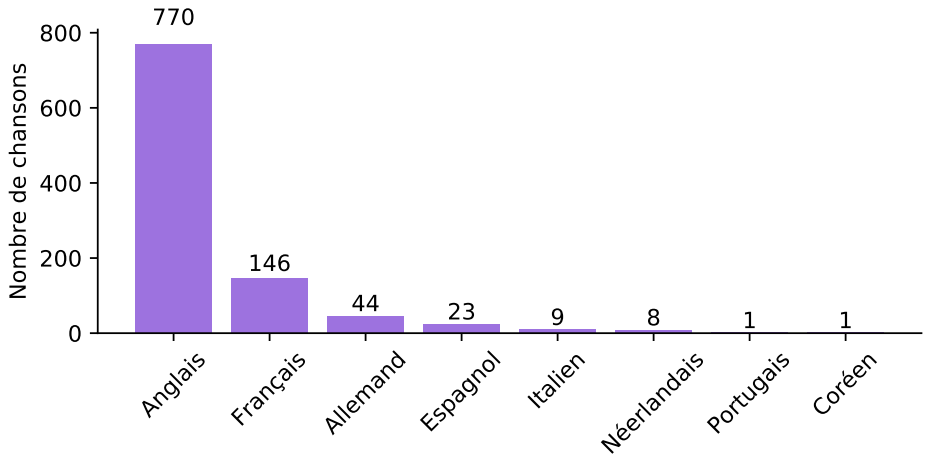


FIGURE 2.4 – Répartition des annotations relatives aux langues dans les chants présents dans Kara1K. Certaines chansons peuvent contenir du chant dans plusieurs langues, c'est pourquoi la somme des annotations de langues dans l'ensemble des chansons n'est pas de 1 000.

Kara1K contient également des annotations en ce qui concerne le ou les genres de chaque morceau. Les dix annotations de genre les plus présentes sont affichées dans la figure 2.5.

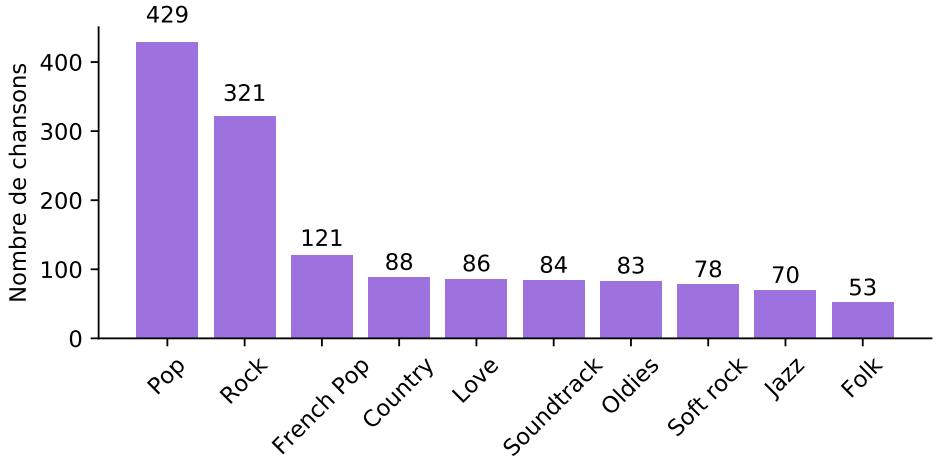


FIGURE 2.5 – Nombre de chansons affichant les dix annotations de genre les plus présentes dans Kara1K. Certaines chansons peuvent afficher plusieurs genres, c'est pourquoi la somme des annotations de genre n'est pas de 1 000.

Après l'écoute attentive de chaque chanson, il a été possible d'intégrer une annotation dans Kara1K concernant la présence de chœurs, de duos, de chanteurs féminins ou masculins ainsi que la simultanéité du chant pour plusieurs chanteurs, comme indiqué dans la figure 2.6.

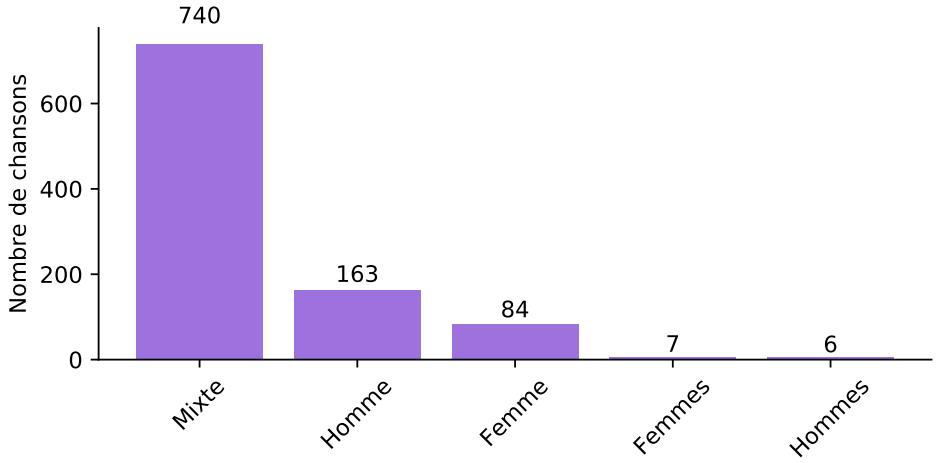


FIGURE 2.6 – Nombre de chansons pour chaque annotation relative au genre du ou des chanteurs et des chœurs dans Kara1K.

La figure 2.7 est une impression d'écran de l'application Karafun sur laquelle le processus de constitution de la base d'annotations a été ajouté.

Conclusion sur la base de données proposée

Pour conclure, la réalisation de Kara1K indique qu'il est possible et bénéfique de développer des partenariats avec des entreprises afin d'avoir accès à des bases de données musicales plus variées et conséquentes que celles utilisées dans l'état de l'art. Sur le plan technique, les sections précédentes ont décrit les étapes de la mise en place d'une base de données musicales et des annotations afférentes assurant la réplicabilité des expériences de recherche.

Le tableau 2.2 résume les principales caractéristiques des deux bases de données proposés durant cette thèse.

Le nombre de morceaux proposés dans les bases Kara1K et SATIN est supérieur à celui utilisé par l'état de l'art mais ce nombre demeure relativement faible comparé aux 150 millions de morceaux disponibles sur SoundCloud. Or, un nombre trop limité de morceaux peut altérer la considération des résultats de classification d'une méthode comme statistiquement significatifs ou proches de la réalité. De plus, la prise en considération de davantage de données permet d'augmenter significativement les résultats des méthodes d'apprentissage machine. La section suivante décrit des méthodes permettant d'augmenter artifi-

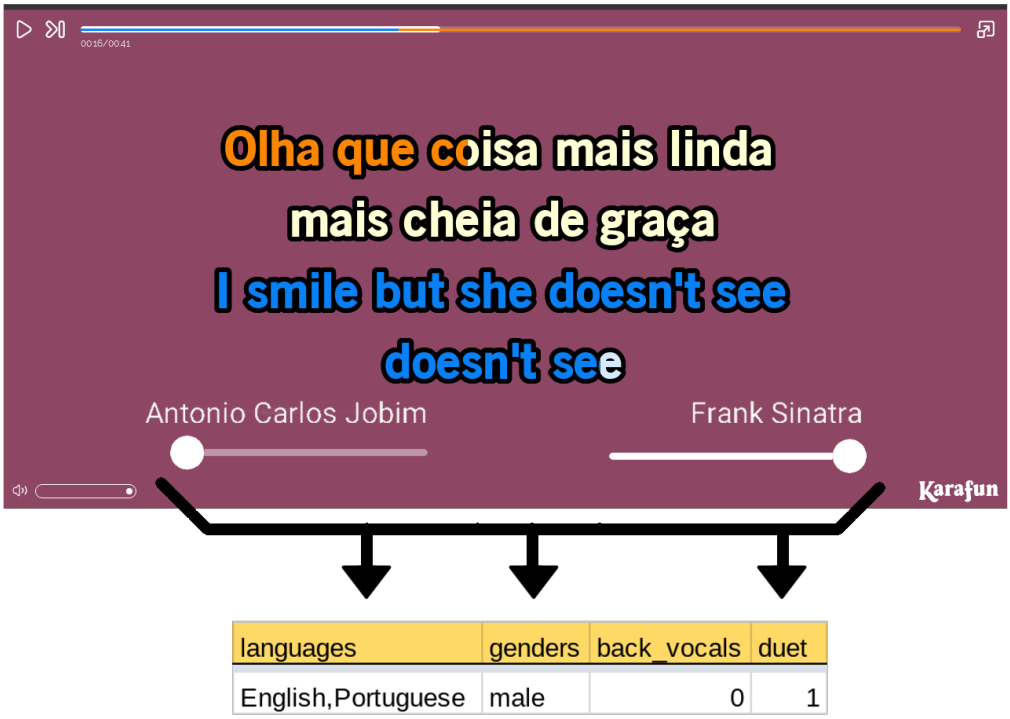


FIGURE 2.7 – Impression d'écran de l'application Karafun. L'utilisateur de l'application peut modifier le volume de la voix de chaque chanteur indépendamment. L'exemple de reprise proposée est un duo d'Antonio Carlos Jobim et de Frank Sinatra. Deux chanteurs sont présents dans cette reprise et ceux-ci chantent successivement, l'annotation associée est donc *Homme* et cette annotation aurait été *Hommes* si le chant avait été simultané.

Tableau 2.2 – Principaux attributs des deux bases de données proposées.

Base de données	SATIN	Kara1K
Nombre de morceaux	408 380	2 000
Code source	GitHub	GitHub
Catégories d'annotations	Genre, Année, Paroles, Humeur, Pays, Activité Présence de chant	Genre, Langue, Chœur, Chanteur, Duo
Caractéristiques audio	Essentia, Marsyas, YAAFE	Essentia, Marsyas, Vamp plugins, YAAFE, harmony-analyser

ciellement la quantité de données d'une base sans avoir à collecter de nouveaux éléments.

2.4 Augmentation artificielle d'une base de données

Les sections précédentes ont présenté les difficultés entourant la constitution de grandes bases de données musicales et des annotations correspondantes. Outre les difficultés liées à l'utilisation de ces bases, ces dernières peuvent afficher des biais parmi les morceaux proposés. Un exemple de biais observé et détaillé dans la littérature [Sturm *et al.*, 2014] concerne la détection des genres musicaux qui a été réalisée lors de la campagne d'évaluation annuelle 2013 [1] du MIREX. La méthode proposée par Pikrakis [2013] a obtenu les meilleurs résultats parmi 13 méthodes proposées en détectant correctement le genre de 77% des morceaux présents [2] dans la base de données musicales du MIREX. Toutefois, Sturm *et al.* [2014] ont reproduit la méthode de Pikrakis [2013] en augmentant artificiellement la base de données grâce à des modifications de tempo et ont mis en évidence qu'il était possible d'obtenir des résultats proches de prédictions aléatoires ou proches de 100% en fonction des modifications de tempo effectuées. Cette constatation suggère que la méthode proposée par Pikrakis [2013] est un Horse puisqu'elle ne détecte pas le genre musical mais le tempo.
La base de données musicales proposée par le MIREX est de plus biaisée puisque les tempos seuls permettent de retrouver le genre d'un morceau, ce qui n'est pas le cas dans la réalité. Le tempo peut en effet être utilisé conjointement avec d'autres caractéristiques audio pour détecter le genre d'un morceau [Gouyon et Dixon, 2004] mais ne constitue pas une caractéristique suffisante pour déterminer un genre musical. La base de données utilisée par le MIREX ne contient donc pas des exemples suffisamment représentatifs des genres musicaux et les résultats de classification sur cette base doivent être considérés avec précaution. Un faible nombre de données dans une base constitue donc un frein à la représentativité statistique et à la qualité des résultats de classification obtenus sur celle-ci. De plus, un trop faible nombre de morceaux diminue la représentativité d'une base de données musicales en ce qui concerne un thème spécifique et peut biaiser la qualité de la classification.

Le biais de tempo décrit n'est qu'un exemple parmi d'autres [Sturm, 2014a] et les bases de données musicales existantes contiennent des biais inhérents à leur constitution qui doivent être pris en compte lors de leur analyse. Afin de réduire le nombre d'éventuels biais présents dans une base de données musicales, une solution consiste à générer de nouvelles données à partir de données existantes. Cette génération est qualifiée d'augmentation de données d'une

1. http://www.music-ir.org/nema_out/mirex2013/results/act/latin_report/, consulté le 19 Avril 2018.
2. http://www.music-ir.org/nema_out/mirex2013/results/act/latin_report/summary.html, consulté le 19 Avril 2018.

base.

La section 2.4.1 décrit des méthodes d'augmentation de données propres à l'audio. La section 2.4.2 détaille différentes méthodes de rééquilibrage de thèmes utilisables pour n'importe quel type de données en entrée.

2.4.1 Augmentation de données à partir de caractéristiques musicales

> *« Car il suffit au faible d'éprouver sa force, et alors il sera fort. »*
> – *Les Aventuriers de la mer*, 1998, Robin Hobb

Les méthodes d'augmentation de données à partir de caractéristiques musicales consistent à modifier simultanément une ou plusieurs caractéristiques audio afin de produire de nouveaux morceaux dérivés du morceau initial.

Parmi ces méthodes, les plus couramment utilisées sont :

- Le décalage de fréquences réalisé par demi-ton [Bayle *et al.*, 2016; Kum et Nam, 2017; Thickstun *et al.*, 2017; Ycart et Benetos, 2017] ou de manière continue [Schlüter et Grill, 2015; Thickstun *et al.*, 2017],
- Le mélange de pistes audio [Schlüter et Grill, 2015; Takahashi *et al.*, 2016],
- La transposition circulaire [Martel Baro, 2017],
- Le mixage [Martel Baro, 2017],
- L'interversion d'une piste d'un morceau avec celle d'un autre morceau [Martel Baro, 2017],
- L'ajout de bruit [Schlüter et Grill, 2015],
- L'étirage temporel [Schlüter et Grill, 2015; Bayle *et al.*, 2016],
- La modification du volume sonore [Schlüter et Grill, 2015; Bayle *et al.*, 2016],
- L'ajout de filtres de fréquences [Schlüter et Grill, 2015],
- La suppression d'échantillons [Schlüter et Grill, 2015],
- La modification du tempo [Bayle *et al.*, 2016].

Lors de l'application de ces modifications aux fichiers audio, il faut également modifier les annotations correspondantes. Si, par exemple, une méthode tente d'estimer le tempo d'un morceau, alors une augmentation de données en utilisant le tempo doit être accompagnée de l'annotation du nouveau tempo généré à chaque nouveau morceau. La modification simultanée de morceaux et

de leurs annotations peut être effectuée grâce au logiciel proposé par McFee *et al.* [2015a] en Python [1].

De plus, toutes les méthodes d'augmentation de données ne peuvent pas être utilisées pour toutes les tâches et certaines n'auront aucun impact sur certains résultats obtenus. Par exemple, l'augmentation de données qui utilise la modification du volume sonore ne doit pas avoir d'impact sur les résultats d'estimation du tempo, à moins que la méthode ne soit un Horse.

Schlüter et Grill [2015] ont réalisé l'étude la plus complète à ce jour de comparaison des différentes méthodes d'augmentation de données musicales. Cette étude a été réalisée afin de déterminer les avantages et inconvénients de leur méthode de détection des instants chantés dans un morceau. Dans ce cadre, l'ajout de bruit et la suppression d'échantillons sonores ont amené plus d'erreurs de détection par leur méthode que lorsqu'ils ont utilisé une base non augmentée. Les résultats de leur méthode de détection du chant n'ont en revanche pas diminué face à des modifications dues à un étirage temporel, à la modification du volume sonore, au mixage, au décalage de fréquences, à l'ajout de filtres de fréquences ainsi qu'à l'utilisation successive de plusieurs méthodes d'augmentation de données.

Il est important d'ajouter que l'augmentation de données à partir de caractéristiques musicales est bénéfique aux méthodes d'extraction d'informations musicales, d'une part lorsque la base de données considérée est restreinte et d'autre part lorsqu'il s'agit d'améliorer les résultats issus de bases plus conséquentes.

La section suivante présente des méthodes d'augmentation de données à partir de caractéristiques musicales. Une expérience de classification des instrumentaux et des chansons a par la suite été réalisée sur cette base augmentée [Bayle *et al.*, 2016].

Expérience de classification des Instrumentaux et des Chansons

L'objectif de cette expérience est de souligner l'importance de l'augmentation de données à partir de caractéristiques musicales afin d'éviter les Horses. L'application considérée consiste à différencier les instrumentaux des chansons dans une base de données musicales [Bayle *et al.*, 2016].

Cette expérience évalue la précision d'une méthode de classification des instrumentaux et des chansons pour deux tailles de bases de données différentes. La méthode considérée est celle de Ghosal *et al.* [2013], qui affiche une *accuracy* de 93% sur une base de données musicales de 540 morceaux qui n'ont pas subis d'augmentation. Dans l'expérience proposée, la base de données utilisée contient des morceaux obtenus de manière synthétique grâce à l'augmentation de données. Onze modifications ont été appliquées à 782 morceaux uniques. Les fichiers audio, d'une durée comprise entre une et six minutes, proviennent d'une

1. https://github.com/bmcfee/muda, consulté le 19 Avril 2018.

bibliothèque personnelle et sont au format WAV, échantillonnés à 44 100 Hz en stéréophonie. Cinq genres musicaux sont représentés, à savoir le Rock, le Blues, le Jazz, l'Électro et le Classique. Les modifications, qui consistent à faire varier la vitesse, le tempo, le volume et la fréquence du morceau, ne sont pas cumulées. La version 2.1.1 du logiciel Audacity [1] a été utilisée afin de réaliser ces modifications, qui sont présentées dans le tableau 2.3.

Tableau 2.3 – Liste des modifications appliquées aux fichiers audio et utilisées pour constituer la base de données musicales de l'expérience de classification des instrumentaux et des chansons [Bayle *et al.*, 2016].

Type de modification du signal audio	Valeur de modification
Décalage de fréquence	+1, +7 demi-tons
Étirage temporel avec conservation des fréquences	x0,5, x1.5, x2
Étirage temporel sans conservation des fréquences	x0,5, x1.5, x2
Modification du volume sonore	-6 dB, -3 dB, +3 dB

La précision de classification des instrumentaux et des chansons de la méthode proposée par Ghosal *et al.* [2013] est donc évaluée sur la base de données musicales ainsi augmentée. Si leur méthode détecte effectivement le chant, alors la précision affichée ne devrait pas diminuer pour des morceaux dont le tempo ou le volume sonore ont été modifiés. En effet, puisque chez l'Homme de telles modifications n'empêchent pas de détecter la présence de chant, il devrait en être de même pour la méthode de Ghosal *et al.* [2013].

La précision obtenue par la méthode de Ghosal *et al.* [2013] affiche 77% pour 1 000 morceaux et chute à 66,9% pour 8 000 morceaux [Bayle *et al.*, 2016]. De plus, les précisions affichées par la méthode de Ghosal *et al.* [2013] dans cette expérience et sur ces morceaux sont inférieures à celles présentées dans leur article. Il semble donc que la méthode de ces auteurs présente certaines des caractéristiques d'un Horse puisqu'elle ne traite pas le problème qu'elle prétend résoudre. Si leur méthode détectait en effet la présence de chant, alors la précision atteinte ne chuterait pas lorsque les morceaux sont modifiés en termes de fréquence, de volume sonore ou de tempo.

Afin de vérifier qu'une méthode n'est pas un Horse, il apparaît donc pertinent d'utiliser des méthodes d'augmentation de données à partir des caractéristiques musicales. Il existe également d'autres méthodes d'augmentation de données, qui sont présentées dans la section suivante.

1. http://audacity.fr/, consulté le 19 Avril 2018.

2.4.2 Augmentation d'une base à partir de propriétés statistiques

Il est possible de créer de nouvelles données dans une base en copiant des caractéristiques de données existantes et les modifiant suivant des propriétés statistiques. Ce type de procédé est qualifié d'augmentation de données à partir d'un ré-échantillonnage statistique. Ce procédé ne doit pas être confondu avec la méthode de ré-échantillonnage d'un signal audio.

Les méthodes statistiques de ré-échantillonnage sont utiles afin de modifier une base de données qui présente des déséquilibres au sein des thèmes qu'elle contient. Elles permettent donc d'améliorer les capacités de généralisation des méthodes de classification par apprentissage [Mani et Zhang, 2003]. Il est, en effet, plus facile pour une méthode de classification d'apprendre ce que représentent le Rock et le Mambo si autant d'exemples de chaque genre lui sont présentés plutôt que si 98 morceaux de Rock et 2 morceaux de Mambo lui sont fournis. Néanmoins, il n'est pas toujours possible de collecter davantage de morceaux de Mambo étant donnée la faible prévalence de ce genre dans les bases de données musicales comparée à celle du Rock. Le ré-échantillonnage de certaines propriétés statistiques des morceaux de Rock et de Mambo permet donc dans ces conditions d'obtenir une base de données musicales équilibrée, c'est-à-dire contenant approximativement autant de morceaux de chacun des deux genres.

Il existe deux catégories de méthodes de ré-échantillonnage, qui peuvent être utilisées indépendamment ou successivement. La première catégorie est composée de méthodes qui sous-échantillonnent les thèmes les plus représentés dans une base de données. La seconde sur-échantillonne en revanche les thèmes les moins représentés. Les principales méthodes de ré-échantillonnage sont décrites ci-après mais il en existe davantage [Branco *et al.*, 2016]. Parmi les méthodes de sous-échantillonnage du thème majoritaire, il existe :

- Condensed Nearest Neighbours (Condensed-NN) [Hart, 1968] qui filtre tous les éléments du thème majoritaire qui n'ont pas pour voisin un élément du thème minoritaire,
- Edited Nearest Neighbours (ENN) [Wilson, 1972] qui supprime un élément lorsqu'au moins deux de ses trois plus proches voisins ne présentent pas le même thème,
- Repeated Edited Nearest Neighbours (RENN) [Tomek, 1976a] qui réitère plusieurs fois ENN. Aucune condition d'arrêt à la répétition n'est présente puisque le nombre de répétitions est fixé au préalable,
- All K-Nearest Neighbours (AllKNN) [Tomek, 1976a] qui, par comparaison à RENN, augmente le nombre de plus proches voisins à chaque itération,

- Extraction of majority-minority Tomek links (Tomek) [Tomek, 1976b] qui supprime des éléments du thème majoritaire qui sont les plus proches du thème minoritaire,
- One-Sided Selection (OSS) [Kubat et Matwin, 1997] qui applique la méthode Tomek suivie de Condensed-NN,
- Neighboorhood Cleaning Rule (NCR) [Laurikkala, 2001] qui modifie ENN afin d'augmenter considérablement la suppression des éléments du thème majoritaire comme suit. Pour chaque élément, ses trois plus proches voisins sont récupérés. Si l'élément appartient au thème majoritaire et que la classification par ses trois plus proches voisins indique le contraire, alors l'élément est supprimé. Si l'élément appartient au thème minoritaire et que la classification par ses trois plus proches voisins indique le contraire, alors les voisins appartenant à la classe majoritaire sont supprimés,
- NearMiss using nearest neighbours (NearMiss) [Mani et Zhang, 2003] qui conserve uniquement les éléments du thème majoritaire qui sont proches des éléments du thème minoritaire,
- Random majority Under-Sampling (RUS) [Kotsiantis *et al.*, 2006] qui est une méthode non-heuristique de suppression aléatoire des éléments du thème majoritaire,
- Instance Hardness Threshold (IHT) [Smith *et al.*, 2014] qui filtre les éléments ayant le moins de probabilité d'appartenir au thème majoritaire.

L'inconvénient majeur du sous-échantillonnage consiste en la possibilité de supprimer des éléments qui sont peut-être cruciaux pour différencier les thèmes présents dans la base de données. Dans certains cas et pour éviter ce problème, il semble donc préférable de sur-échantillonner les éléments du thème minoritaire. Parmi les méthodes de sur-échantillonnage couramment utilisées on retrouve :

- Synthetic Minority Over-sampling TEchnique (SMOTE) [Chawla *et al.*, 2002] qui crée aléatoirement de nouveaux éléments de la classe minoritaire en interpolant les caractéristiques des éléments minoritaires proches,
- borderline Synthetic Minority Over-sampling TEchnique (bSMOTE) [Han *et al.*, 2005] qui améliore SMOTE en supprimant la part de hasard dans la constitution de nouveaux éléments : à chaque élément est attribuée une catégorie en fonction de ses plus proches voisins. Un élément est considéré comme étant du bruit si ses plus proches voisins appartiennent à un autre thème. Un élément est considéré comme étant en danger si au moins la moitié de ses plus proches voisins appartiennent au même thème. Un élément est considéré comme étant en sécurité si tous ses voisins appartiennent au même thème que lui,

- Random minority Over-Sampling (ROS) [Kotsiantis *et al.*, 2006] qui est une méthode non-heuristique répliquant aléatoirement les éléments de la classe minoritaire,
- ADAptive SYNthetic sampling approach for imbalanced learning (ADASYN) [He *et al.*, 2008] qui est une extension de SMOTE. Le nombre de nouveaux éléments générés pour chaque élément existant est proportionnel au nombre d'éléments du thème opposé dans le voisinage,
- Support Vector Machine and Synthetic Minority Over-sampling TEchnique (SVMSMOTE) [Nguyen *et al.*, 2009] qui utilise une SVM comme méthode de détection des plus proches voisins dans la méthode SMOTE,

Par ailleurs et puisque le sur-échantillonnage peut permettre d'augmenter considérablement la taille d'une base de données, il peut par conséquent augmenter le temps de calcul des méthodes de classification. Le temps de calcul peut constituer un facteur limitant lors des analyses musicales puisque certaines méthodes de classification possèdent des complexités[1] en $\Omega(n^3)$. Le second inconvénient inhérent à l'utilisation d'un sur-échantillonnage concerne le sur-apprentissage, c'est-à-dire la production d'une analyse qui correspond trop bien à un jeu de données en particulier et qui produira difficilement une analyse fiable sur un nouveau jeu de données[2]. Afin de diminuer l'impact des inconvénients générés par l'emploi des sur- et sous-échantillonnages, il est possible d'utiliser successivement ces deux méthodes sur une même base de données. Parmi les articles dans lesquels deux méthodes de ré-échantillonnage sont employées, le sur-échantillonnage s'effectue avant le sous-échantillonnage. Les méthodes les plus utilisées sont :

- Synthetic Minority Over-sampling TEchnique followed by Tomek links (SMOTETOMEK) [Batista *et al.*, 2003] qui sur-échantillonne tous les thèmes avant d'appliquer un sous-échantillonnage afin de supprimer les éléments qui sont à la limite entre les thèmes,
- Synthetic Minority Over-sampling TEchnique followed by Edited Nearest Neighbours (SMOTEENN) [Batista *et al.*, 2004] qui utilise ENN censé supprimer davantage d'éléments du thème majoritaire que ne le fait Tomek.

Plusieurs études comparatives des méthodes de ré-échantillonnage ont été effectuées pour d'autres domaines de recherche [Batista *et al.*, 2004] et ont montré des résultats variables selon le domaine considéré. Le tableau 2.4 compare

1. https://cathyatseneca.gitbooks.io/data-structures-and-algorithms/analysis/notations.html, consulté le 19 Avril 2018.
2. https://en.oxforddictionaries.com/definition/overfitting, consulté le 19 Avril 2018.

différentes méthodes de ré-échantillonnage appliquées à une base de données musicales. La tâche choisie consiste à identifier le genre de chaque morceau présent dans la base de données musicales FMA présentée dans la section 2.1.

L'algorithme de classification supervisé considéré est une Machine à Vecteurs de Support qui sera détaillée dans la section 4.1.2. L'ensemble des caractéristiques audio proposées par le logiciel *Essentia* ont été utilisées pour cette expérience. Puisque l'objectif de cette section est la description des méthodes d'augmentation de données à partir d'un ré-échantillonnage, les méthodes d'extraction de caractéristiques audio et de classification ne sont pas détaillées davantage car elles font l'objet du chapitre suivant. Ces caractéristiques restent les mêmes pour toutes les comparaisons de méthodes de ré-échantillonnage réalisées dans cette expérience. Les différences de résultats ainsi affichés ne peuvent être imputés qu'à la méthode de ré-échantillonnage et non aux caractéristiques audio utilisées.

Les méthodes de ré-échantillonnage sont comparées sur leurs valeurs de *f-score* et de *loss*, qui sont des métriques définies dans la section 1.3. Les méthodes sont d'autant plus performantes que leur valeur de *f-score* est proche de 100% et que celle de *loss* se rapproche de 0.

L'implémentation des méthodes de ré-échantillonnage a été réalisée par Lemaître *et al.* [2017] en Python [1].

Le tableau 2.4 indique que les résultats obtenus par les méthodes de sous-échantillonnage sont inférieurs pour les deux métriques considérées à ceux obtenus par les méthodes de sur-échantillonnage. Cette constatation est en accord avec les données de la littérature [Mani et Zhang, 2003; Nanayakkara et Caldera, 2016]. Elle explique également pourquoi une succession de sur-échantillonnage puis de sous-échantillonnage ne produit pas de meilleur résultat que l'utilisation seule d'une méthode de sur-échantillonnage. D'après le tableau 2.4, les méthodes de sur-échantillonnage peuvent de plus améliorer la valeur de *loss* jusqu'à 5% et celle de *f-score* jusqu'à 3%.

Les méthodes de ré-échantillonnage peuvent par ailleurs améliorer les résultats de classification mais dépendent du type de tâche à accomplir et des données considérées. Toutefois, ces méthodes ont rarement été utilisées dans le domaine de l'extraction automatique d'informations musicales [Martın *et al.*, 2009; Lin *et al.*, 2011; Nanayakkara et Caldera, 2016]. Dans ce domaine, en effet, la majorité des articles proposent de nouvelles caractéristiques musicales, de nouvelles bases de données ou encore une comparaison des méthodes de classification existantes. Des impacts positifs forts sur les résultats de classification en utilisant des méthodes de ré-échantillonnage ont en revanche été observés en reconnaissance d'image, en détection de bactérie ou en reconnaissance de matériaux à partir de ses composés chimiques [Batista *et al.*, 2004]. Dans ce contexte, il semble donc prometteur de mener davantage d'expériences afin

1. https://github.com/scikit-learn-contrib/imbalanced-learn, consulté le 19 Avril 2018.

d'étudier l'impact des méthodes de ré-échantillonnage sur les bases de données musicales.

Tableau 2.4 – Comparaison des valeurs de *loss* et de *f-score* obtenues sans et avec différentes méthodes de ré-échantillonnage. Les nombres en gras indiquent les meilleurs résultats obtenus, toutes méthodes confondues. Sur la version en couleur de ce mémoire, le nombre en rouge indique un résultat moins satisfaisant que celui des prédictions aléatoires et les nombres en vert indiquent les résultats qui améliorent les résultats obtenus par la méthode de base.

Méthode	*Loss*	*F-score* (%)
Références sans ré-échantillonnage		
Cas parfait	0.0000	100.00
Aléatoire	2.8122	6.17
Base	0.9712	69.83
Sous-échantillonnage		
Condensed-NN [Hart, 1968]	1.7439	49.44
ENN [Wilson, 1972] et RENN [Tomek, 1976a]	1.9307	54.19
AllKNN [Tomek, 1976a]	1.2776	62.41
Tomek [Tomek, 1976b]	0.9746	69.73
OSS [Kubat et Matwin, 1997]	1.3523	57.38
NearMiss [Mani et Zhang, 2003]	3.0400	12.83
RUS [Kotsiantis *et al.*, 2006]	2.0759	38.80
IHT [Smith *et al.*, 2014]	2.4178	37.16
Sur-échantillonnage		
SMOTE [Chawla *et al.*, 2002]	0.9519	71.40
bSMOTE-1 [Han *et al.*, 2005]	**0.9219**	71.66
bSMOTE-2 [Han *et al.*, 2005]	0.9654	70.78
ROS [Kotsiantis *et al.*, 2006]	0.9487	**72.02**
SVMSMOTE [Nguyen *et al.*, 2009]	0.9236	71.70
Sur- puis sous-échantillonnage		
SMOTETOMEK [Batista *et al.*, 2003]	0.9519	71.44
SMOTEENN [Batista *et al.*, 2004]	2.0072	51.51

2.5 Conclusion sur les bases de données musicales

« Il existe une musique de style, et celui qui ne la possède pas, ne saura jamais écrire. »
– Le dictionnaire universel, 1843, Pierre-Claude-Victor Boiste

La base de données musicales constitue le premier élément à considérer pour le développement de méthodes d'extraction d'informations musicales. Ce chapitre a souligné les difficultés inhérentes à la constitution de bases de données musicales afin d'assurer la réplicabilité des expériences en classification thématique musicale. Il en est de même pour la constitution des bases d'annotations qui souffrent de la subjectivité dans la définition de certains thèmes et de l'aspect chronophage de la réalisation de ces annotations. Ce chapitre a également souligné l'intérêt des méthodes permettant d'augmenter artificiellement une base de données ainsi que l'importance du rééquilibrage des thèmes que contient celle-ci.

À partir des connaissances introduites dans ce chapitre, il est possible de détailler davantage la figure 1.1 présentée dans l'introduction et qui détaille la chaîne de traitement des morceaux afin de créer une liste de lecture musicale. La figure 2.8 détaille ces nouveaux éléments.

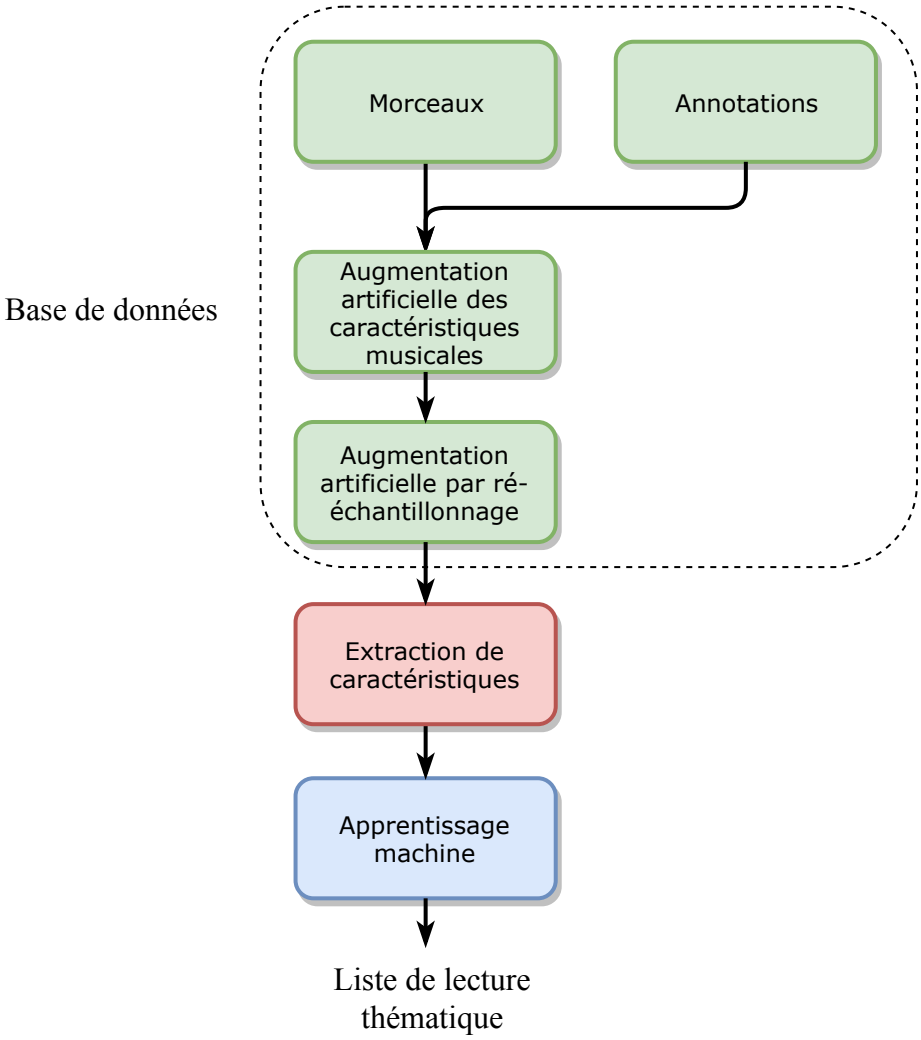


FIGURE 2.8 – Détails de la chaîne de traitement de morceaux permettant de constituer des listes de lecture musicale.

3

Extraction de caractéristiques musicales du signal audio

> *« It is so easy to mistake the superficial aspects of a piece for its beauty, and to imitate them, when the beauty itself is locked deep inside the music, in a way which seems always to elude analysis. »*
>
> – *Gödel, Escher, Bach: an Eternal Golden Braid*, 1979, Douglas R. Hofstadter

Le chapitre précédent a été consacré à la description de la constitution de bases de données musicales pour une utilisation académique. Ce chapitre introduit les concepts et méthodes permettant d'extraire des informations audio et musicales des morceaux contenus dans ces bases de données.

Afin d'appréhender les modalités des méthodes d'extraction de caractéristiques audio, la section 3.1 décrit d'abord les mécanismes physiques de production des sons ainsi que la dimension psychoacoustique liée à leur perception. Cette section présente notamment des expériences psychoacoustiques qui ont été menées afin d'étudier les mécanismes de perception auditive [Demany *et al.*, 2017].

La section 3.2 permet ensuite d'expliquer l'utilisation des phénomènes psychoacoustiques dans le cadre du traitement de signaux audio et de l'extraction de caractéristiques musicales.

3.1 De la production d'un son à sa perception

3.1.1 Introduction à la psychoacoustique

> *« D'une manière générale nous pouvons définir le son comme un coup donné par l'air à travers les oreilles au cerveau et au sang et arrivant jusqu'à l'âme. Le mouvement qui s'ensuit, lequel commence à la tête et se termine dans la région du foie, est l'ouïe. Ce mouvement est-il rapide, le son est aigu ; s'il est plus lent, le son est plus grave. »*
>
> – *Timée* (traduction par Luc Bruisson), 300 av. J.-C., Platon

Si l'intérêt de Platon pour le son au IV^{ème} siècle av. J.-C. démontre déjà une connaissance rudimentaire des mécanismes auditifs durant l'Antiquité, de nombreux phénomènes sonores impliquant le conscient et le subconscient demeurent encore aujourd'hui à découvrir et à comprendre. Le système auditif traite automatiquement et simultanément les informations sonores provenant de multiples sources du milieu extérieur et qui sont indissociées lors de leur arrivée aux tympans. Depuis les années 1930, la psychoacoustique a permis de découvrir de nombreux processus concernant cette perception auditive qui met en jeu des mécanismes complexes.

La psychoacoustique est un domaine d'étude qui repose principalement sur de la recherche fondamentale et dont certaines découvertes ont façonné le monde moderne occidental. L'un des exemples marquants de cette influence concerne l'invention en 1993 du format de compression audio Moving Picture Experts Group Phase 1 Audio Layer III, plus connu sous le nom de MP3. Le taille réduite des fichiers MP3 a en effet permis d'accélérer la diffusion et l'échange de morceaux en ligne ainsi que leur écoute sur des baladeurs portables. L'une des techniques de compression utilisée dans le format MP3 tire avantage d'une limitation perceptuelle de l'audition humaine étudiée en psychoacoustique sous le nom de masquage auditif [Mayer, 1894; Ehmer, 1959; Terhardt *et al.*, 1982]. Un masquage auditif survient en effet lorsque la perception d'un son est affectée par la présence d'un autre son. Par exemple, si deux fréquences d'intensités différentes sont présentes en même temps, celle d'intensité inférieure peut ne pas être perçue si, comme le souligne la figure 3.1, sa fréquence est trop proche de la première.

En utilisant le principe du masque auditif, il est possible de supprimer d'un signal audio les fréquences qui sont masquées tout en minimisant l'impact de cette suppression sur la perception humaine. Cette suppression permet d'alléger le signal audio afin de le compresser. La mise en évidence du masquage auditif a donc eu un impact applicatif fort sur la compression numérique audio. Néanmoins, il existe des lacunes dans les connaissances actuelles concernant les capacités de l'Homme à identifier les caractéristiques sonores d'une source

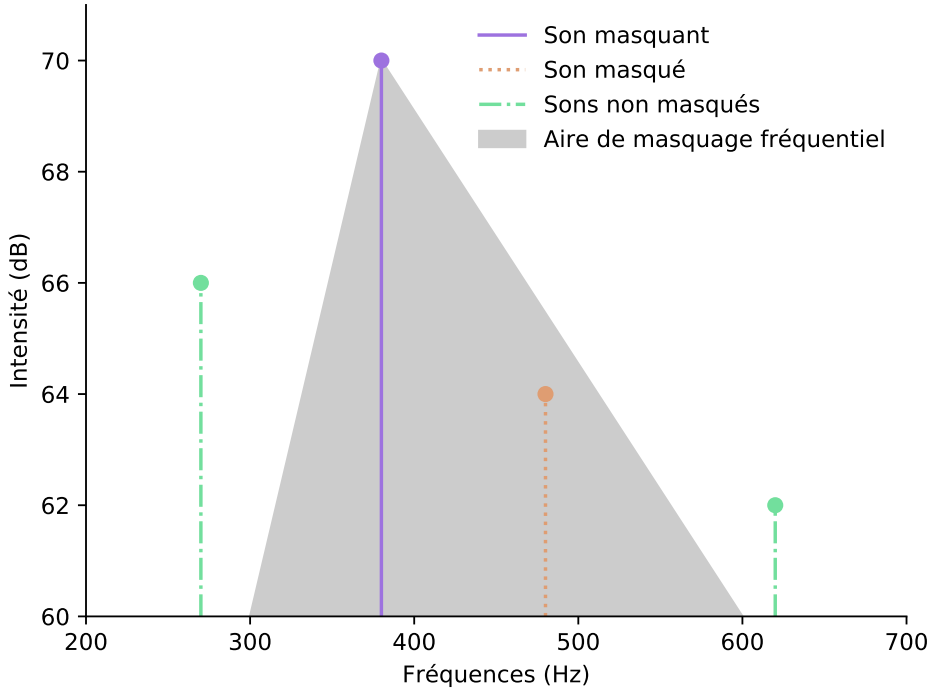


FIGURE 3.1 – Exemple de masquage fréquentiel. L'abscisse représente les fréquences et les ordonnées correspondent aux intensités des sons. Le son masquant empêche la perception auditive chez l'Homme du son masqué de par son intensité et sa proximité fréquentielle. Les sons non masqués ont des fréquences suffisamment distantes de celle du son masquant pour être perçus par l'Homme.

parmi plusieurs autres. La découverte de nouveaux phénomènes psychoacoustiques pourrait par ailleurs permettre la création de nouveaux outils d'analyse et de caractérisation audio. De nouvelles caractéristiques permettraient d'améliorer les méthodes de classification musicale thématique ainsi que la génération de listes de lecture musicale. Des expériences psychoacoustiques ont donc été réalisées afin d'étudier de nouvelles capacités à identifier et caractériser une source sonore. L'objectif consiste à déterminer dans quelle mesure l'oreille humaine est sensible à un changement de fréquence de modulation. Il s'agit donc de prouver l'existence de détecteurs de changement de rythme dans l'oreille humaine [Demany *et al.*, 2017].

La section 3.1.2 détaille d'abord le contexte qui a amené à l'étude des détecteurs de changement de rythme. Les sections 3.1.3 et 3.1.4 décrivent ensuite les deux principales expériences qui ont permis de caractériser ces détecteurs. La section 3.1.5 discute enfin des résultats obtenus dans un contexte fonda-

mental psychoacoustique puis dans un contexte plus applicatif de génération de listes de lecture musicales.

3.1.2 Contexte de l'étude des détecteurs de changement de rythme

> *« We can no longer maintain any distinction between music and discourse about music, between the supposed object of analysis and the terms of analysis. »*
>
> – *Discourse*, 1999, Bruce Horner

Rappel sur les sons

Un son est une vibration mécanique qui se propage de manière longitudinale. Il peut être produit par un instrument de musique, une voix humaine ou tout autre élément présent dans l'environnement. La fréquence et le niveau de pression constituent ses principales caractéristiques physiques. L'intensité, la hauteur, le timbre et le rythme en constituent les principales caractéristiques perceptibles par l'être humain.

La fréquence, exprimée en Hertz (Hz), correspond au nombre de répétitions par seconde de la vibration mécanique due à la propagation d'un son. La hauteur correspond au caractère plus ou moins aigu d'une fréquence. L'oreille humaine est capable de distinguer des sons graves de basses fréquences allant jusqu'à 20 Hz et des sons aigus de hautes fréquences allant jusqu'à 20 000 Hz. En comparaison, la fréquence fondamentale de la note la plus grave de la guitare basse est par exemple de 41 Hz et celle de la note la plus aiguë du piano est de 4 186 Hz. Un ensemble de notes jouées simultanément est qualifié d'accord.

Le niveau de pression est quantifiable selon plusieurs échelles et celle retenue est le décibel (dB) Sound Pressure Level (SPL) qui exprime le rapport de la pression acoustique par rapport à la valeur de référence de 0 dB SPL. Cette référence correspond au minimum de pression acoustique perçu par l'oreille humaine à 1 000 Hz et qui est de 20 μPa.

L'intensité perçue se mesure quant à elle en sone et dépend de la pression acoustique ainsi que de la fréquence. Pour exemples, la pression acoustique reçue lorsqu'une personne chuchote à 1 mètre est de 30 dB tandis que celle-ci est d'environ 110 dB dans une discothèque, ce qui correspond respectivement à 0,4 et 128 sones.

Le timbre est à différencier de la hauteur dans la mesure où il est dépendant de la répartition des fréquences du spectre sonore ainsi que de leur évolution temporelle. Par exemple, pour une même note de musique, le timbre permet de différencier celle jouée par un piano de celle jouée par une trompette.

Il n'existe actuellement pas de consensus quant à la définition du rythme. Liddell et Scott [1996] qualifient en effet de rythmique tout mouvement régulier qui se répète temporellement tandis que Bispham [2006] définit le rythme comme étant la perception interne d'une pulsation externe. Par ailleurs, un rythme peut se manifester à différentes échelles, de la milliseconde pour des oscillations de la note Ré à l'échelle de la seconde pour le tempo d'un morceau.

Une légère variation de l'un des paramètres d'un son est perçue par l'oreille humaine et peut modifier la perception de celui-ci de façon importante. L'une des conséquences de cette dernière observation est qu'il est possible d'identifier et de suivre plusieurs sources sonores dans un même environnement. Bien que les sons présents dans la nature ne sont que rarement composés d'une seule fréquence, l'utilisation de sons composés de peu de fréquences est préférable en psychoacoustique puisque les interactions d'un grand nombre de fréquences sont encore mal comprises.

Introduction au fonctionnement de l'audition humaine

Lorsqu'un son est produit puis capté par l'oreille, plusieurs mécanismes sont mis en jeu et permettent par la suite la représentation mentale de ce son.

La psychoacoustique étudie la perception des *stimuli* sonores par le cerveau et de nombreux processus demeurent encore à analyser concernant les mécanismes de perception sous-jacents.

Le système auditif humain fonctionne par bancs de filtre [Munkong et Juang, 2008], c'est-à-dire que chaque fibre du nerf auditif répond à la présentation d'une fréquence. Les fréquences aiguës sont analysées au niveau de la base de la cochlée et les fréquences graves au niveau de l'apex. Cette relation entre un point de l'espace et une fréquence est qualifiée de tonotopique.

La précision des fibres du nerf auditif n'est pas parfaite. En effet, ces fibres répondent également à des fréquences proches de leur fréquence caractéristique mais de manière moins importante. Cependant, leur sélectivité fréquentielle augmente lorsque la fréquence caractéristique diminue, c'est-à-dire que deux fréquences graves proches sont plus précisément différenciables que ne le sont deux fréquences aiguës proches. L'analyse des rythmes est également tonotopique [Giraud *et al.*, 2000]. Le colliculus inférieur [Van Noort, 1969] est par exemple plus sensible à des rythmes compris entre 50 et 120 Hz tandis que le cortex auditif primaire répond mieux à des rythmes compris entre 10 et 20 Hz.

Un son constitue un spectrogramme pour le système auditif, c'est-à-dire qu'il constitue une évolution d'une ou de plusieurs fréquences dans le temps. Lorsqu'un son composé de plusieurs fréquences est présenté à l'oreille, la fusion spectrale complexifie la perception individuelle des fréquences [De Cheveigné, 1997]. Cet effet n'est observé que lorsque toutes les fréquences sont synchrones, c'est-à-dire qu'elles commencent et se terminent au même instant. Si cet effet

est attendu pour des fréquences proches entre elles, il a été constaté que la fusion spectrale s'appliquait également aux fréquences éloignées [De Cheveigné, 1997]. De plus, la perception de la fusion spectrale est plus importante lorsque les fréquences présentent des relations harmoniques.

Introduction aux détecteurs de changement de fréquence

Dans un contexte de fusion spectrale, il est difficile d'identifier chaque fréquence indépendamment, il semble à première vue difficile de détecter la modification d'une fréquence qui n'a pas été perçue au préalable. La capacité à détecter une modification de fréquence qui n'a pourtant pas été perçue consciemment a néanmoins été prouvée et est optimale pour un changement de fréquence d'environ un demi-ton [Demany et Ramos, 2005]. Un demi-ton correspond à un rapport de fréquences de $2^{\frac{1}{12}}$, soit environ 1,059. La fonction des détecteurs de changement de fréquence est l'établissement de liens entre des sons successifs, qui permettent donc de regrouper ces sons entre eux. Ces détecteurs de changement de fréquence sont activés en permanence chez l'être humain, ils ne nécessitent pas de concentration de la part de l'individu et permettent à ce dernier de détecter automatiquement les modifications de son environnement sonore. L'existence de ces détecteurs de changement de fréquence autorise de plus le suivi simultané de plusieurs sources sonores. Les détecteurs de changement de fréquence ne sont toutefois pas sensibles à la localisation du son puisqu'ils sont mis en jeu dans la représentation de la fréquence des sons provenant des deux oreilles [Carcagno *et al.*, 2011]. Afin d'étudier les détecteurs de changement de fréquence, il est par conséquent possible de présenter le même *stimulus* aux deux oreilles. Aucun équivalent des détecteurs de changement de fréquence n'a actuellement été découvert en ce qui concerne les changements de rythme, l'étude de tels équivalents est par conséquent proposé.

Introduction à la perception du rythme

La perception du rythme ne semble pas être propre à l'audition [Levitan *et al.*, 2015]. Pourtant, certains mécanismes ne s'expriment qu'à travers l'audition, comme par exemple le regroupement de sources sonores qui partagent un rythme identique [Moore et Shailer, 1992]. La présence de plusieurs rythmes simultanés crée toutefois des interférences qui empêchent la perception nette de chacun de ces rythmes distincts [Moore et Shailer, 1992]. Cet effet, dénommé Modulation Discrimination Interference (MDI) est d'autant plus important que les rythmes sont proches les uns des autres ou qu'il possèdent des relations harmoniques. L'effet des MDI a été pris en compte en établissant une distance minimale entre chaque rythme et en évitant les relations harmoniques entre les rythmes utilisés.

L'écoute d'un son contenant un rythme complexifie les tâches de distinction fréquentielle par rapport à des sons ne contenant pas de rythme [Jones *et al.*,

2002]. L'impact de l'attention accordée à un rythme sur ces tâches n'a pas été étudié mais l'attention permet à l'auditeur de séparer des sons contenant des rythmes différents [Andreou *et al.*, 2011]. Lorsqu'une source a clairement été identifiée, il est en effet possible de suivre son évolution grâce à son rythme. Le système auditif est un modèle de prédiction temporelle aux suites isochrones d'événements. Toute rupture rythmique est détectable [Schröger *et al.*, 2014] car la rupture brise la prédiction du système auditif. Une telle rupture permet de repérer la fin de l'émission de sons par une source. Les effets d'une accélération ou d'un ralentissement de rythme sur la sensibilité à la rupture rythmique n'ont néanmoins pas encore été caractérisés, l'étude de ce phénomène est donc présenté ci-après.

Étude de la perception d'un changement de rythme

Le système de perception auditive humain est capable de détecter automatiquement des changements de fréquence. L'un des intérêts des expériences présentées dans ce chapitre est de déterminer dans quelle mesure ce même système est sensible à une modification de rythme et donc s'il existe des détecteurs de changement de rythme dans l'oreille humaine. Afin d'étudier cette possibilité, le rythme peut être étudié sous plusieurs formes. L'étude de la sensibilité au rythme *via* une modulation d'amplitude sinusoïdale est pertinente [Formby, 1985] et est retenue pour les expériences présentées dans cette section. La modulation d'amplitude consiste à créer une variation périodique du volume sonore d'un signal. Les deux expériences mesurent la sensibilité à un changement de rythme chez des personnes volontaires et utilisent des paradigmes similaires à ceux décrits pour l'observation des détecteurs de changement de fréquence [Demany et Ramos, 2005].

La compréhension de la sensibilité humaine aux rythmes permet notamment d'approfondir les connaissances concernant la perception des sentiments d'autrui *via* les intonations vocales. Les sentiments se traduisent en effet, dans le domaine de la parole, par des variations rythmiques *via* la modulation d'amplitude [Mencattini *et al.*, 2014]. La modulation d'amplitude présente dans la voix est généralement comprise entre 0,6 et 12 Hz [Payton et Braida, 1999]. L'unité rythmique utilisée en musique est généralement comprise entre 36 et 300 Battements Par Minute (BPM), ce qui correspond respectivement à des fréquences comprises entre 0,6 et 5 Hz. Ces intervalles de rythme sont donc considérés dans les expériences proposées mais des rythmes plus rapides sont également étudiés afin de comparer leur impact sur la perception auditive humaine des rythmes. Il s'agit donc de déterminer s'il est possible pour l'Homme de détecter une accélération ou une décélération rythmiques de façon automatique, c'est-à-dire sans que l'attention soit impliquée dans cette détection. Dans le domaine de l'audition, l'étude la plus semblable à celle proposée concerne la perception des événements sonores réguliers qui surviennent ou disparaissent

de l'environnement sonore [Jones *et al.*, 1981].

La compréhension de la sensibilité humaine aux rythmes permet également de développer de nouveaux descripteurs musicaux qu'il est possible d'extraire des morceaux afin de générer des listes de lecture musicale thématiques. Il est possible d'imaginer une liste de lecture dont le rythme des morceaux accélère afin d'accompagner le réveil ou bien une liste de lecture qui contient uniquement des morceaux dont le rythme est stable afin de faciliter la concentration au travail. Cet aspect est développé dans la section 3.2.

Des exemples de *stimuli* sonores présentés aux participants sont disponibles [1] afin d'aider à la compréhension des expériences.

3.1.3 Mise en évidence de l'existence des détecteurs de changement de rythme

> *« Ce qui est écrit sans peine est lu sans plaisir. »*
> – *Lueurs*, 1951, Louis-Philippe Robidoux

Il a été démontré que la détection d'un changement de hauteur, matérialisé par un changement de fréquence porteuse, ne nécessite pas l'attention de l'auditeur [Demany *et al.*, 2011]. Il s'agit dans cette expérience de déterminer s'il en est de même en ce qui concerne un changement de fréquence de modulation, utilisé ici pour produire un rythme. Le but de cette expérience est donc de savoir s'il existe des détecteurs de changement de rythme comme il existe des détecteurs de changement de fréquence dans l'oreille humaine. Cette expérience tente de fournir des réponses à ces questions. Pour cela, les performances obtenues pour une détection de changement de fréquence porteuse ou de modulation sont comparées. La fréquence de modulation est utilisée pour matérialiser un rythme.

Matériels et Méthodes

Cette expérience utilise des Sons Purs définis dans l'équation 3.1.

$$S_p(t) = A_p sin(2\pi f_p t + \varphi_p) \tag{3.1}$$

où

- t est le temps,
- A_p est l'amplitude du signal porteur,
- f_p est la fréquence du son porteur,
- φ_p est la phase à l'origine du signal porteur.

1. http://yannbayle.fr/french/psychoacoustic.php, consulté le 19 Avril 2018.

Le paramètre t possède une précision de $\frac{1}{Fe}$, où Fe est la fréquence d'échantillonnage fixée à 44 100 Hz. Chaque Son Pur est ensuite modulé en amplitude par une modulante définie par l'équation 3.2.

$$S_m(t) = sin(2\pi f_m t + \varphi_m) \tag{3.2}$$

où

- t est le temps,
- f_m est la fréquence du son modulant,
- φ_m est la phase à l'origine du signal modulant.

Le signal final est donc défini dans l'équation 3.3.

$$S(t) = A_p sin(2\pi f_p t + \varphi_p)\Big(1 + sin(2\pi f_m t + \varphi_m)\Big) \tag{3.3}$$

où

- t est le temps,
- A_p est l'amplitude du signal porteur,
- f_p est la fréquence du son porteur,
- φ_p est la phase à l'origine du signal porteur,
- f_m est la fréquence du son modulant,
- φ_m est la phase à l'origine du signal modulant.

Le signal $S(t)$ est par la suite nommé Son et la somme de plusieurs Sons dénote un Accord. Les Sons qui composent un Accord possèdent chacun des fréquences porteuses et fréquences modulantes différentes. Afin d'étudier l'importance de l'attention dans la détection d'un changement de fréquence porteuse ou de modulation, l'expérience met en jeu quatre conditions qui sont résumées dans le tableau 3.1.

Tableau 3.1 – Définition des quatre conditions de la première expérience en fonction des modifications du signal prises en compte.

Modifications	Fréquence porteuse	Fréquence modulante
Son puis Accord	Condition 1	Condition 3
Accord puis Son	Condition 2	Condition 4

Les conditions 1 et 2 mettent en jeu un Son, respectivement précédé ou suivi d'un Accord. Les fréquences porteuses de l'Accord sont choisies aléatoirement entre 300 et 3 000 Hz. Afin de garantir une certaine consistance avec les rapports de fréquences porteuses, les fréquences modulantes correspondantes

sont sélectionnées aléatoirement entre 4 et 50 Hz. Les fréquences modulantes doivent être séparées au minimum par 800 cents soit huit demi-tons et au maximum par 1 100 cents soit onze demi-tons. Il en est de même en ce qui concerne les fréquences porteuses. Lors de chaque nouvel essai, de nouvelles fréquences sont choisies aléatoirement. Dans les conditions où la fréquence porteuse est modifiée, elle est sélectionnée de façon équiprobable parmi l'une des quatre fréquences porteuses de l'Accord et est augmentée de 800 à 1 100 cents. La fréquence modulante correspondante demeure inchangée.

Les troisième et quatrième conditions impliquent de modifier la fréquence de modulation de l'un des Sons de l'Accord et de conserver une fréquence porteuse identique. Les durées du Son et de l'Accord sont de deux secondes et aucun silence n'est présent entre eux. Le tableau 3.2 détaille le spectrogramme d'un exemple de *stimulus* pour chaque condition.

Pour les conditions 1 et 3, la première moitié du *stimulus* est constituée d'un seul Son alors que la seconde moitié du *stimulus* est composée d'un Accord de quatre Sons.

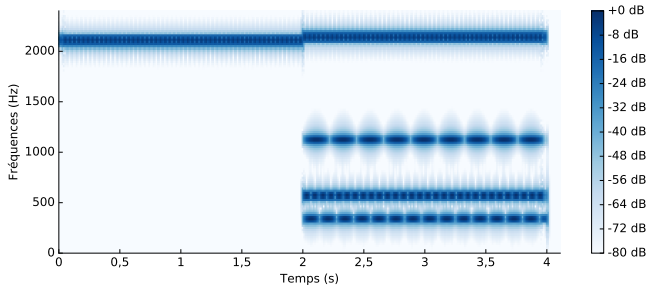
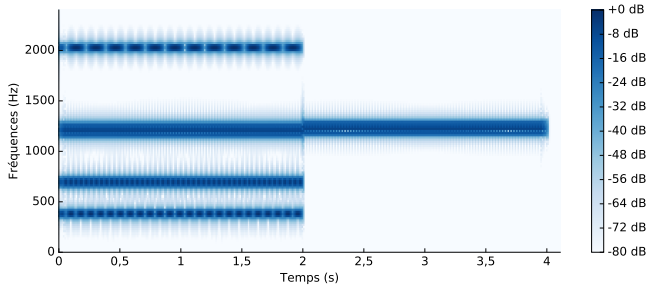
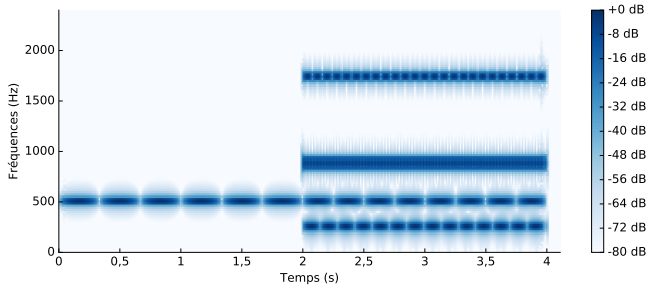
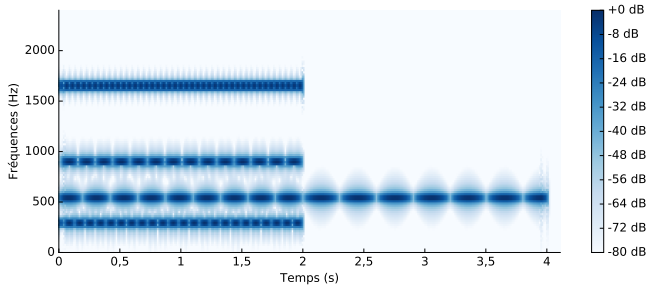
Pour les conditions 2 et 4, cet ordre est inversé puisque le participant entend d'abord l'Accord puis le Son. Pour les conditions 1 et 2, le Son seul a une fréquence porteuse qui est supérieure ou inférieure à l'un des quatre Sons de l'Accord mais possède la même fréquence de modulation.

Pour les conditions 1 et 2, la tâche du participant consiste à juger si la fréquence porteuse perçue subit une augmentation ou une diminution, ce qui se traduit chez l'expérimentateur par la perception d'une modification de hauteur. Dans l'exemple de la condition 1, la fréquence porteuse du premier Son en partant du haut de l'Accord est supérieure à la fréquence porteuse du Son seul et dans ce cas la réponse correcte à fournir est « Monte ». Dans l'exemple de la condition 2, la fréquence porteuse du deuxième Son en partant du haut de l'Accord est inférieure à la fréquence porteuse du Son seul et dans ce cas la réponse correcte à fournir est également « Monte ».

Pour les conditions 3 et 4, le Son seul a une fréquence modulante qui est supérieure ou inférieure à l'un des quatre Sons de l'Accord mais il possède la même fréquence porteuse que le Son de l'Accord considéré. Lorsque le rythme augmente, les stries observées sur les spectrogrammes sont plus rapprochées et elles s'éloignent lorsque le rythme ralentit. Pour les conditions 3 et 4, le sujet doit juger si la fréquence de modulation subit une augmentation ou une diminution, ce qui se traduit par une modification du rythme perçu. Dans l'exemple de la condition 3, le rythme du Son seul est plus lent que celui du troisième Son en partant du haut de l'Accord. Dans ce cas, la réponse correcte est « Accélère ». Dans l'exemple de la condition 4 en revanche, le rythme du troisième Son en partant du haut de l'Accord est plus rapide que celui du Son seul, la réponse correcte est donc « Ralentit ».

Afin d'évaluer la capacité de chaque participant à réaliser l'expérience, deux entraînements éliminatoires leur sont proposés. Ces entraînements ont permis

Tableau 3.2 – Exemple de spectrogrammes de *stimuli* pour chacune des quatre conditions de la première expérience en fonction des modifications du signal effectuées.

Modifications	Spectrogramme d'exemple de *stimuli*
Condition 1 : modification de la fréquence porteuse et présentation du Son puis de l'Accord	
Condition 2 : modification de la fréquence porteuse et présentation de l'Accord puis du Son	
Condition 3 : modification de la fréquence modulante et présentation du Son puis de l'Accord	
Condition 4 : modification de la fréquence modulante et présentation de l'Accord puis du Son	

de vérifier leur capacité à détecter un changement de fréquence porteuse.

Au cours de l'entraînement 1, l'Accord est composé de trois Sons Purs et le Son Pur qui suit a une fréquence porteuse qui est 100 cents plus aiguë ou plus grave que l'un des trois Sons Purs de l'Accord. Le choix de l'un des trois Sons Purs de l'Accord est effectué aléatoirement pour chaque écoute d'un *stimulus*.

En ce qui concerne l'entraînement 2, l'Accord comprend quatre Sons Purs et la modification de fréquence porteuse du Son Pur qui suit est de 50 cents plus aiguë ou plus grave que l'un des quatre Sons Purs de l'Accord. De même, le choix de l'un des quatre Sons Purs de l'Accord est effectué aléatoirement pour chaque écoute d'un *stimulus*. Outre l'aspect éliminatoire, les deux entraînements permettent aux participants de se familiariser avec le matériel utilisé ainsi qu'avec les *stimuli* auditifs.

L'objectif de ces deux entraînements pour les participants consiste à déterminer si la fréquence augmente ou diminue, c'est-à-dire respectivement si elle devient plus aiguë ou plus grave. La perception des dix participants est testée sur 100 présentations du *stimulus* au cours de chaque entraînement. Le score de réussite requis afin de participer à l'expérience est de 90%. Dix participants parmi onze ont atteint ce score et ont donc été retenus pour l'expérience.

L'étape suivante est une condition de contrôle permettant de vérifier l'aptitude des participants à détecter correctement une modification de fréquence de modulation. Cette phase n'est pas éliminatoire. Pour cette condition, le *stimulus* est composé de deux Sons successifs ayant la même fréquence porteuse mais des fréquences de modulation différentes. La tâche consiste à déterminer si la fréquence de modulation du premier Son est plus rapide ou plus lente que celle du second. Ces augmentations ou diminutions sont de 500 cents. L'ampleur de ce changement est par la suite nommé delta et l'ensemble de la condition correspondante est nommé Son-Son. Il est possible de mesurer l'impact de l'attention dans la détermination d'un changement de rythme en comparant les performances obtenues dans les conditions Son-Son, Son-Accord et Accord-Son.

S'il existe des détecteurs de changement de rythme chez l'Homme, alors les performances de détection dans la condition Accord-Son devraient être inférieures ou égales à celles obtenues dans la condition Son-Accord. En effet, dans cette dernière condition, l'auditeur peut se concentrer sur la fréquence qui est modifiée, ce qui n'est pas le cas dans la condition Accord-Son puisqu'il ignore au préalable quel est le Son dont la fréquence sera modifiée. Afin de comparer efficacement les conditions contenant un changement de fréquence porteuse ou modulante, il est préférable d'ajuster les deltas de changement de fréquence porteuse et modulante pour chaque participant afin que ceux-ci fournissent les mêmes performances dans la condition Son-Accord. Le pourcentage de réussite de cette tâche d'ajustement est fixé à 75% pour tous les participants. Chacun d'entre eux participe à l'expérience durant douze heures, qui sont réparties

comme suit : une heure est consacrée aux entraînements 1 et 2, deux heures à la condition Son-Son, quatre heures à la recherche du delta et cinq heures aux conditions Son-Accord et Accord-Son.

La perception des participants est testée dans une cabine insonorisée. L'objectif consiste, pour chaque participant, à répondre à une question relative aux *stimuli* entendus dont les deux réponses possibles sont indiquées sur des boutons virtuels. Un *stimulus* contenant une augmentation ou une diminution de la fréquence porteuse est déclenché respectivement par l'appui sur le premier bouton s'intitulant « Monte » et sur le second s'intitulant « Baisse ». De même, un *stimulus* contenant une augmentation ou une diminution de la fréquence modulante est déclenché respectivement par l'appui sur le premier bouton s'intitulant « Accélère » et sur le second s'intitulant « Ralentit ». Le temps de réponse est illimité. Au lancement du programme, le participant doit sélectionner l'une des quatre conditions définies dans le tableau 3.1. Un essai est défini comme étant la présentation au participant d'un *stimulus* relatif à la condition choisie et l'acquisition de la réponse du participant. Le caractère correct ou erroné de toute réponse est communiqué aux participants *via* la coloration –verte ou rouge– des boutons de réponse. Chaque essai est présenté 400 ms après que le participant a répondu à l'essai précédent et un bloc regroupe 50 essais. Chaque participant répond à huit blocs dans chacune des quatre conditions. Pour l'expérience, l'Accord est composé de quatre Sons modulés.

Les dix participants ont reçu une éducation musicale et ne présentent pas d'anomalies auditives. La moyenne d'âge est de 22,3 ans, l'écart-type de 1,7 an et la répartition des genres est de sept femmes pour trois hommes. Les participants ont signé un accord de consentement de participation et ont été rémunérés.

Programmation des *stimuli*

Les développements des *stimuli*, de l'interface graphique ainsi que de l'analyse des résultats ont été intégralement effectués à l'aide du logiciel Matlab [1] R2012a.

Cinq fonctions, présentées sur la figure 3.2, sont utilisées afin de générer les *stimuli*.

Le générateur de fréquences fournit tout d'abord un Son. Le générateur d'accords renvoie ensuite un Accord de quatre Sons et gère les relations minimales qu'il doit exister entre les différentes fréquences utilisées pour ces quatre Sons. Le générateur d'essais fournit le signal final composé de l'Accord précédemment généré et d'un Son positionné avant ou après celui-ci. Les paramètres de ce dernier sont modifiés en fonction de la condition choisie par l'utilisateur.

1. https://fr.mathworks.com/products/matlab.html, consulté le 19 Avril 2018.

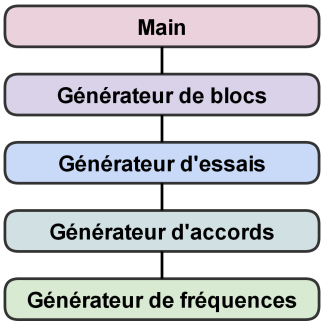


FIGURE 3.2 – Fonctions utilisées pour la chaîne de génération des sons.

Le générateur de blocs gère ensuite la succession des 50 essais ainsi que l'enregistrement de l'intégralité des paramètres. Il crée un fichier par bloc, qui est utilisé pour l'analyse *a posteriori*. La fonction *main* récupère enfin les informations concernant les choix de chaque utilisateur *via* l'interface graphique. Ces informations comprennent l'ordre Son-Accord ou Accord-Son, la modification de fréquence porteuse ou de modulation, le delta ainsi que les initiales du participant.

Afin d'éviter toute discontinuité dans le signal audio, les phases des fréquences porteuse et modulante des Sons sont par construction positionnées à zéro lors du changement entre le Son et l'Accord. Une discontinuité dans le signal audio produit en effet un bruit susceptible de perturber l'identification d'un changement de fréquence. Une telle discontinuité proviendrait des phases des sinus qui, lors du passage d'un son à l'autre, seraient toutes différentes. Il est par conséquent préférable d'assurer la continuité des phases des fréquences porteuses et de modulation lors de la transition entre le Son et l'Accord et *vice-versa*. Si les phases sont identiques lors du changement, cela implique qu'au début et qu'à la fin du *stimulus* les phases seront également toutes différentes de zéro. Afin de pallier toute discontinuité, une rampe de 50 ms est appliquée au début et à la fin du *stimulus* afin de conduire progressivement le volume total à zéro.

La diffusion des sons aux participants est réalisée au moyen d'un casque Sennheiser HD650, sélectionné pour sa courbe de réponse fréquentielle. Cette courbe montre que le casque a la capacité de délivrer la même intensité sonore quel que soit le courant fourni à ses bornes. Cette particularité est uniquement vraie concernant la gamme de fréquences utilisée dans les expériences décrites. Le niveau maximal de chaque Son est fixé à 53 dB SPL.

Le fichier créé pour chaque bloc contient l'ensemble des paramètres utilisés pour les *stimuli* et qui sont :

- Les initiales du participant,
- Les positions relatives du Son et de l'Accord (Son-Accord ou Accord-Son),
- La fréquence à modifier,
- La valeur du delta en cents,
- L'horodatage, c'est-à-dire la date et l'heure de début et de fin du bloc,
- Les réponses de l'utilisateur,
- Les fréquences porteuses et modulantes utilisées pour l'Accord et le Son.

Les réponses de l'utilisateur sont enregistrées dans un vecteur avec la convention de codage présentée dans le tableau 3.4.

Tableau 3.4 – Convention d'enregistrement des réponses de chaque utilisateur pour la première expérience.

Modification réalisée par le programme	Réponse « Monte »	Réponse « Baisse »
Augmentation	1	2
Diminution	4	3

Ce codage permet un calcul rapide du nombre de réponses correctes puisque celles-ci sont représentées par des nombres impairs. Afin d'obtenir le nombre total de réponses correctes, il suffit de quantifier les nombres impairs dans le vecteur stockant les réponses de chaque utilisateur.

L'interface graphique doit présenter une taille de police suffisamment importante pour que le texte puisse être lu par le sujet depuis un écran situé à l'extérieur de la cabine insonorisée et à travers une vitre de plusieurs centimètres d'épaisseur. L'interface met en jeu deux fenêtres qui permettent de sélectionner les paramètres du bloc et de répondre à l'essai en cours. La première fenêtre de l'interface présentée dans la figure 3.3 permet à l'utilisateur de choisir s'il désire effectuer l'entraînement 1, l'entraînement 2 ou bien le test principal.

Si l'utilisateur choisit l'entraînement 1 ou l'entraînement 2, il est directement redirigé vers la seconde fenêtre. S'il choisit d'effectuer le test principal, il doit au préalable renseigner différents paramètres, indiqués dans la capture d'écran présentée dans la figure 3.4.

Afin que tout utilisateur puisse comprendre les paramètres de la fenêtre présentée dans la figure 3.4, les termes « fréquence porteuse » et « fréquence

FIGURE 3.3 – Partie supérieure de la première fenêtre de l'interface graphique.

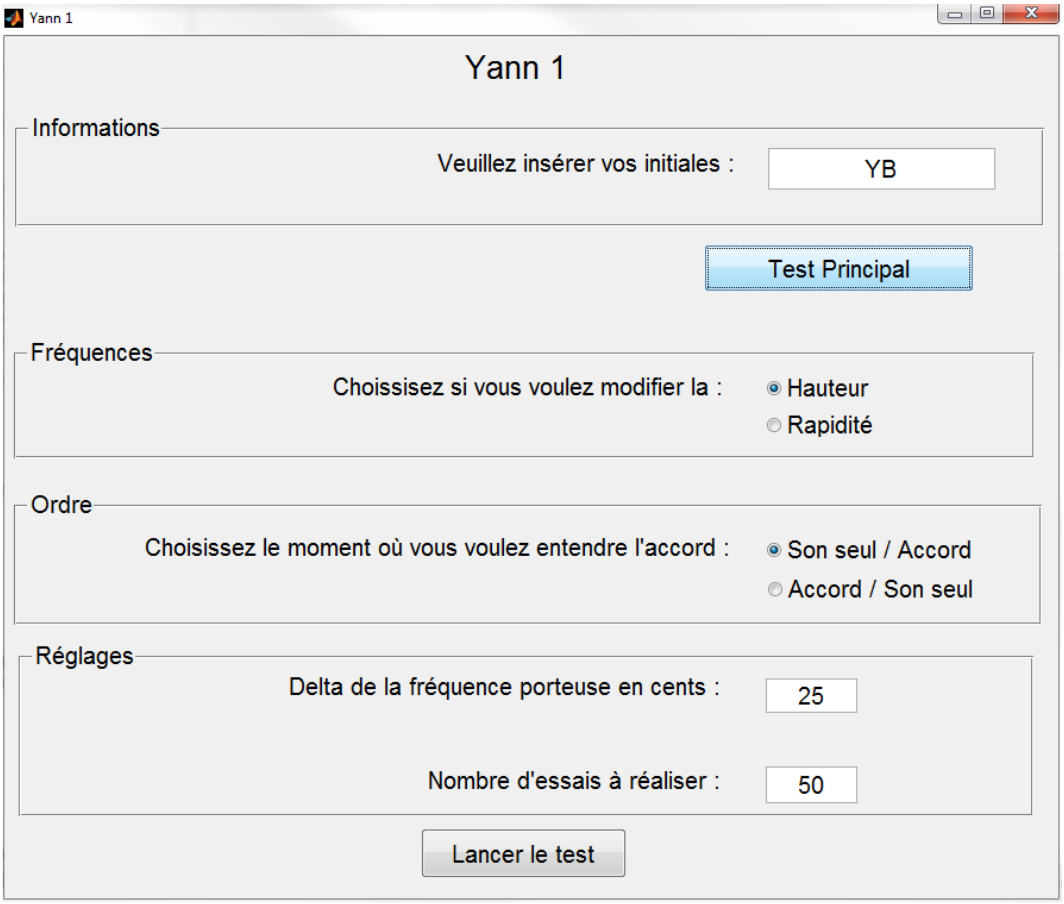


FIGURE 3.4 – Intégralité de la première fenêtre de l'interface graphique.

de modulation » ont respectivement été modifiés en « Hauteur » et « Rapidité ». Cette vulgarisation permet à l'utilisateur d'identifier correctement et sans aucun doute la tâche à accomplir. De plus et comme indiqué dans les parties précédentes, le participant doit également choisir l'ordre dans lequel il entend le Son et l'Accord. Le panneau inférieur de l'interface permet au participant d'indiquer son propre delta en fonction de la condition choisie. Après

avoir cliqué sur le bouton « Lancer le test », l'utilisateur commence une session d'écoute d'un bloc constitué de 50 essais durant lesquels il entend les *stimuli* et doit indiquer la modification de fréquence perçue pour chacun d'eux.

Le tableau 3.5 présente la nomenclature utilisée pour désigner les réponses des participants en fonction de la modification de fréquence réalisée par le programme.

Tableau 3.5 – Définition du vocabulaire utilisé pour le calcul des métriques d'évaluation en fonction de la modification réalisée par le programme et du choix de l'utilisateur.

Modification réalisée par le programme	Réponse « Monte »	Réponse « Baisse »
Augmentation	$AugMonte$	$AugBaisse$
Diminution	$DimMonte$	$DimBaisse$

Afin d'évaluer les capacités de détection des changements par les participants, deux métriques sont calculées à partir des réponses fournies. La première métrique calcule le pourcentage de réussite P_r défini dans l'équation 3.4.

$$P_r = \frac{AugMonte + DimBaisse}{AugMonte + DimBaisse + AugBaisse + DimMonte} \tag{3.4}$$

La deuxième métrique est l'index de sensibilité *d'* [Macmillan et Creelman, 2004] défini dans la formule 3.5.

$$d' = Z(\frac{AugMonte}{AugMonte + AugBaisse}) - Z(\frac{DimMonte}{DimMonte + DimBaisse}) \tag{3.5}$$

où Z(x) est l'inverse de la fonction de répartition de la loi Normale, avec x $\in$ [0 ;1] [Bonnet, 1986].

L'index de sensibilité *d'* est sans dimension et mesure les capacités perceptives d'une personne à détecter un événement particulier produit dans un signal sonore. L'événement à détecter dans cette expérience est un changement de fréquence. Il est primordial de noter que la valeur de *d'* est indépendante d'éventuels biais de réponse des participants, contrairement au pourcentage de réussite. Par ailleurs, une valeur de *d'* proche de zéro indique que le participant répond aléatoirement alors qu'une augmentation de la valeur de *d'* indique une augmentation de la capacité du participant à percevoir correctement les changements de fréquence.

Les résultats des participants en termes de pourcentages de réussite et de valeurs de *d'* sont décrits dans la section qui suit.

Résultats

Les résultats obtenus suite aux entraînements ne sont pas détaillés puisque l'ensemble des participants a atteint 90% de réussite lors de la réalisation de deux blocs consécutifs et que cette performance ne peut pas être imputée au hasard.

La figure 3.5 représente le pourcentage de réponses correctes et la valeur du *d'* moyennés sur l'ensemble des participants en fonction des conditions d'ordre entre l'Accord et le Son.

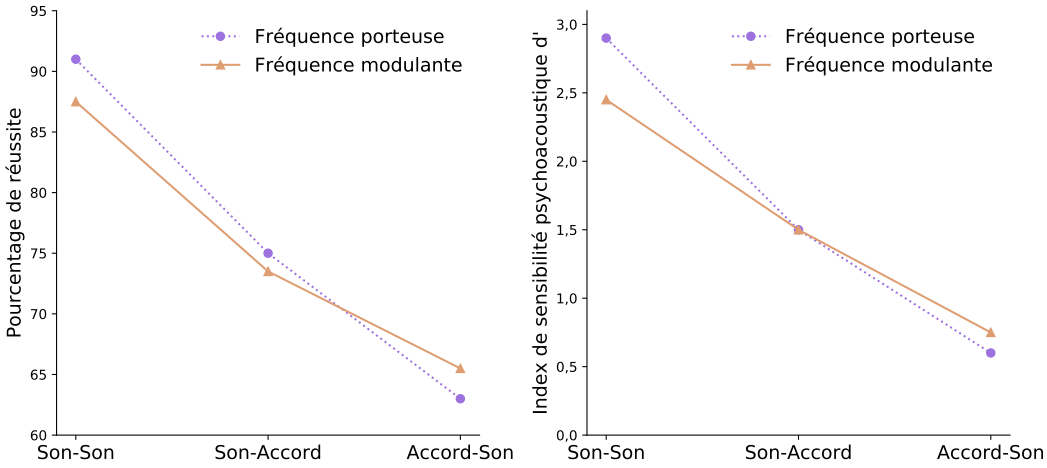


FIGURE 3.5 – Pourcentage de réussite (à gauche) et index de sensibilité psychoacoustique *d'* (à droite) des participants en fonction de la condition contrôle Son-Son, de l'ordre Son-Accord et de l'ordre Accord-Son.

La figure 3.5 montre un pourcentage de réussite moyen –sur les deux types de modification des fréquences– des participants dans la condition Son-Son de 89%. Dans la condition Son-Accord, le pourcentage de réponses correctes diminue de 14 points par rapport à la condition de référence Son-Son. De même, le pourcentage de réussite moyen dans la condition Accord-Son diminue de 24 points par rapport à la condition Son-Son.

L'index de sensibilité *d'* est en moyenne –sur les deux types de modifications des fréquences– de 2,7 pour tous les participants pour la condition Son-Son. Cette sensibilité est presque divisée par deux dans la condition Son-Accord et divisée par quatre dans la condition Accord-Son. En ce qui concerne l'ordre Son-Accord, il existe une faible différence entre les performances de détection de changements de fréquence porteuse ou modulante. Ceci est la conséquence de l'ajustement des performances décrit précédemment. De même, il existe une faible différence de la valeur de *d'* dans la condition d'ordre Accord-Son pour la détection d'une modification des fréquences porteuse ou modulante. La figure 3.6 détaille plus finement ces résultats puisqu'elle représente les per-

formances obtenues pour un changement de fréquence modulante en fonction des valeurs de fréquences de modulation utilisées.

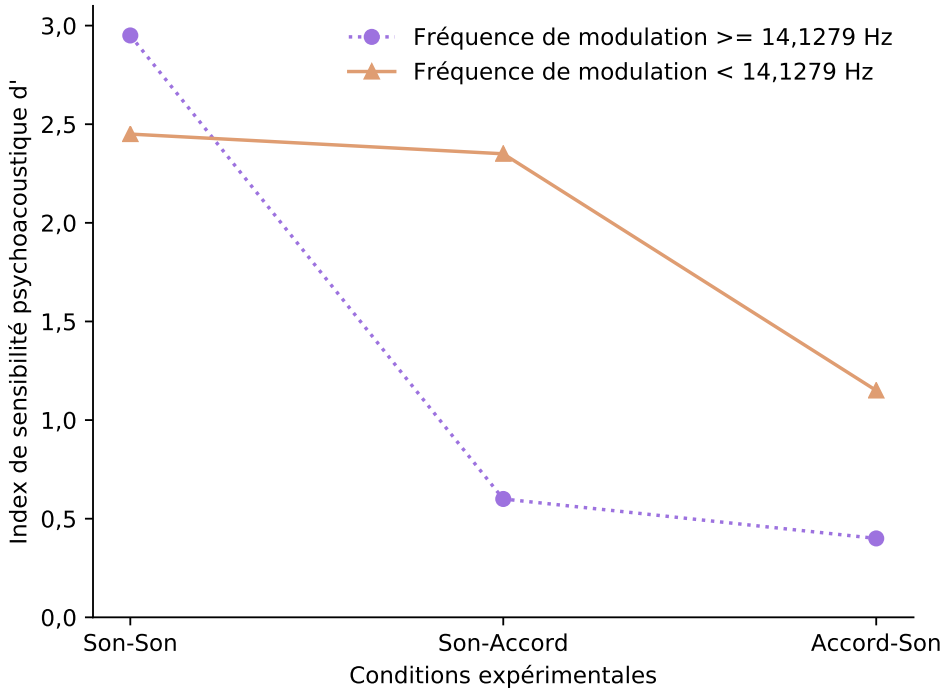


FIGURE 3.6 – Index de sensibilité *d'* en fonction de la condition d'ordre et de l'étendue des fréquences de modulation utilisées. L'analyse des fréquences de modulation est réalisée sur deux plages de fréquences séparées par la moyenne géométrique.

Il y a peu de différences entre les performances obtenues pour les conditions Son-Son et Son-Accord pour les fréquences basses de modulation. Les performances des fréquences de modulation élevées sont en revanche divisées par six entre les conditions Son-Son et Son-Accord. Il y a peu de différences dans les performances des conditions Son-Accord et Accord-Son pour les fréquences de modulation élevées alors que les performances pour les fréquences basses sont divisées par deux par rapport aux fréquences élevées.

Discussion

Cette expérience recherche l'existence de détecteurs de changement de rythme chez l'Homme en les comparant aux détecteurs de changement de fréquence déjà mis en évidence chez ce dernier. L'existence des détecteurs de changement de fréquence est avérée [Demany et Ramos, 2005] et leur mani-

festation se traduit par des performances similaires dans les conditions Son-Accord et Accord-Son. Or, les résultats de l'expérience montrent une différence de perception entre les ordres Son-Accord et Accord-Son dans la détection d'un changement de fréquence porteuse. Ceci indique que les détecteurs de changement de fréquence ne se manifestent pas dans cette expérience ou qu'ils sont supplantés par un mécanisme auditif inconnu. Cette absence de manifestation des détecteurs de changement de fréquence peut provenir du fait que l'existence de ceux-ci a été prouvée en utilisant des *stimuli* composés de sons stationnaires, ce qui n'est pas le cas ici avec la modulation d'amplitude. L'expérience ne permet donc pas de comparer les détecteurs de changement de fréquence aux détecteurs de changement de rythme et donc de confirmer l'existence de ces derniers. Cette expérience prouve néanmoins que l'attention favorise grandement la détection d'un changement de fréquence de modulation.

En outre, un effet remarquable transparaît suite à l'analyse des fréquences utilisées. Lors de l'écoute d'une fréquence de modulation basse, il est en effet aussi aisé de suivre son évolution lorsqu'elle est seule que lorsqu'elle est présentée au sein d'un Accord. Les participants ont en revanche montré des difficultés de détection d'un changement de fréquence de modulation élevée. Les MDI, qui constituent les interférences dans la distinction de fréquences de modulation, n'ont donc pas d'impact sur la détection des fréquences basses alors qu'elles impactent la détection des fréquences élevées.

L'expérience qui suit propose une nouvelle approche qui a pour objectif d'étudier la sensibilité des auditeurs à un changement de fréquence de modulation d'amplitude.

3.1.4 Étude de la perception d'un changement de rythme

> *« Je voye un abysme de science... Mais science sans conscience n'est que ruine de l'ame. »*
>
> – *Pantagruel*, 1554, François Rabelais

Les constats proposés par l'expérience précédente ne permettent pas de valider ni de réfuter l'existence de détecteurs automatiques de changement de rythme. La question de l'existence de ces détecteurs ainsi que de leurs caractéristiques, s'ils existent, est toujours soulevée dans cette section. Une nouvelle expérience est proposée qui utilise des *stimuli* différents afin de rechercher un indice de l'existence des détecteurs de changement de rythme dans l'oreille humaine.

Deux aspects différents sont étudiés afin de caractériser plus précisément la sensibilité auditive à un changement de rythme. Ce changement de rythme est concrétisé par un changement de fréquence de modulation.

Le premier aspect concerne la différence qui peut exister entre la détection d'une augmentation et d'une diminution de la fréquence de modulation.

Si le participant est attentif à une modification de rythme, l'hypothèse est qu'il est peu probable qu'une telle différence puisse exister. Cependant, si des mécanismes de détection automatique tels que les détecteurs de changement de rythme existent, alors les résultats de l'expérience montreront une asymétrie de perception entre les augmentations et les diminutions de la vitesse du rythme.

L'impact d'une rupture temporelle *via* un silence au sein de la modulation constitue le second aspect étudié. L'ajout d'une rupture temporelle *via* un silence à l'instant du changement de rythme prévient l'auditeur de l'instant exact de la modification rythmique. L'hypothèse envisagée considère qu'un auditeur doit pouvoir détecter plus aisément les changements de rythme en présence d'un tel indice. Toutefois, un silence provoque une rupture de rythme et la mémoire à court terme de l'auditeur est donc mise en jeu afin de mémoriser tout changement de rythme. L'hypothèse posée implique que cette rupture temporelle entraîne une diminution des performances de détection d'un changement de rythme. S'il existe des détecteurs de changement de rythme chez l'Homme, alors les performances lors de l'expérience montreront que les participants détectent mieux un changement de rythme lorsqu'aucune rupture temporelle n'est présentée dans le *stimulus*. De plus, lorsqu'un silence est présenté au cours du *stimulus*, l'ensemble des Sons subit une rupture identique, ce silence devrait donc perturber l'activation des détecteurs de changement de rythme. En revanche, lorsqu'aucun silence n'est présenté lors du changement rythmique, l'intégralité des Sons de l'Accord poursuit sa progression tandis que le Son modifié subit une rupture temporelle. Une telle rupture devrait être aisément détectable par des détecteurs de changement de rythme.

Matériels et méthodes

L'objectif consiste à évaluer la capacité des participants à identifier la fréquence de modulation qui est modifiée parmi les cinq fréquences présentes dans un Accord. La différence de performance chez les participants en fonction des accélérations et des ralentissements rythmiques est évaluée.

Chaque *stimulus* est composé de deux Accords successifs puis d'un Son Pur. Les Accords sont chacun composés de cinq Sons. Le Son Pur a une fréquence identique à l'une des fréquences porteuses de l'un des cinq Sons de l'Accord.

Une condition est ajoutée à l'expérience afin d'identifier l'impact d'une rupture rythmique lors du *stimulus*. Le participant sélectionne en effet lors de chaque initiation de bloc s'il souhaite entendre deux Accords séparés par un silence de 100 ms ou si la transition entre ces Accords doit être continue.

La durée de l'Accord est de 2 500 ms, un silence de 100 ms sépare le dernier Accord du Son Pur s'il est choisi par le participant et le Son Pur a une durée de 500 ms.

Le delta de modification de la fréquence modulante est fixé à 600 cents

pour l'ensemble des participants.

La figure 3.7 représente le spectrogramme d'un exemple de *stimulus*, c'est-à-dire l'évolution des fréquences d'un signal sonore au cours du temps.

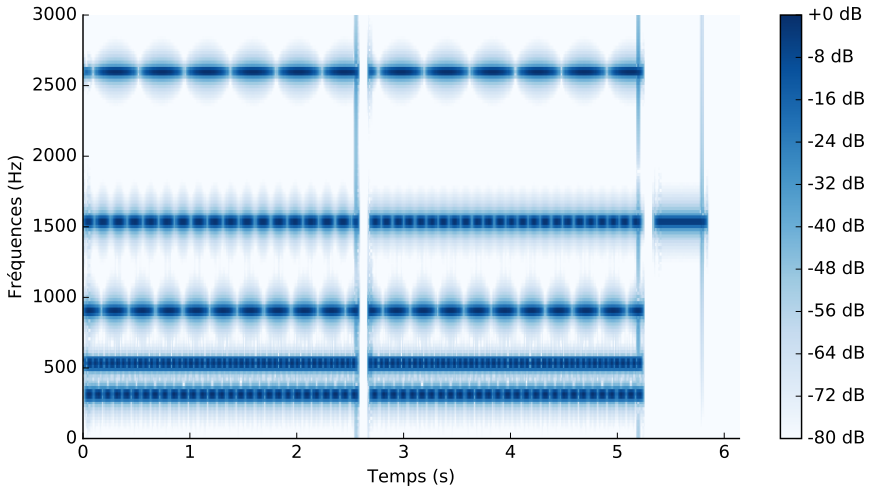


FIGURE 3.7 – Spectrogramme d'un exemple de *stimulus* utilisé lors de l'expérience.

Les paramètres utilisés afin d'obtenir ce spectrogramme sont les suivants :

- Un Son Pur après les deux Accords,
- Un silence entre les deux Accords,
- Un rythme plus rapide dans le second Accord.

Dans cet exemple, le quatrième Son en partant du bas de l'Accord a une fréquence porteuse proche de 1 500 Hz et est celui qui subit le changement de rythme. Dans le premier Accord, ce quatrième Son a un rythme plus lent que dans le second Accord. Ce changement rythmique du quatrième Son est visible sur le spectrogramme puisque la longueur des stries est inférieure pour le second Accord par rapport au premier. Ce changement est également visible dans les 18 stries qui caractérisent le quatrième Son du premier Accord alors que 24 stries caractérisent le quatrième Son du second Accord. Le Son Pur a une fréquence porteuse similaire à celle du quatrième Son de l'Accord qui subit la modification rythmique. Dans cet exemple, le participant doit donc confirmer que le Son Pur a la même fréquence porteuse que le Son des Accords qui a subi un changement de rythme.

Le programme décrit pour l'expérience précédente est réutilisé pour cette expérience. Cependant, trois nouveaux choix sont proposés aux participants et ajoutés à l'interface graphique, à savoir :

- L'ordre relatif entre le Son Pur et les deux Accords (Son Pur-Accord-Accord ou Accord-Accord-Son Pur),
- La durée de la transition entre les deux Accords (0 ou 100 ms),
- La nature de la modification de rythme (Accélération ou Ralentissement).

Résultats

L'analyse de la valeur du *d'* est présentée dans la figure 3.8.

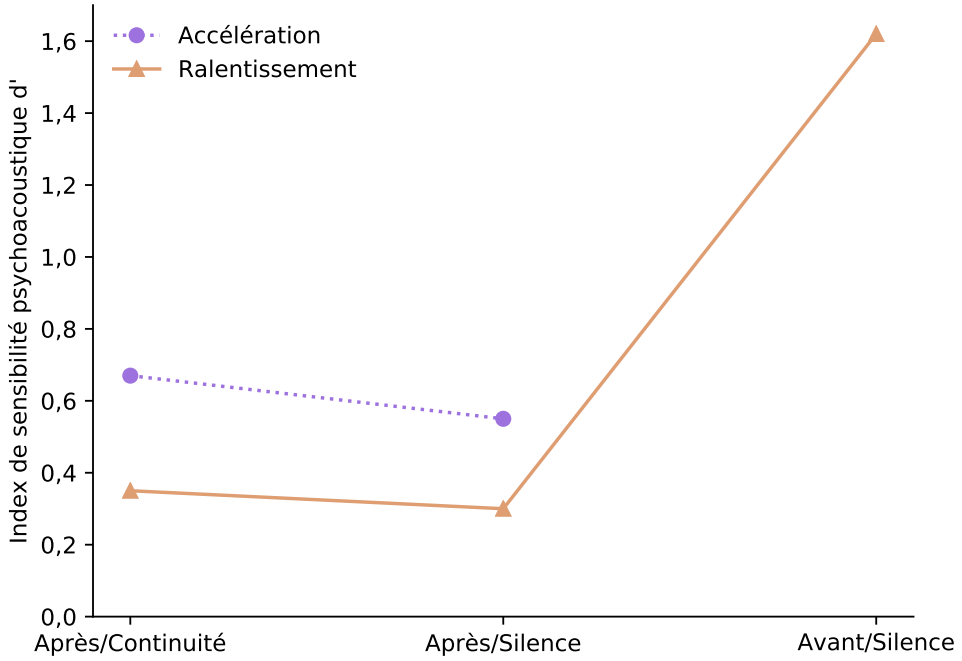


FIGURE 3.8 – Analyse de l'index de sensibilité *d'* en fonction de la position du Son Pur Avant ou Après les deux Accords et de la condition Continuité ou Silence. Dans la condition de contrôle Avant/Silence, une valeur de *d'* est affichée pour un Ralentissement mais l'expérience n'a pas été réalisée pour une Accélération. Des résultats similaires sont en effet attendus dans le cas d'un Ralentissement et d'une Accélération et cette condition de contrôle permet uniquement d'avoir une référence de la valeur de *d'* lorsque le Son Pur est entendu avant les deux Accords.

Lorsque le Son Pur est positionné avant les deux Accords, la valeur du *d'* est multipliée par 3,5 par rapport à la condition où le Son Pur est positionné après les Accords. Ce résultat est obtenu en comparant les performances du parti-

cipant pour la condition Son Pur-Accord-Accord avec celles pour la condition Accord-Accord-Son Pur.

Les performances obtenues par les participants lorsque la fréquence de modulation augmente –donc quand le rythme est plus rapide– dans le second Accord sont deux fois supérieures à celles obtenues lorsque cette fréquence diminue. Afin de mieux comprendre la différence de valeur de *d'* pour des accélérations et des ralentissements, la figure 3.9 représente le détail de la valeur de *d'* en fonction des fréquences de modulation.

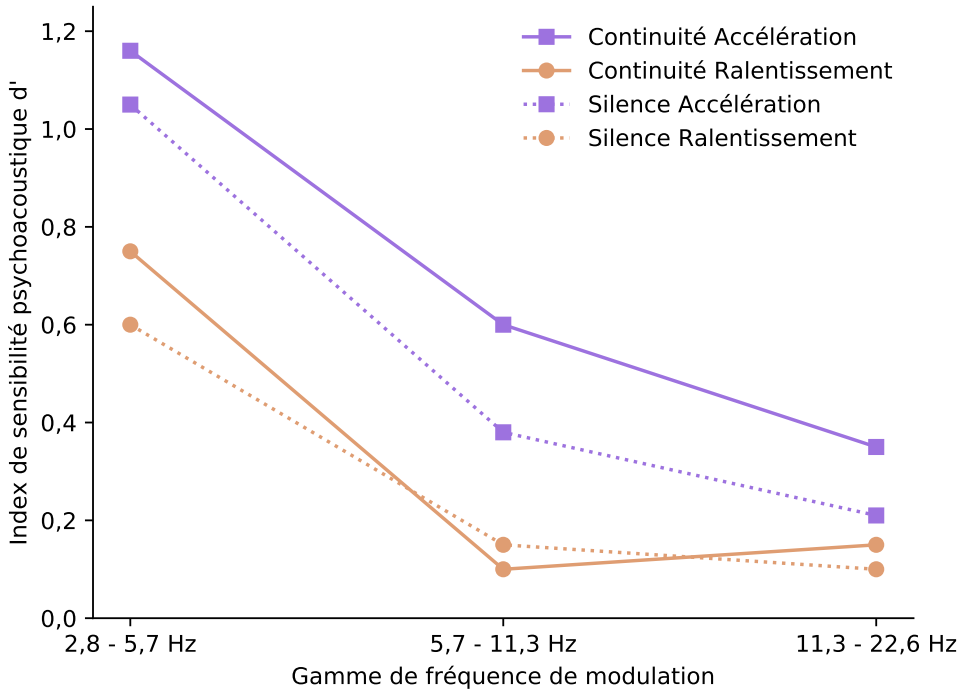


FIGURE 3.9 – Analyse de l'index de sensibilité *d'* en fonction du registre de la fréquence de modulation pour la condition Accord-Accord-Son Pur.

À l'instar des résultats présentés dans la figure 3.8, la valeur du *d'* est supérieure dans le cas d'une Accélération par rapport à un Ralentissement du rythme dans le second Accord. Le second résultat remarquable concerne la valeur du *d'* qui est, pour les quatre conditions, plus importante pour des fréquences de modulation inférieures à 5,7 Hz que pour des fréquences de modulation comprises entre 5,7 et 22,6 Hz.

Discussion

Tout comme décrit pour la première expérience, les différences de résultats dans l'ordre de présentation du Son Pur par rapport aux Accords montrent que l'attention favorise grandement la détection d'un changement de rythme. Cette dernière observation constitue un premier indice en faveur de l'inexistence des détecteurs de changement de rythme chez l'être humain.

Les faibles performances obtenues pour la détection d'un ralentissement de rythme et pour les fréquences de modulation élevées indiquent par ailleurs la présence d'un effet plancher, ce qui signifie que le système auditif humain n'est pas adapté à détecter ce type de changement. La faible sensibilité auditive à des rythmes supérieurs à 12 Hz pourrait néanmoins être expliquée par le fait que la voix humaine est composée de rythmes évoluant entre 0,6 et 12 Hz [Payton et Braida, 1999].

La faible différence des performances entre les conditions continuité et silence fournit un second argument en faveur de l'inexistence des détecteurs de changement de rythme chez l'être humain. Un détecteur automatique de changement de rythme serait en effet beaucoup plus stimulé par un rythme brutalement modifié que par un rythme modifié après un silence.

Par ailleurs, l'asymétrie observée dans les performances entre la détection d'une accélération et d'un ralentissement constitue un effet qui n'a jamais été décrit dans la littérature. Trois hypothèses sont proposées afin d'expliquer les causes de cette asymétrie.

La première hypothèse propose un rapprochement entre l'asymétrie observée dans l'expérience avec celle rapportée par Tajadura-Jiménez *et al.* [2011] qui montre que l'Homme est plus sensible à un son en approche plutôt qu'à un son qui s'éloigne. L'explication fournie pourrait constituer un avantage évolutif qui impliquerait qu'on son en approche constitue un avertissement de l'arrivée d'un prédateur. Cependant, de nouvelles études devraient être menées afin de confirmer ou d'infirmer cette hypothèse.

La deuxième hypothèse implique le delta utilisé. Ce dernier, fixé à 600 cents, est identique pour l'ensemble des participants, ce qui implique que l'écart entre les fréquences lors d'une accélération ou d'un ralentissement est différent en termes de Hertz. Pour une fréquence de modulation initiale de 10 Hz par exemple, une augmentation de 600 cents produit en effet une fréquence de 14,14 Hz alors qu'une diminution d'autant de cents fournit une fréquence de 7,07 Hz. Dans le cas d'une accélération, une différence de fréquence de 4,14 Hz est donc obtenue alors que pour un ralentissement cette différence est de 2,93 Hz. Cette observation tend à montrer que l'oreille humaine détecte d'autant mieux une modification de rythme que celle-ci est importante. L'asymétrie des performances entre accélération et ralentissement pourrait être considérée comme étant la conséquence de cette différence qui, de plus, augmente avec la fréquence initiale. Une telle considération est néanmoins incorrecte puisqu'elle

impliquerait que les performances des participants soient meilleures lors de la détection de rythmes rapides que lorsque ceux-ci sont plus lents, ce qui n'est expérimentalement pas le cas.

Une dernière hypothèse permet d'expliquer partiellement l'asymétrie des performances obtenues lors des détections d'accélération et de ralentissement de rythmes. Le colliculus inférieur constitue un ensemble de neurones du cerveau qui permettent la représentation mentale des rythmes [Schreiner et Langner, 1988]. Or, une accélération et une décélération rythmiques provoquent respectivement une augmentation et une diminution de l'activité neuronale du colliculus inférieur. De plus, l'Homme est plus sensible à une augmentation de l'activité neuronale qu'à sa diminution [Constantino *et al.*, 2012], ce qui semble entraîner une asymétrie de perception des rythmes. Les neurones du colliculus inférieur sont par ailleurs activés uniquement lorsque le participant n'est pas attentif au *stimulus* qu'il entend. Lors de l'écoute attentive d'un *stimulus* en revanche, aucune asymétrie de détection de rythme ne devrait subsister puisque d'autres zones du cerveau sont activées lorsque l'attention est mise en jeu. La présence d'une asymétrie de détection des rythmes pourrait par conséquent constituer un argument en faveur de l'existence des détecteurs de changement de rythme. Toutefois et au vu des résultats obtenus, cet argument est uniquement valide pour des rythmes lents. En dépit de cette dernière restriction, cette hypothèse semble actuellement être la plus probable parmi les trois proposées afin d'expliquer la présence de l'asymétrie entre la détection d'une accélération et celle d'un ralentissement de rythme.

3.1.5 Discussion sur les détecteurs automatiques de changement de rythme

> *« Once you get into the flow of things, you're always haunted by the way that things could have turned out. This outcome, that conclusion. You get my drift. The uncertainty is what holds the story together, and that's what I'm going to talk about. »*
>
> – *Rhythm Science*, 2004, Paul Miller

Les deux expériences proposées fournissent des informations relatives à la sensibilité auditive humaine et plus précisément à la capacité de l'oreille humaine à détecter des changements de fréquence de modulation et donc de rythme. Deux arguments en faveur de l'inexistence des détecteurs de changement de rythme ont été proposés, un troisième cependant maintient un doute quant à leur existence. De nouvelles interrogations ont été soulevées suite aux données des expériences, notamment en ce qui concerne l'interprétation de l'asymétrie de la sensibilité humaine à des rythmes qui accélèrent ou décélèrent.

Les connaissances apportées par ces expériences ne permettent actuellement pas de proposer de nouvelles caractéristiques audio. Les deux expériences menées offrent toutefois plusieurs pistes d'étude en psychoacoustique afin de caractériser plus en détail la perception des rythmes chez l'Homme. Afin d'explorer de telles pistes, une nouvelle expérience pourrait consister à soumettre la perception des participants à la succession de deux Accords dont les fréquences de modulation seraient échangées au sein de l'un des deux Accords. Une autre expérience permettrait par ailleurs d'étudier l'impact de la détection d'une modification de fréquence porteuse lorsqu'une modification de fréquence de modulation est également effectuée.

3.2 Des modèles psychoacoustiques pour l'extraction des caractéristiques audio

> *« Il n'y a pas de création ex nihilo, de saut brusque. Toute invention, toute innovation, n'est que la combinaison nouvelle d'éléments préexistants empruntés [...] aux techniques déjà connues. »*
>
> – *La Technologie, science humaine*, 1987, André-Georges Haudricourt

Les analyses psychoacoustiques jouent un rôle prépondérant dans la conception de caractéristiques audio pertinentes qui permettent de décrire les signaux musicaux. Le système auditif humain ne possède par exemple pas la même sensibilité à toutes les fréquences qui lui sont audibles. La sensibilité fréquentielle affiche en effet un pic autour de 1 kHz puis diminue au-delà de cette fréquence. Il s'est par ailleurs avéré que la modélisation de cette non-linéarité améliore la qualité des caractéristiques audio extraites du signal [Gaikwad *et al.*, 2010]. L'échelle mel [Stevens *et al.*, 1937] figure parmi les modélisations non-linéaires les plus utilisées en audio [Gaikwad *et al.*, 2010] et est définie par l'équation 3.6.

$$mel(f) = 1127 \ln(1 + \frac{f}{700}) \tag{3.6}$$

où f est la fréquence en Hertz.

En outre, il est possible de concevoir une méthode de détection du chant qui se fonde uniquement sur l'échelle mel afin d'extraire des caractéristiques audio d'un signal [Ghosal *et al.*, 2013; Gouyon *et al.*, 2014]. Toutefois, plusieurs autres caractéristiques audio existent et permettent de caractériser un signal de façon plus détaillée. La section suivante introduit des caractéristiques audio pertinentes à extraire des morceaux et notamment celles relatives à la classification des instrumentaux et des chansons.

3.2.1 Caractéristiques audio extraites à l'échelle locale

« Écrire, c'est l'art des choix, comme on dit à Privas. »
– Les pensées de San-Antonio, 1996, Frédéric Dard

L'étape d'extraction de caractéristiques musicales nécessite de sélectionner des informations pertinentes contenues dans un signal audio temporel. La modification de la phase des fréquences présentes dans un morceau modifie par exemple sa représentation temporelle sans en modifier le contenu musical. Bergstra *et al.* [2006b] détaillent quatre raisons qui soulignent les difficultés de l'analyse directe du signal audio sous sa forme temporelle :

- Un signal audio numérique brut contient en moyenne 44 100 valeurs de pression acoustique par seconde,
- Une méthode d'analyse devrait être robuste au décalage global d'un signal musical brut tel que celui dû au retard d'un échantillon,
- De légères différences présentes dans les composantes spectrales peuvent avoir un impact perceptif fort mais demeurer imperceptibles au sein du signal brut,
- De faibles changements globaux dans la fréquence d'un signal brut ne sont pas forcément perceptibles contrairement à un léger changement de fréquence d'un instrument.

Le signal audio brut enregistré dans un vecteur d'échantillons ne semble donc pas être la représentation la plus pertinente à utiliser lors de l'application de méthodes d'apprentissage machine. Il est donc préférable d'utiliser des techniques heuristiques qui permettent d'extraire des caractéristiques perceptivement représentatives du signal audio.

Par ailleurs, les méthodes de traitement du signal s'appliquent rarement à l'intégralité des quatre minutes d'un morceau. Elles sont en effet plutôt appliquées à des segments consécutifs de l'ordre de 30 ms appelés trames. Plusieurs raisons justifient le fait de traiter un signal en fonction des trames qu'il contient plutôt que de le considérer dans sa globalité :

- L'analyse en temps réel du signal,
- La quantification de l'évolution temporelle d'une caractéristique musicale,
- L'analyse suivant plusieurs échelles temporelles.

Toutefois, afin de caractériser un morceau dans sa globalité, le traitement du signal par trame requiert par la suite un processus d'agrégation qui est détaillé dans la section 3.2.2.

Les principales caractéristiques audio qui permettent la détection du chant dans un signal musical sont présentées ci-après. En ce qui concerne les caractéristiques audio utilisées pour caractériser d'autres thèmes musicaux, plusieurs

états de l'art ont été proposés [Aucouturier et Pachet, 2003; Orio, 2006; Casey *et al.*, 2008].

De nombreux travaux ont été menés afin de distinguer les trames contenant de la voix de celles purement instrumentales [Nwe *et al.*, 2004; Markaki *et al.*, 2008; Regnier et Peeters, 2009; Lehner *et al.*, 2014; Leglaive *et al.*, 2015; Lehner *et al.*, 2015]. Ces travaux, menés afin d'étudier la présence de chant dans les trames, ont montré que certains descripteurs pouvaient être particulièrement pertinents pour cette tâche [West et Cox, 2004].

Parmi l'ensemble de ces descripteurs, les Mel-Frequency Cepstral Coefficients (MFCC) [Rabiner et Juang, 1993] sont apparus comme étant les plus pertinents afin de détecter le chant dans une piste musicale [Foote, 1997; Eronen et Klapuri, 2000; West et Cox, 2004; Rocamora et Herrera, 2007; Lehner et Widmer, 2015]. Les MFCC contiennent les coefficients du « spectre du spectre » (appelé cepstre) affichés sur une échelle mel des fréquences. Les MFCC permettent donc de représenter de manière compacte les informations fréquentielles contenues dans une trame audio. Généralement, 10 à 20 coefficients suffisent à différencier un bruit de la parole, du chant ou de la musique. De plus, les MFCC présentent l'avantage d'être robustes à la dégradation de la qualité d'un signal audio [Urbano *et al.*, 2014].

Cependant, les MFCC seuls ne fournissent pas d'information quant à la direction d'évolution fréquentielle dans le temps du contenu musical. Afin de représenter l'évolution des MFCC, le delta est couramment utilisé et calcule la dérivée du premier ordre entre les trames [Furui, 1986]. La dérivée du second ordre est également couramment utilisée et est notée double delta [Livshin et Rodet, 2009].

Puisque l'extraction de caractéristiques audio constitue une composante essentielle de nombreux systèmes d'analyse musicale, il est par conséquent important d'assurer la réplicabilité de cette étape. Dans ce cadre, plusieurs logiciels ont été proposés et Moffat *et al.* [2015] ont comparé les performances des 10 outils les plus utilisés et cités.

3.2.2 Processus d'agrégation des caractéristiques audio locales

> *« The only difference between screwing around and science is writing it down. »*
>
> – Émission *MythBusters*, 2012, Alex Jason (popularisation par Adam Savage)

Les caractéristiques musicales sont extraites à l'échelle de la trame audio, c'est-à-dire pour chaque trame d'un morceau. Ces caractéristiques permettent de décrire et d'annoter un morceau à l'échelle locale mais ne peuvent être utilisées telles quelles pour annoter la globalité d'un morceau. Afin de déterminer si

un morceau est un instrumental ou une chanson il est donc possible d'agréger ces caractéristiques locales afin d'en générer de nouvelles à l'échelle globale. Ces caractéristiques globales sont ensuite utilisées par des algorithmes d'apprentissage machine afin de classer les morceaux en fonction des annotations qui les caractérisent et de générer des listes de lecture musicale thématiques.

État de l'art

Ghosal *et al.* [2013] ont proposé une méthode d'agrégation qui calcule la moyenne des caractéristiques audio locales afin de fournir des caractéristiques globales, c'est-à-dire à l'échelle du morceau. Pour cela, leur algorithme calcule tout d'abord 13 MFCC par trame puis les moyenne afin d'obtenir 13 MFCC pour chaque morceau.

Afin de parvenir au même objectif, Gouyon *et al.* [2014] ont quant à eux proposé d'ajouter l'écart type des caractéristiques audio locales à la moyenne des caractéristiques audio locales. Néanmoins, le calcul de la moyenne et de l'écart type des MFCC effectué par les méthodes de Ghosal *et al.* [2013] et Gouyon *et al.* [2014] semble restreindre les informations contenues dans les trames audio [Bayle *et al.*, 2018b]. Une importante dynamique de chant sur de courts instants du morceau ne serait en effet pas décelée sur la moyenne ou l'écart type des MFCC puisque les fréquences du chant sont proches de celles d'instruments comme le violon par exemple.

Afin de conserver cette dynamique, la méthode proposée par Langlois et Marques [2009] et améliorée par Gouyon *et al.* [2014] utilise une chaîne de Markov du premier ordre en tant qu'agrégateur de caractéristiques locales. Cette méthode modélise les probabilités d'observer une séquence donnée de caractéristiques audio pour chaque classe. Cette modélisation probabiliste semble toutefois être restreinte lorsque de nouveaux morceaux lui sont présentés qui affichent des séquences non observées auparavant ou qui sont présentes avec un ratio différent dans un morceau.

Afin de pallier les problèmes des méthodes existantes, deux nouvelles méthodes d'agrégation de caractéristiques audio locales ont été proposées [Bayle *et al.*, 2016, 2018b] et sont détaillées dans les sections suivantes.

Agrégation fondée sur la probabilité de présence de la voix chantée

Afin de créer une liste de lecture musicale instrumentale, il faut tout d'abord identifier les instrumentaux et les chansons dans une base de données. Or, ces types de morceaux contiennent tous deux des sons produits par des instruments. Le seul moyen de distinguer une chanson d'un instrumental consiste donc à identifier la présence de chant dans le morceau. De plus, le fait de détecter une seule trame contenant du chant suffit à affirmer qu'un morceau est une chanson. Il est en revanche nécessaire de vérifier l'absence de chant dans l'intégralité des trames d'un morceau avant de pouvoir affirmer qu'il s'agit

d'un instrumental. Dans le cas idéal, une méthode peut identifier correctement toutes les trames en fonction de la présence de chant et il est donc uniquement requis de vérifier la présence d'une trame contenant du chant pour affirmer qu'un morceau est une chanson. Puisque les approches qui tentent de prédire la présence de voix à l'échelle de la trame ne sont actuellement pas fiables à 100%, cette section propose une stratégie de prise de décision fondée sur les prédictions par trames et qui prend en compte ce manque de précision.

La première étape de cette méthode nécessite d'avoir à disposition des annotations indiquant la présence ou l'absence de chant à l'échelle de la trame pour chaque morceau. Ces annotations à l'échelle de la trame permettent au modèle de classification de distinguer efficacement les trames vocales de celles qui contiennent des instruments perceptivement proches du chant humain. Les deux bases de données musicales utilisées afin de fournir ces annotations sont *Jamendo* et *RWC*, qui ont été présentées dans la section 2.3.1. Pour chaque trame de 200 ms, les annotations de présence du chant ont été générées à la main par Ramona *et al.* [2008] pour les morceaux de *Jamendo* et par Bernhard Lehner [1] pour les morceaux de *RWC*. Un modèle de classification par trame est ensuite généré à partir de ces annotations. Pour cela, les annotations concernant chacune des trames sont utilisées afin de constituer une base d'apprentissage à l'échelle de la trame. Le logiciel Weka [Hall *et al.*, 2009; Frank *et al.*, 2016] permet de générer un tel modèle d'apprentissage. Il est donc possible de prédire la présence de chant dans chaque trame de chaque morceau de la base de données à l'aide du modèle généré par Weka. Pour chaque trame de chaque morceau, Weka fournit en effet une prédiction comprise entre 0 et 1 qui indique la probabilité pour qu'une trame contienne de la voix. Pour chaque morceau, un fichier contenant le vecteur des prédictions de l'ensemble des trames est créé. Dans un cas idéal, aucune trame d'un morceau instrumental n'est détectée comme contenant du chant, ce qui n'est pas le cas expérimentalement. L'hypothèse envisagée par la méthode proposée prend donc en compte le manque de précision de l'algorithme de détection du chant à l'échelle de la trame. Pour ce faire, la méthode présentée dans cette section utilise le vecteur de prédictions de chaque morceau contenant les probabilités de présence de voix pour chaque trame. Cette méthode suppose ensuite qu'il existe un ratio de trames prédites comme contenant du chant qui est différent pour les chansons par rapport aux instrumentaux.

Pour chaque morceau, un histogramme de dimension N est généré qui représente la répartition linéaire des probabilités de présence du chant équitablement répartie entre 0,5 et 1. Par construction, les probabilités inférieures ou égales à 0,5 ne sont pas prises en compte. Cette probabilité de 0,5 constitue en effet le seuil à partir duquel la présence de chant est considérée. La valeur de N a un impact sur la précision de classification des instrumentaux comme

1. Communication personnelle

indiqué dans la figure 3.10.

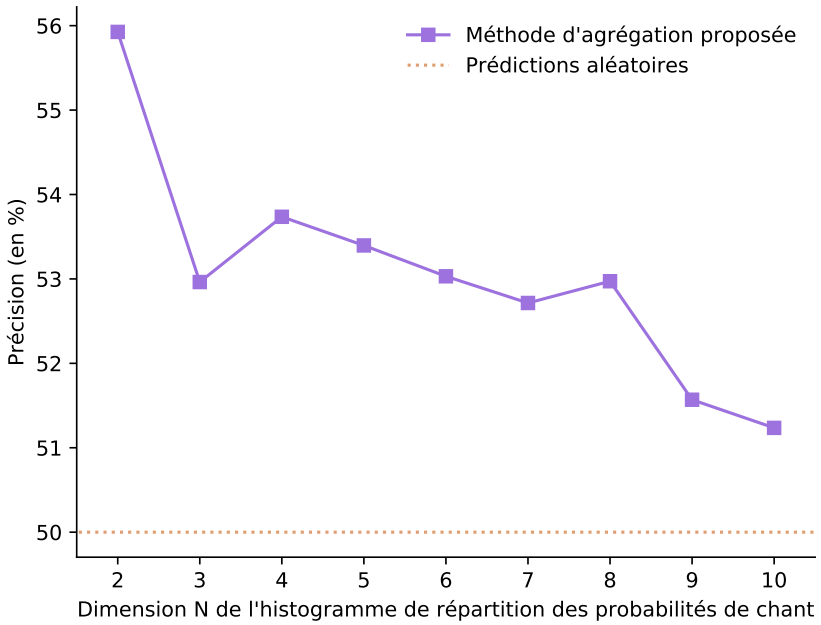


FIGURE 3.10 – Précision de classification des instrumentaux en fonction de la valeur de N. La valeur de N représente la dimension de l'histogramme de prédictions des trames contenant du chant.

La figure 3.10 montre que la meilleure précision de classification des instrumentaux est obtenue en utilisant un histogramme à 2 valeurs (N = 2). Cette précision est comparée à ProbaH2-10, qui utilise le vecteur de concaténation des histogrammes avec chacune des valeurs de N de 2 à 10 et qui affiche 54 dimensions. Pour N = 2, la précision obtenue pour la classification des instrumentaux est de 55,9% alors qu'elle est de 67,4% pour ProbaH2-10. La précision obtenue avec ProbaH2-10 est encourageante mais ne dépasse celle des prédictions aléatoires que de 6 points au maximum. Il est donc préférable de rechercher une méthode complémentaire de prise de décision qui utilise différemment le vecteur de prédictions de chaque morceau. La section suivante propose une autre méthode d'agrégation des prédictions de présence de chant dans les trames.

Agrégation utilisant les trames vocales consécutives

Cette section propose une nouvelle méthode d'agrégation à partir du vecteur de prédictions des morceaux, qui est détaillée dans la figure 3.11.

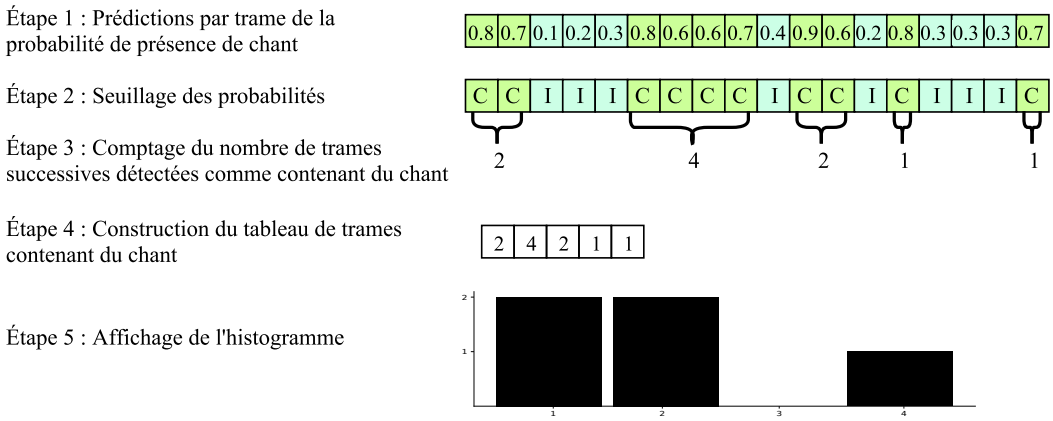


FIGURE 3.11 – Processus d'agrégation des prédictions des trames afin de constituer une caractéristique audio haut-niveau. Ce processus s'inspire du codage par plages, ou *run-length encoding*, qui est un algorithme de compression de données.

La méthode proposée sélectionne uniquement les trames pour lesquelles la probabilité de présence du chant est supérieure à 50%. Par construction, les trames affichant une probabilité de présence de chant inférieure à 50% sont considérées comme étant des trames instrumentales. Les trames instrumentales n'apportent en effet pas d'information discriminante en ce qui concerne les chansons et les instrumentaux puisque ces deux types de morceaux contiennent des parties instrumentales. Les trames qui ne sont pas censées contenir de chant ne sont donc pas utilisées puisque la présence d'instruments n'est pas corrélée à la présence de chant. Pour chaque morceau, un vecteur est créé afin de comptabiliser le nombre de fois consécutives pour lesquelles une trame a été identifiée comme contenant de la voix. Il est supposé ici qu'un morceau musical contenant du chant présente davantage de trames vocales consécutives qu'un morceau strictement instrumental. Un histogramme de ce vecteur est ensuite généré en accord avec cette hypothèse. La figure 3.12 présente un exemple d'histogrammes des vecteurs obtenus pour un instrumental et pour une chanson.

Les histogrammes correspondant aux vecteurs obtenus pour des chansons possèdent davantage de longues suites de trames identifiées comme contenant du chant que n'en présentent les instrumentaux. Il est donc possible d'utiliser les valeurs des histogrammes en tant que descripteurs afin de différencier les instrumentaux des chansons.

Sous cette forme, la taille du vecteur varie avec chaque morceau. Or, afin de pouvoir comparer les morceaux entre eux à partir de leurs caractéristiques audio, ces caractéristiques doivent posséder les mêmes dimensions. La dimension de l'histogramme est donc fixée à 30 valeurs par construction. Les suites de plus de 30 trames vocales consécutives sont ajoutées à la comptabilisation

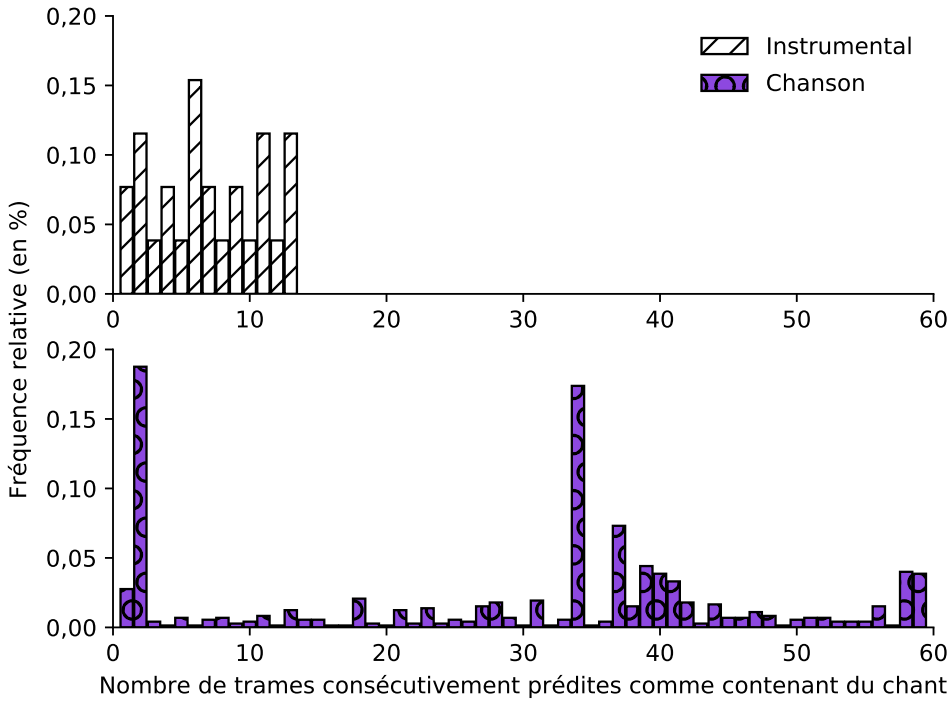


FIGURE 3.12 – Exemples d'histogrammes des vecteurs d'estimation du chant obtenus pour un instrumental et une chanson.

de celle pour 30 valeurs. Ce descripteur qui utilise un encodage des trames à 30 dimensions est par la suite noté ET30.

Un descripteur D à une dimension est également utilisé et est défini dans l'équation 3.7. Ce descripteur D est le rapport du nombre de fois où plus de i trames consécutives ont été identifiées comme contenant du chant sur le nombre de fois où moins de i trames consécutives l'ont été. La valeur du descripteur D est généralement nulle pour un instrumental et généralement positive si le morceau contient de la voix.

$$D = \frac{\sum_{i=1}^{R-1} r_i}{\sum_{i=R}^{30} r_i} \tag{3.7}$$

où

- R est le seuil du nombre de rangs à considérer,
- r_i est le nombre de fois pour lesquelles i trames consécutives ont été identifiées comme contenant du chant.

Le choix du rang R impacte la précision de classification, comme l'illustre la figure 3.13.

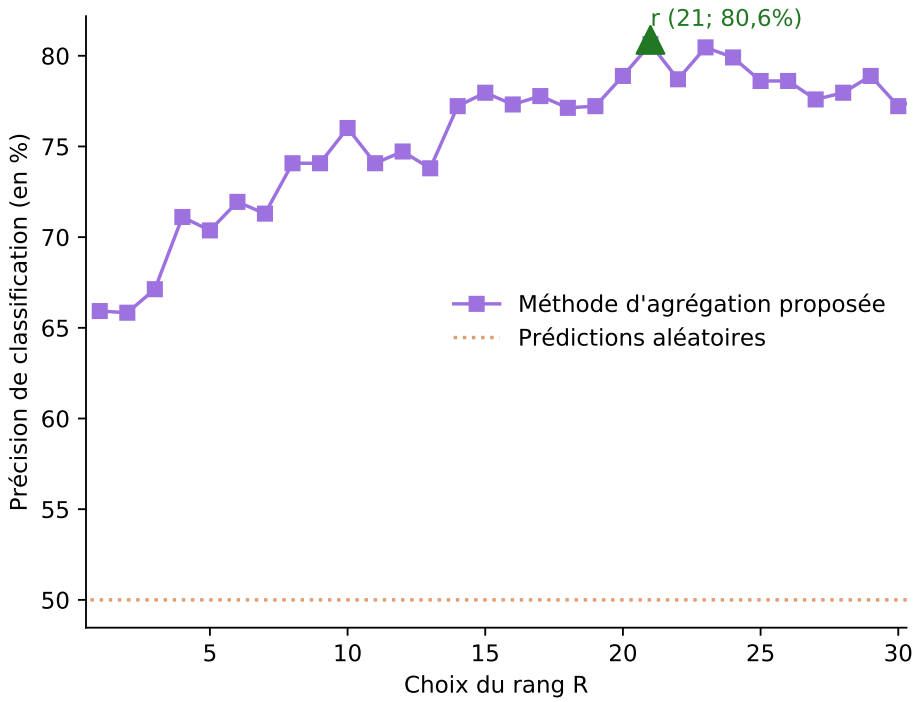


FIGURE 3.13 – Évolution de la précision du résultat de classification des chansons et des instrumentaux en fonction du rang R. Les morceaux proviennent de *ccMixter*, *MedleyDB*, *QUASI*, *RWC*, *Jamendo* et d'une bibliothèque personnelle.

La figure 3.13 indique que la courbe présente un maximum de précision de 80,6% lorsque R = 21. Ce descripteur est par la suite noté ETR21. Un autre descripteur est proposé, noté ETRAll, qui concatène les vecteurs de tous les rangs R de telle sorte que $ETRAll = [ETR1, ETR2, ..., ETR30]$.

Résultats de l'expérience de comparaison des méthodes d'agrégation de caractéristiques audio locales

Les méthodes proposées analysent des aspects différents d'un même morceau. La méthode ProbaH2-10 fournit une analyse de la présence de chant à l'échelle d'une trame tandis que la méthode ETR21 se fonde sur l'étude des blocs de trames . Il apparaît donc pertinent d'étudier l'impact de la concaténa-

tion des caractéristiques de ces deux méthodes sur les résultats de classification des chansons et des instrumentaux. Les méthodes d'apprentissage machine utilisées ne sont pas détaillées puisqu'elles le seront dans le chapitre 4 et que leur impact n'est pas l'objet de cette expérience.

Le tableau 3.6 indique les résultats de classification obtenus par ProbaH2-10, ET30, ETR21 et ETRAll pour la classification des instrumentaux. Le choix de ne considérer que les instrumentaux est motivé par le fait que ce thème est plus complexe à détecter que les chansons [Bayle *et al.*, 2018b].

Tableau 3.6 – Comparaison du pourcentage de précision de classification des instrumentaux obtenu pour cinq méthodes d'agrégation différentes face à une méthode de prédiction aléatoire. La valeur en gras indique la meilleure précision atteinte parmi toutes les précisions des méthodes référencées.

Méthode	Précision (en %)
Aléatoire	50,0
ProbaH2-10	67,4
ET30	87,3
ETR21	80,7
ETRAll	85,9
[ProbaH2-10, ET30]	88,0
[ProbaH2-10, ETRAll]	**89,3**

Les résultats de classification obtenus lorsque les caractéristiques audio de chaque méthode sont concaténées sont supérieurs à ceux obtenus lorsque les caractéristiques audio extraites par chaque méthode sont utilisées individuellement.

Les résultats de classification obtenus en utilisant les caractéristiques audio de la méthode ET30 affichent seuls une meilleure précision que ceux de ETRAll. Toutefois, la précision obtenue par la concaténation des caractéristiques audio pour le couple [ProbaH2-10, ETRAll] est légèrement supérieure à celle du couple [ProbaH2-10, ET30]. Les méthodes ProbaH2-10 et ETRAll semblent donc analyser des éléments différents des morceaux et leurs caractéristiques audio peuvent être concaténées avec succès. La première a en effet recours à la probabilité de présence du chant de chaque trame, tandis que la seconde se fonde sur la présence de trames consécutives contenant du chant.

Discussion

Les méthodes ProbaH2-10 et ETRAll proposées décrivent de nouvelles approches permettant de distinguer les chansons des instrumentaux au sein d'une base de données musicales. Ces deux méthodes fournissent de nouvelles agrégations d'éléments locaux qui permettent de constituer des caractéristiques globales.

Bien que la méthode d'agrégation fondée sur la probabilité de présence de chant, ProbaH2-10, apparaît peu performante lorsqu'elle est utilisée seule, elle permet néanmoins d'améliorer les résultats de classification lorsqu'elle est utilisée conjointement avec la méthode ETRAll.

Les méthodes d'agrégation proposées utilisent des MFCC à l'échelle de la trame. Il existe toutefois d'autres caractéristiques audio qu'il serait pertinent d'évaluer, telles que les i-vecteurs [Eghbal-zadeh et Widmer, 2016] ou la variance vocale [Lehner *et al.*, 2014] [1].

Il serait, de plus, possible d'ajouter une étape de traitement préalable à l'analyse par trames afin de maximiser les chances de détecter des trames vocales et de réduire le nombre de faux positifs. L'amélioration de la différenciation entre instrumentaux et chansons serait par exemple facilitée par l'utilisation de méthodes de séparation de sources. Pour cela, des méthodes de séparation du chant [Roma *et al.*, 2016; McVicar *et al.*, 2016; Fan *et al.*, 2017; Jansson *et al.*, 2017; Mimilakis *et al.*, 2017; Stoller *et al.*, 2018] ou de séparation des contenus mélodique et rythmique [Tachibana *et al.*, 2010] sont envisageables [Weninger *et al.*, 2011; Liutkus *et al.*, 2014] [2]. Ces méthodes demeurent toutefois imparfaites et il existe un risque de cumul des erreurs de plusieurs algorithmes si ceux-ci sont utilisés consécutivement. L'étude de l'impact de ces pré-traitements sur la pertinence de classification semble néanmoins pertinente et cette utilisation successive n'est donc pas complètement exclue.

Une dernière démarche consiste à reconsidérer le système de classification en deux classes. L'objectif est en effet actuellement de classer correctement un maximum de morceaux comme contenant du chant ou non. Cet objectif n'est cependant pas indispensable par exemple pour un site de streaming musical. Les sites de streaming possèdent en effet des bases de données de plusieurs millions de morceaux et l'annotation de tous les morceaux est considérée comme moins importante que le fait de garantir l'exactitude des annotations, même si ces annotations sont en faible nombre. Dans ce contexte, il est donc préférable de proposer un système capable de classer avec certitude un nombre restreint de morceaux dans une base de données et donc de garantir une précision de classification de 100% pour chacun de ces morceaux. Le nouvel objectif consiste par conséquent à augmenter la proportion de morceaux de la base de données qui sont correctement classés par notre approche. Cet objectif pourrait être atteint grâce aux méthodes d'apprentissage machine présentées dans le chapitre 4.

1. https://github.com/f0k/ismir2015, consulté le 19 Avril 2018.
2. Voir également https://github.com/Js-Mim/mss_pytorch, https://github.com/posenhuang/singingvoiceseparationrpca, https://github.com/Xiao-Ming/UNet-VocalSeparation-Chainer, https://github.com/EdwardLin2014/CNN-with-IBM-for-Singing-Voice-Separation et https://github.com/andabi/music-source-separation, consultés le 19 Avril 2018.

3.3 Conclusion sur les méthodes d'extraction de caractéristiques audio

« Ne rien livrer au hasard, c'est économiser du travail. »
– L'Art d'écrire enseigné en vingt leçons, 1899, Antoine Albalat

Ce chapitre a détaillé les étapes à suivre afin d'extraire des informations audio d'un signal musical. À partir des connaissances introduites dans ce chapitre, la figure 3.14 détaille les nouveaux éléments de la chaîne de traitement permettant de créer une liste de lecture musicale thématique. De la qualité des caractéristiques audio extraites des morceaux dépendront les résultats d'identification des thèmes par les méthodes d'apprentissage machine. Le chapitre 4 détaille comment ces caractéristiques audio sont utilisées afin de générer des listes de lecture musicale thématiques.

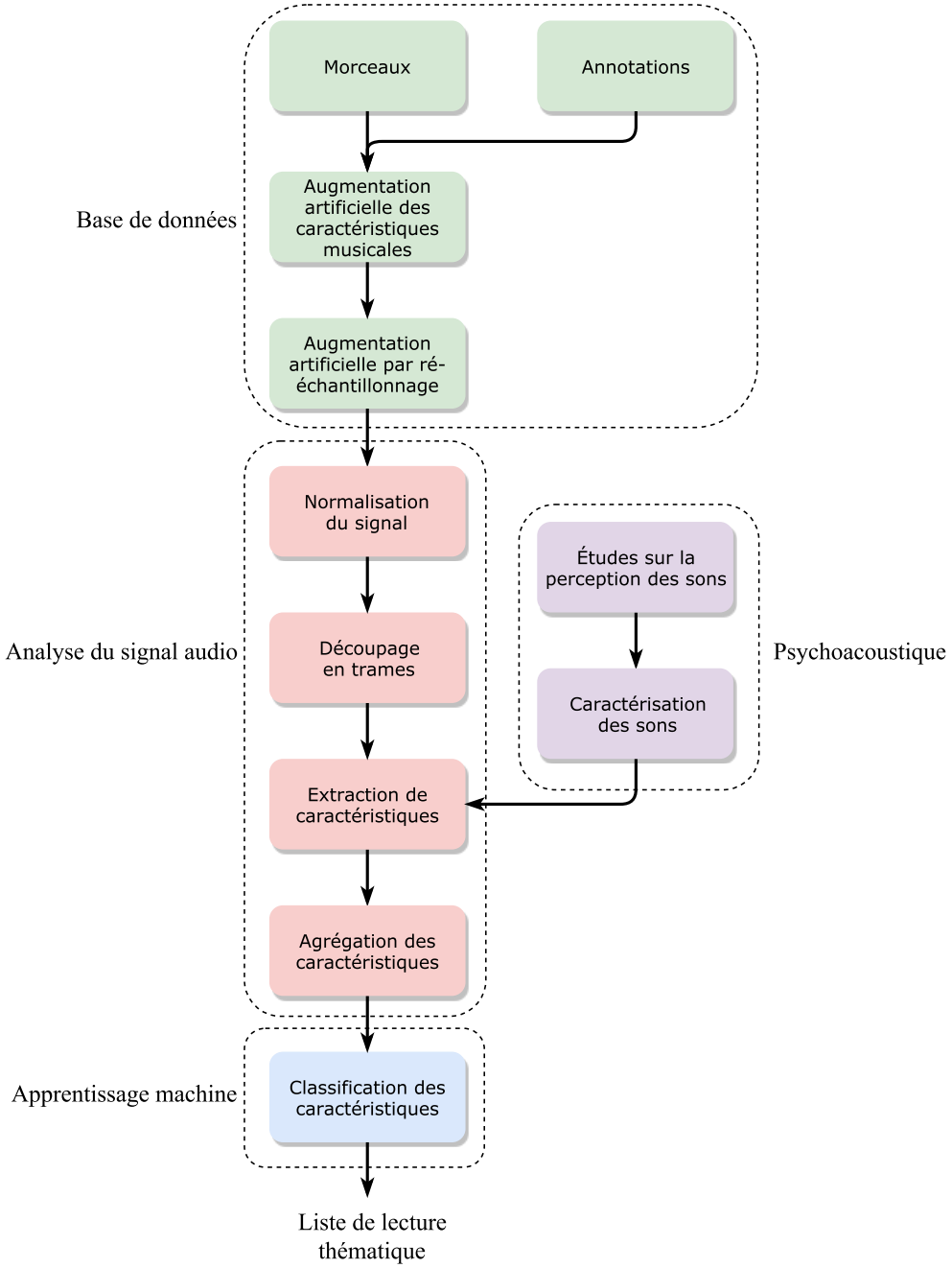


FIGURE 3.14 – Détails des étapes d'analyse audio à appliquer à des morceaux afin de générer des listes de lecture musicale.

4

Algorithmes d'apprentissage automatique

> *« Artificial intelligence is the future not only of Russia but of all of mankind ... Whoever becomes the leader in this sphere will become the ruler of the world. »*
>
> – *Cérémonie d'ouverture de la rentrée scolaire russe*, Septembre 2017, Vladimir Putin

Les méthodes de constitution des bases de données musicales ont été détaillées dans le chapitre 2 et le chapitre 3 a introduit les algorithmes d'extraction de caractéristiques audio pour l'estimation de présence de chant. Ce chapitre 4 décrit l'utilisation des caractéristiques audio des morceaux afin d'identifier leur thème. Les thèmes identifiés facilitent alors la génération de listes de lecture musicale thématiques. Plusieurs algorithmes d'apprentissage automatique, ou de *machine learning*, sont pour cela considérés et détaillés.

Les concepts clés de l'apprentissage machine, qui est un sous-domaine de l'intelligence artificielle, sont introduits dans la section 4.1.1. Les principes des méthodes de classification sont détaillés dans la section 4.1.2, qui souligne également les écueils à éviter lors de la conception et de l'utilisation de méthodes de classification. La section 4.3 décrit les expériences de classification musicale menées au cours de cette thèse. La section 4.4 détaille les méthodes d'apprentissage profond qui bouleversent actuellement tous les domaines de recherche utilisant des méthodes d'apprentissage automatique. Une expérience de classification musicale est également réalisée afin d'évaluer quel est l'impact de l'utilisation de l'apprentissage profond dans le domaine musical. La section 4.5 conclut enfin ce chapitre sur les algorithmes d'apprentissage automatique.

4.1 Introduction à l'apprentissage automatique

4.1.1 Concepts clés de l'intelligence artificielle

> *« Flow. Machines that describe other machines, texts that absorb other texts, bodies that absorb other bodies. »*
> – *Rhythm Science*, 2004, Paul Miller

La presse dédiée au grand public et non spécialisée tend à utiliser les termes et concepts clés de l'intelligence artificielle[1] de manière interchangeable[2] et grandiloquente[3]. La figure 4.1 expose une hiérarchie globale des termes les plus couramment utilisés en intelligence artificielle.

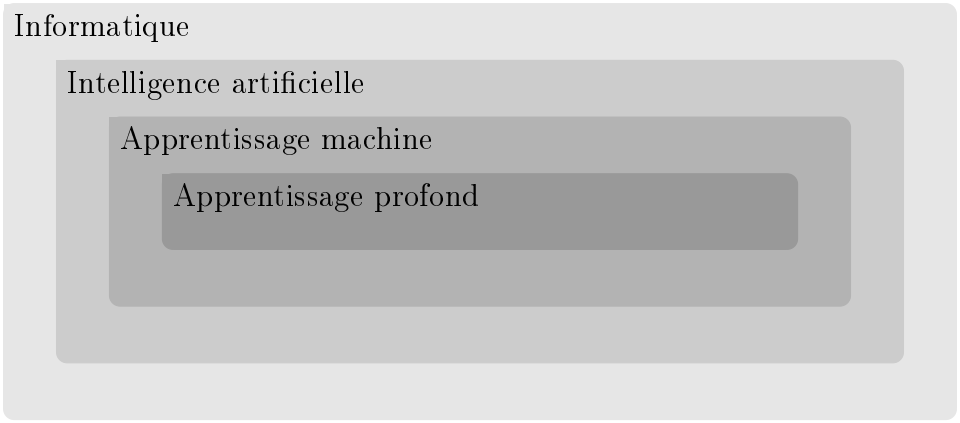


FIGURE 4.1 – Représentation du positionnement de l'apprentissage profond au sein de l'intelligence artificielle.

Comme indiqué dans la figure 4.1, l'intelligence artificielle est un sous-domaine de l'informatique qui cherche à simuler les capacités cognitives humaines par des ordinateurs. Au sein de l'intelligence artificielle, on retrouve par exemple des algorithmes permettant de trouver le meilleur itinéraire parmi un ensemble de chemins possibles mais également des méthodes qui tentent de reconnaître des visages dans des images. Au sein de l'intelligence artificielle, il existe également un domaine dédié à l'apprentissage automatique qui est d'intérêt dans le contexte de cette thèse.

Parmi les méthodes d'apprentissage automatique, on peut noter l'existence des algorithmes de régression sur des jeux de données. Un exemple d'utilisation

1. https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7?token=lQABrq73732J8a65, consulté le 19 Avril 2018.
2. https://www.latribune.fr/opinions/tribunes/ia-ne-dites-pas-apprentissage-profond-dites-apprentissage-statistique-772890.html, consulté le 19 Avril 2018.
3. https://www.bankinfosecurity.com/interviews/whats-artificial-intelligence-heres-solid-definition-i-3939, consulté le 19 Avril 2018.

de ces algorithmes consiste à prédire le prix d'une maison en fonction de sa superficie et de son ancienneté. Un autre type de méthodes d'apprentissage vise à classer les données d'une base en fonction de caractéristiques similaires dans des groupes. On parle alors de méthode de classification non supervisée ou de *clustering* lorsque les groupes de données ne sont pas définis *a priori* et que l'algorithme utilisé doit déceler une structure sous-jacente de regroupement dans la base de données.
La classification supervisée concerne en revanche le fait qu'un algorithme a accès à l'appartenance de chaque donnée à chacun des groupes avant d'effectuer son analyse. Dans ce dernier cas, l'objectif est d'associer une nouvelle donnée observée à un groupe existant. La classification supervisée constitue donc la méthode d'apprentissage automatique utilisée pour identifier les groupes musicaux, appelés thèmes, dans une base de données musicales. Comme présenté dans le tableau 4.1, une méthode de classification supervisée recherche la meilleure relation entre les caractéristiques d'un ensemble de données et le thème associé à chacune de ces données.

Tableau 4.1 – Exemple de représentation matricielle des caractéristiques et annotations d'une base de données.

Élément	Caractéristique 1	Caractéristique 2	...	Caractéristique c	Annotation
ID 1	$\text{Valeur}_{1,1}$	$\text{Valeur}_{1,2}$	...	$\text{Valeur}_{1,c}$	Thème 3
ID 2	$\text{Valeur}_{2,1}$	$\text{Valeur}_{2,2}$	...	$\text{Valeur}_{2,c}$	Thème 4
⋮	⋮	⋮	⋮	⋮	⋮
ID i	$\text{Valeur}_{i,1}$	$\text{Valeur}_{i,2}$	...	$\text{Valeur}_{i,c}$	Thème 3

Un sur-apprentissage des caractéristiques par la méthode de classification constitue l'un des facteurs propres aux Horses. Afin d'éviter un sur-apprentissage, l'un des algorithmes communément utilisés consiste à appliquer une validation croisée à la base de données. Pour cela, la base de données est divisée en N jeux de données. La méthode de classification recherche une relation entre les données provenant de N-1 jeux et cette relation est par la suite testée sur le jeu de données restant. On différencie alors le jeu d'entraînement, ou *train set*, (généralement composé des données provenant des N-1 jeux de données) du jeu de test, ou *test set*, (composé des données restantes), comme indiqué dans la figure 4.2. Dans la configuration de la validation croisée, la méthode de classification n'a donc pas accès aux données cachées du jeu de test, ce qui limite le phénomène de sur-apprentissage.

La figure 4.2 fournit un exemple de validation croisée à trois jeux de données (N = 3). Plus généralement, le choix du nombre de jeux de données dépend de la taille initiale de la base de données. Le nombre de jeux varie le plus souvent entre deux et dix [Ghosal *et al.*, 2013; Hespanhol, 2013; Zhang et Kuo, 2013; Gouyon *et al.*, 2014; Bayle *et al.*, 2016].

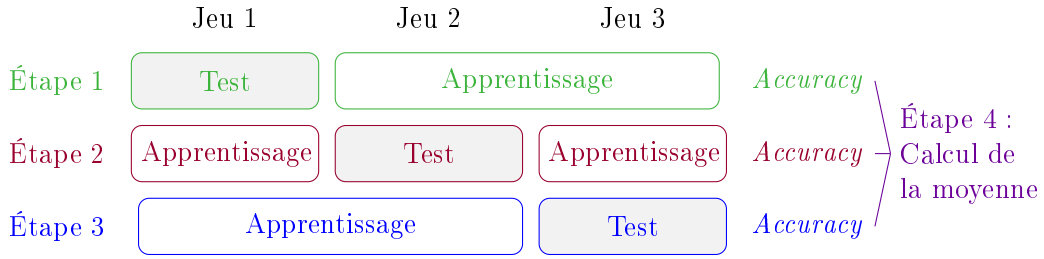


FIGURE 4.2 – Exemple de fonctionnement de la validation croisée à trois jeux de données.

Il est possible de mesurer le taux de sur-apprentissage en comparant la valeur d'une métrique calculée sur le jeu d'apprentissage et sur celui de test. Lorsque la métrique indique un résultat supérieur pour le jeu d'apprentissage que pour celui de test, alors il y a un sur-apprentissage. Inversement, lorsque la métrique fournit un résultat inférieur pour le jeu d'apprentissage que pour celui de test, alors il s'agit d'un sous-apprentissage. Lorsqu'une méthode de classification supervisée a atteint le point d'équilibre entre les résultats obtenus pour la base d'apprentissage et ceux obtenus pour la base de test, alors les paramètres utilisés sont optimaux pour cette méthode.

La recherche d'un résultat similaire, pour une métrique donnée, sur le jeu d'apprentissage et sur celui de test se justifie d'une seconde manière. Afin d'expliquer cela, on considère une méthode de classification fondée sur des arbres décisionnels. Si cette méthode utilise un nombre important de branches conditionnelles, elle risque un sur-apprentissage sur le jeu d'apprentissage. Au contraire, si cette méthode utilise un trop faible nombre de branches de conditions, elle risque un sous-apprentissage du jeu d'apprentissage. Il est donc impératif de trouver un compromis quant au nombre de branches décisionnelles à utiliser qui minimise les effets de sur- et sous-apprentissage. L'atteinte de ce point d'équilibre est justifié par la loi de parcimonie – ou rasoir d'Ockham –, qui indique qu'il est plus probable qu'un nombre minimum de causes élémentaires explique un phénomène donné. La loi de parcimonie constitue donc un outil permettant de réduire le risque qu'une méthode soit un Horse.

L'utilisation de validation croisée ne garantit toutefois pas les capacités de généralisation d'une méthode puisque les morceaux inclus dans une base de données ne sont pas forcément représentatifs de l'ensemble des morceaux existants [Ng, 1997]. De même, une validation croisée sur des bases de données de petite taille est sujette aux biais de par la nature des morceaux présents dans la base de données [Herrera *et al.*, 2003; Livshin et Rodet, 2003; Bogdanov *et al.*, 2011]. Pour pallier ce problème, il est possible d'utiliser deux bases de données distinctes, l'une pour la phase d'apprentissage et la seconde pour la phase de test [Chudáček *et al.*, 2009; Bekios-Calfa *et al.*, 2011; Llamedo *et al.*, 2012; Fernández *et al.*, 2015]. L'utilisation de deux bases de données distinctes

permet alors de limiter le biais de rassemblement des données apporté par les algorithmes ainsi que par chacune des équipes de recherche ayant constitué ces bases.

L'utilisation d'une validation croisée sur plusieurs jeux d'une base ou sur plusieurs bases de données ne suffit néanmoins pas à garantir une classification sans erreurs. Une méthode de classification peut en effet afficher des faux positifs, ce qui introduit des morceaux hors-thème dans une liste de lecture. Or, la génération de prédictions ayant une précision de 100% permet d'assurer la qualité d'une liste de lecture musicale. La section suivante présente des outils pertinents pour atteindre cet objectif.

4.1.2 Principes des méthodes de classification supervisées

> *« [...] human categorization should not be considered the arbitrary product of historical accident or of whim but rather the result of psychological principles of categorization, which are subject to investigation. »*
>
> – *Principles of Categorization*, 1978, Eleanor Rosch

Les principales méthodes de classification supervisées qui ont été utilisées en musique sont introduites ci-après. L'ensemble des éléments qui les composent n'est pas détaillé puisque cela a fait l'objet de nombreuses revues de l'état de l'art [Dietterich, 1998; Kotsiantis *et al.*, 2007; Wu *et al.*, 2008].

Avant d'utiliser une méthode d'apprentissage automatique, il est préférable de normaliser les données considérées en leur soustrayant leur moyenne puis en les divisant par leur écart type [Stolcke *et al.*, 2008]. Cette étape de normalisation permet aux méthodes d'apprentissage automatique de comparer pertinemment les caractéristiques considérées.

La méthode de classification supervisée la plus simple consiste à imposer des valeurs de seuil à chacune des caractéristiques des données afin de regrouper ces dernières par thèmes. Ce type de méthode est qualifié d'arbre de décision [Breiman *et al.*, 1984], dans lequel chaque branche permet de vérifier qu'une caractéristique est supérieure ou inférieure à un seuil donné. Dans ces arbres, les nœuds terminaux constituent donc les thèmes dans lesquels sont regroupées les données. Les résultats des arbres de décision sont simples à expliquer puisqu'ils utilisent des vérifications successives de seuils afin de déterminer la classe d'un morceau.

Les arbres de décision permettent également de modéliser des interactions entre différentes caractéristiques. Ces interactions correspondent à des successions de branches. De plus, il est à noter que les algorithmes utilisant des arbres de décision sont non-paramétriques. Ceci implique qu'aucune vérification *a priori* n'est à effectuer quant à la séparabilité linéaire des données ni à la

présence de données aberrantes ou *outliers*. Toutefois, les arbres de décision doivent être reconstruits lorsque de nouvelles données sont ajoutées à la base d'apprentissage.
Les arbres de décision ont également tendance à sur-apprendre, c'est la raison pour laquelle il est préférable d'utiliser la méthode dite de forêt aléatoire ou *Random Forest* proposée par Breiman [2001]. La première étape de la réalisation d'une forêt aléatoire consiste à calculer un *bootstrap* [Efron, 1979], c'est-à-dire à appliquer N arbres de décision à N sous-ensembles de la base de données initiale. La seconde étape effectue un vote majoritaire entre les résultats de chacun de ces arbres afin d'obtenir le thème de chaque élément.
De plus, la méthode des forêts aléatoires ne requiert pas l'affinage précis d'un grand nombre de paramètres et peut donc aisément passer à l'échelle. Les forêts aléatoires ont par exemple été utilisées avec succès par Lehner *et al.* [2014, 2015] dans la détection du chant à l'échelle de la trame. Les forêts aléatoires conservent toutefois l'un des défauts majeurs des arbres de décision. Les coefficients d'appartenance d'un élément à une classe sont en effet distribués dans l'arbre [Wu *et al.*, 2008]. Quinlan [1993] propose alors l'algorithme C4.5 censé améliorer leur interprétation. Quinlan [1993] regroupe pour cela les règles par thème à classer sous la forme « Si conditions A et B et C alors classer l'élément dans le thème T ».

La Machine à Vecteurs de Support ou Support Vector Machine (SVM), proposée par Cortes et Vapnik [1995] fait partie des méthodes de classification supervisées les plus utilisées [Wu *et al.*, 2008]. Le principe d'une SVM est de détecter un hyperplan dans l'espace des caractéristiques des données de telle sorte que ce plan maximise la distance entre les données appartenant à des thèmes différents.
En musique, l'utilisation des SVM affiche des atouts, notamment en détection et reconnaissance du genre musical [Sadjadi *et al.*, 2007; Gouyon *et al.*, 2014]. Les principaux avantages d'une SVM résident dans le fait qu'elle permet d'atteindre une importante *accuracy*, qu'elle présente une faible tendance au sur-apprentissage sur un jeu de données et qu'elle permet d'utiliser des données possédant des caractéristiques qui ne sont pas linéairement séparables. Toutefois, afin d'atteindre une *accuracy* importante, une SVM nécessite l'affinage de nombreux paramètres ainsi que l'utilisation de mémoire vive dont la taille est proportionnelle au nombre de données à traiter. Les résultats d'une SVM demeurent en outre complexes à interpréter de par les hyperplans proposés.

Un troisième type de méthode de classification supervisée consiste à utiliser une modélisation statistique probabiliste des données. La méthode la plus connue parmi celles de ce type est la classification naïve bayésienne [Nilsson, 1965] qui fournit une valeur R définie dans l'équation 4.1.

$$R = \frac{P(i|x)}{P(j|x)} \tag{4.1}$$

où

- i et j sont deux thèmes,
- x est une matrice contenant les caractéristiques d'un élément de la base de données.

Si $R > 1$, alors la classe i est assignée à x, dans le cas contraire la classe j lui est assignée. Une méthode de classification naïve bayésienne affiche un biais important et une variance faible, ce qui la rend simple d'utilisation notamment sur de petits jeux de données. Cette méthode ne peut toutefois pas inférer d'interactions entre des caractéristiques et elle requiert la vérification, au préalable, de l'absence de corrélation entre les caractéristiques considérées. La méthode naïve bayésienne a été utilisée dans le cadre de la reconnaissance d'émotions et de genres musicaux [Pohle *et al.*, 2005]. Afin d'obtenir davantage de détails techniques, il est possible de se référer à Kotsiantis *et al.* [2007].

La méthode des plus proches voisins ou K-Nearest Neighbours (k-NN) décrite dans la section 2.4.2 peut également être utilisée en tant que méthode de classification supervisée [Hastie et Tibshirani, 1996]. Dans ce cadre, elle affiche une importante variance et un faible biais, notamment pour de grands jeux de données, puisqu'elle introduit une faible erreur asymptotique dans les résultats.

La méthode RANdom Sample And Consensus (RANSAC) est une méthode itérative d'estimation de paramètres liant des caractéristiques de données à leur annotation. Elle a été proposée par Fischler et Bolles [1981] et a notamment été utilisée dans un contexte musical par Ghosal *et al.* [2013]. Cette méthode repose sur deux étapes qui sont menées de manière itérative. Au cours de la première étape, un sous-ensemble de données de la base est choisi aléatoirement et l'algorithme propose un modèle sur ce sous-ensemble. Au cours de la seconde étape, le modèle est testé sur le reste des données de la base afin de fournir un score de modélisation. La méthode réitère la première étape jusqu'à ce qu'un score de modélisation des données suffisant soit obtenu. Le critère d'arrêt peut être fixé en considérant un certain nombre d'itérations ou bien un nombre minimum d'éléments de la base qui doivent être correctement identifiés.

Le perceptron [Rosenblatt, 1957] constitue une autre méthode de classification supervisée. L'idée principale du perceptron est de trouver une matrice W qui modélise une variable y en appliquant une fonction non-linéaire σ à une variable x en entrée, comme indiqué dans l'équation 4.2.

$$y = \sigma(Wx) \tag{4.2}$$

où

- y est la sortie du système et représente la probabilité que x appartienne à un thème donné,
- σ est une fonction non-linéaire,
- W est une matrice de pondération,
- x est une matrice représentant un élément tel qu'un morceau ou une image.

Les fonctions non-linéaires les plus utilisées sont représentées dans la figure 4.3.

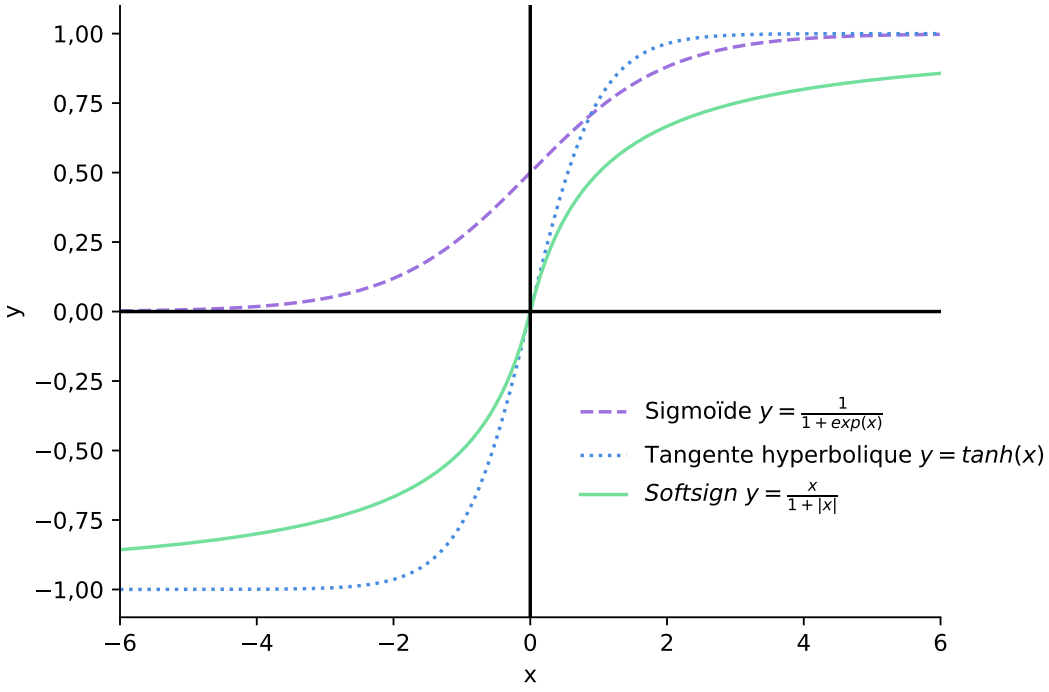


FIGURE 4.3 – Présentation de trois fonctions non-linéaires couramment utilisées par les perceptrons.

Les méthodes de classification décrites possèdent leurs propres avantages et inconvénients et fournissent des résultats différents en fonction des para-

mètres utilisés. Une solution consiste à combiner plusieurs de ces méthodes en utilisant des paramètres différents pour chacune d'elles afin d'obtenir une méthode globale plus performante que chacune des méthodes utilisées indépendamment. Le principe de combinaison de plusieurs méthodes de classification a été proposé par Freund et Schapire [1997] sous le nom de *Boosting*. L'une des méthodes les plus connues de *Boosting* est AdaBoost et elle a notamment été utilisée avec succès en classification thématique musicale [Bergstra *et al.*, 2006a].

Les principales méthodes de classification existantes sont implémentées dans le logiciel *scikit-learn* [Pedregosa *et al.*, 2011] écrit en Python. Ce logiciel constitue aujourd'hui une possibilité de garantir la réplicabilité des expériences dans la communauté de recherche des algorithmes d'apprentissage machine.

4.1.3 Vers une liste de lecture musicale thématique homogène

> *« De la musique encore et toujours ! »*
>
> – *Art poétique*, 1874, Paul Verlaine

La création d'une liste de lecture musicale ne contenant que des morceaux appartenant à un thème considéré revient à garantir une précision de 100% dans la classification d'un thème donné tout en maximisant le rappel pour ce thème. Afin d'atteindre cet objectif, la méthode de classification supervisée doit trouver une relation entre les caractéristiques et les annotations des morceaux, de telle sorte que la précision de détection pour un thème soit de 100%. Toutefois, le chapitre 3 a introduit le fait que les caractéristiques audio ne constituent pas nécessairement des données suffisantes pour décrire un thème donné et les méthodes d'apprentissage machine présentées dans la section 4.1.2 affichent des limites dans l'utilisation des caractéristiques audio. Par conséquent, dans ce contexte, comment garantir 100% de précision dans la détection des thèmes musicaux afin de générer une liste de lecture sans erreurs ?

Il est d'abord possible de considérer différemment les morceaux provenant du thème d'intérêt de tous les autres morceaux contenus dans le jeu d'apprentissage. La section 2.4 a présenté des méthodes d'augmentation de données et de ré-échantillonnage qui permettent d'augmenter le nombre d'éléments appartenant à un thème dans une base de données. L'utilisation d'une telle méthode devrait permettre à la méthode de classification supervisée d'obtenir davantage de morceaux décrivant un thème et donc d'obtenir de meilleurs résultats de détection de celui-ci. L'augmentation de données peut cependant se révéler coûteuse en terme de temps de calcul pour la production de nouveaux morceaux.

Une solution alternative consiste à pondérer différemment les éléments d'un thème par rapport à ceux d'autres thèmes. L'équivalence entre la modification

de distribution d'un thème et celle du coût de classification a été prouvée par Elkan [2001]. Ces deux derniers types de modifications présentent toutefois des avantages et inconvénients propres. Le sous-échantillonnage d'un thème majoritaire peut supprimer des données utiles alors que le sur-échantillonnage, qui crée des copies exactes d'un élément, peut entraîner un sur-apprentissage. Le sur-échantillonnage augmente également le temps de calcul de la phase d'apprentissage puisque de nouveaux éléments sont créés et traités. Par ailleurs, il n'est pas toujours possible d'appliquer un coût à un thème puisque toutes les méthodes de classification ne peuvent pas pondérer un élément durant la phase d'apprentissage. Ceci est notamment la cas de la méthode de classification supervisée C4.5.

De plus, la modification du coût ainsi que l'augmentation de données affichent un même inconvénient en terme de temps de calcul utilisé. En effet, le coût à appliquer ainsi que la modification du ratio des éléments dans chaque thème permettant de garantir 100% de précision n'est pas connu à l'avance. Dans les deux cas, l'utilisation d'une méthode de recherche paramétrique du coût et du ratio permet d'observer une éventuelle augmentation de la précision.

4.2 Classification des instrumentaux et des chansons

> *« When we combine great music with the continued investment, passion, innovation and support of the women and men at music companies, the results are profound. »*
>
> – *Global Music Report of the IFPI*, 2018, Sir Lucian Grainge, Chairman and Chief Executive Officer of Universal Music Group

Cette section compare une nouvelle méthode de classification des instrumentaux et des chansons avec celle proposée par Ghosal *et al.* [2013] détaillée dans la section 2.4.1. La nouvelle méthode, nommée *PERF*, utilise les caractéristiques audio [ProbaH2-10, ETRAll] ainsi que la méthode *Random Forest* Breiman [2001] en tant qu'algorithme d'apprentissage machine. L'expérience d'évaluation de ces deux méthodes consiste à mesurer la précision de classification des instrumentaux et des chansons en fonction de la taille de la base de données considérée. Les morceaux de cette base proviennent de *ccMixter*, *Jamendo*, *MedleyDB*, *QUASI*, *RWC* et d'une bibliothèque personnelle. La base a été augmentée suivant plusieurs critères tels que le décalage de fréquence, l'étirage temporel du signal audio, la modification de tempo et la modification du volume sonore. Le test de la méthode proposée dans de telles conditions permet d'évaluer la propension de la méthode à être un Horse. Les précisions atteintes par les deux méthodes sont affichées dans la figure 4.4.

La figure 4.4 indique que la précision des deux méthodes chute lorsque la

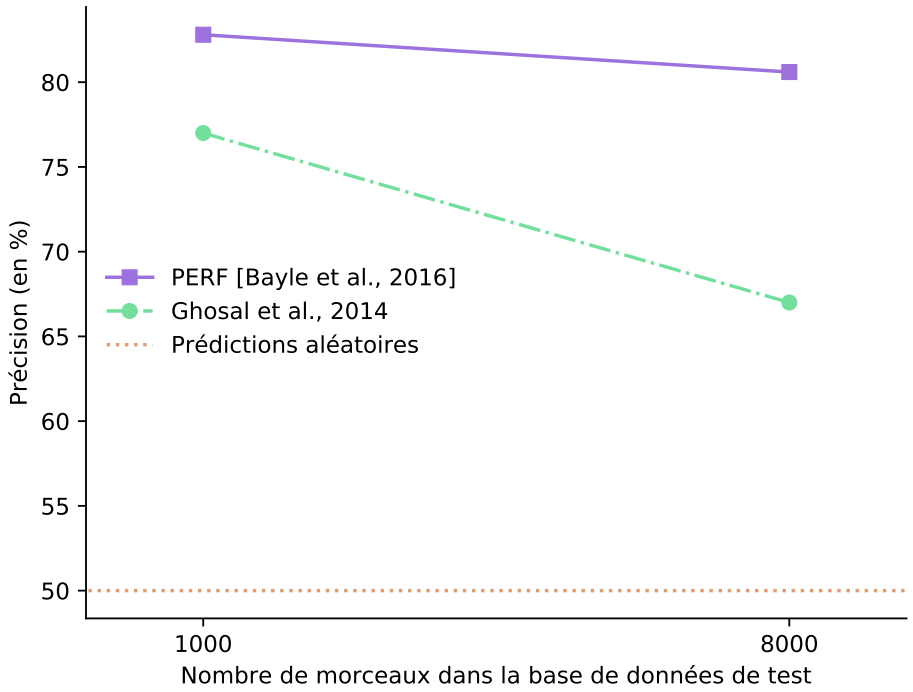


FIGURE 4.4 – Précision de classification des instrumentaux par la méthode de Ghosal *et al.* [2013] comparée à la méthode *PERF* proposée [Bayle *et al.*, 2016].

taille de la base de données augmente. La précision de la méthode de Ghosal *et al.* [2013] chute de 10,1 points lorsque la base de données augmente de 1 000 à 8 000 morceaux alors que la précision de la méthode *PERF* ne diminue que de 2,2 points sur ce même intervalle. Pour la méthode *PERF*, les précisions de 82,8% et de 80,6% sont en effet obtenues respectivement pour 1 000 et 8 000 morceaux. Sur la base de données de test, la nouvelle approche semble plus robuste à un passage à l'échelle que celle de Ghosal *et al.* [2013] puisque sa précision est moins affectée lorsque la taille de la base de données utilisée est multipliée par huit. Il est possible que la précision des deux méthodes diminue également lors d'un nouveau passage à l'échelle plus conséquent. Dans les conditions expérimentales proposées, la méthode *PERF* ne semble pas être un Horse et ce grâce à l'utilisation des caractéristiques audio proposées, qui sont ProbaH2-10 et ETRAll. Il est en effet possible que des conditions expérimentales différentes permettent de trouver d'autres aspects d'un Horse pour la méthode *PERF*. Néanmoins, les performances de la méthode *PERF* dans les conditions proposées justifient son utilisation dans la section suivante afin de générer automatiquement une liste de lecture instrumentale.

4.3 Expérience de génération automatique d'une liste de lecture musicale instrumentale

> *« [...] if you're doing an experiment, you should report everything that you think might make it invalid –not only what you think is right about it: other causes that could possibly explain your results ; and things you thought of that you've eliminated by some other experiment, and how they worked– to make sure the other fellow can tell they have been eliminated. »*
>
> *– Surely You're Joking, Mr. Feynman !: Adventures of a Curious Character*, 1997, Richard Feynman

Les méthodes proposées dans la littérature afin de différencier les instrumentaux des chansons [Ghosal *et al.*, 2013; Hespanhol, 2013; Gouyon *et al.*, 2014] sont comparées avec une méthode utilisant le couple d'agrégation de caractéristiques [ProbaH2-10, ETRAll]. Cette section détaille les algorithmes et les implémentations des méthodes existantes et de celle proposée.

L'algorithme de Ghosal *et al.* [2013] (GA) utilise la méthode de classification RANSAC afin de trier une base de données privées contenant 540 morceaux équitablement répartis entre instrumentaux et chansons. Leur méthode atteint ainsi une *accuracy* de 92,96% en utilisant une validation croisée à deux jeux mais les auteurs ne fournissent pas le code source de leur expérience. Afin de reproduire l'algorithme de Ghosal *et al.* [2013], le logiciel YAAFE [Mathieu *et al.*, 2010] a été utilisé afin d'extraire les MFCC. La méthode RANSAC implémentée dans le logiciel *scikit-learn* [Pedregosa *et al.*, 2011] a quant à elle été utilisée comme méthode d'apprentissage automatique.

La méthode de Gouyon *et al.* [2014], que ceux-ci notent Support Vector Machine applied to Bags of Frames of Features (SVMBFF), utilise une SVM pour classer les morceaux provenant de trois bases de données musicales distinctes, qui contiennent entre 502 et 2 349 morceaux. Pour les trois bases utilisées, le *f-score* de cette méthode varie entre 89% et 95% en ce qui concerne la détection des chansons. La détection des instrumentaux est moindre puisque cette même méthode affiche un *f-score* compris entre 45% et 80%. Le code source fourni par les auteurs est utilisé afin de tester leur algorithme dans le cadre de notre expérience.

La méthode proposée par Langlois et Marques [2009] et améliorée par Gouyon *et al.* [2014] utilise un modèle markovien comme méthode de classification. Cette méthode, notée Vector Quantization and Markov Model (VQMM), est utilisée pour trier les morceaux provenant de trois bases de données musicales distinctes qui contiennent entre 502 et 2 349 morceaux. Pour les trois bases, le *f-score* de VQMM varie entre 83% et 95% en ce qui concerne la détection des chansons. La détection des instrumentaux est inférieure à celle des chansons puisque la méthode affiche un *f-score* compris entre 54% et 66%. De

même que pour la méthode de Gouyon *et al.* [2014], le code source fourni par les auteurs est utilisé afin de tester leur algorithme dans le cadre de notre expérience.

La nouvelle méthode proposée est notée PEA (ProbaH2-10, ETRAll et AdaBoost). Pour chaque morceau, la méthode PEA extrait les caractéristiques audio ProbaH2-10 et ETRAll ainsi que 13 MFCC avec les delta et double delta correspondants. La méthode de classification supervisée choisie est la version d'AdaBoost implémentée dans le logiciel *scikit-learn* [Pedregosa *et al.*, 2011]. Le code source en Python est disponible sur GitHub[1].

4.3.1 Matériels et méthodes

L'objectif de cette expérience est de comparer les performances de détection des méthodes existantes avec celles de la méthode PEA proposée et ce sur une même base de données musicales. L'expérience proposée définit trois conditions qui sont détaillées dans le tableau 4.2.

Tableau 4.2 – Définition des trois conditions de l'expérience de classification supervisée des instrumentaux et des chansons.

	Jeu de test	Validation croisée
Condition 1	Petite base équilibrée (186 morceaux)	5 jeux
Condition 2	Base équilibrée (8 912 morceaux)	Aucune
Condition 3	Base déséquilibrée (41 491 morceaux)	Aucune

Dans la première condition, la base de test est la même que celle d'apprentissage et les résultats de classification sont observés pour une validation croisée sur cinq jeux. La deuxième condition constitue une mise à l'épreuve du passage à l'échelle des algorithmes évalués puisque la base de test est constituée de 8 912 morceaux équitablement répartis entre les instrumentaux et les chansons provenant de SATIN. Dans cette deuxième condition, en effet, une base de test contenant 48 fois plus de morceaux que la base d'apprentissage est utilisée. Dans la troisième condition, l'intégralité des morceaux de SATIN est utilisée en tant que base de test.

Dans les trois conditions, les méthodes utilisent une petite base d'apprentissage composée de 186 morceaux équitablement répartis entre les instrumentaux et les chansons. La base d'apprentissage est composée de plusieurs bases de données musicales couramment utilisées dans la littérature [Ramona *et al.*, 2008; Bittner *et al.*, 2014; Lehner *et al.*, 2014; Liutkus *et al.*, 2014; Schlüter et Grill, 2015; Schlüter, 2016]. Cette base d'apprentissage inclut notamment les morceaux de *ccMixter*, de *Jamendo* et de *MedleyDB*.

1. https://github.com/ybayle/ReproducibleResearchCode, consulté le 19 Avril 2018

4.3.2 Résultats

La comparaison des résultats d'*accuracy* et d'écart-type obtenus dans l'expérience sont indiqués dans la figure 4.5.

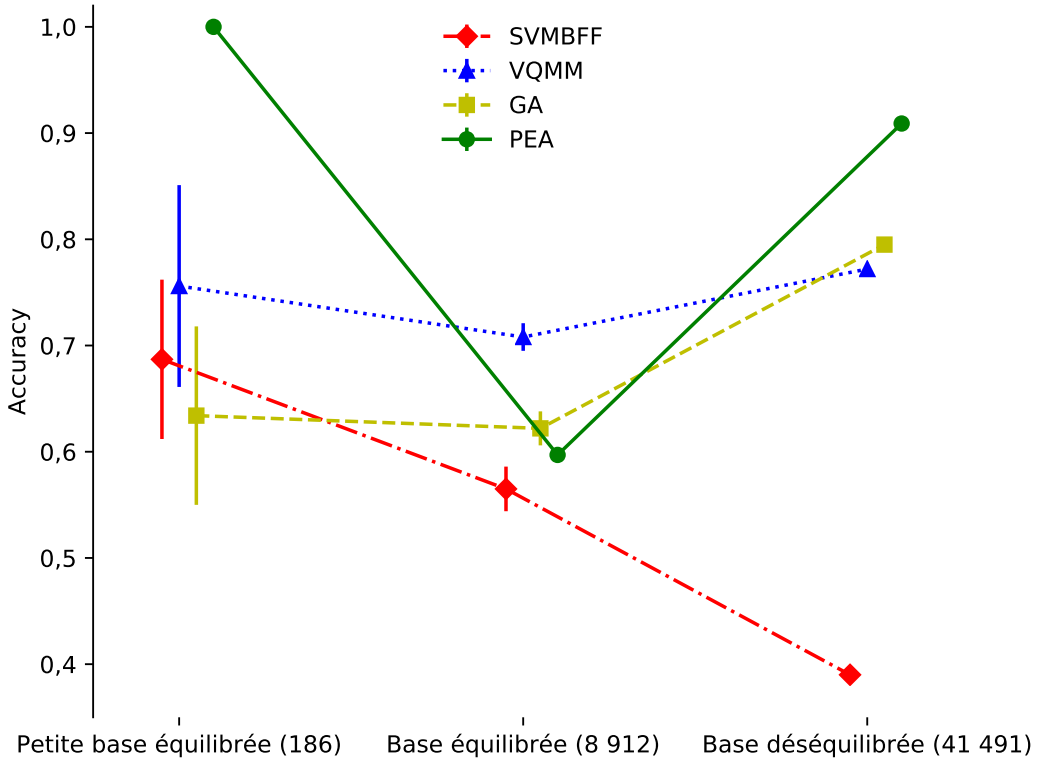


FIGURE 4.5 – Comparaison de l'*accuracy* obtenue par différentes méthodes de classification des instrumentaux et des chansons en fonction de trois bases de données de test différentes.

La méthode PEA proposée affiche une meilleure *accuracy* que les méthodes GA, SVMBFF et VQMM pour une base déséquilibrée. L'intégralité des méthodes voit chuter l'*accuracy* lorsque la taille de la base équilibrée de test est augmentée.

La différence entre le nombre d'instrumentaux et de chansons complexifie l'interprétation de l'*accuracy* dans la base de données déséquilibrée. Le tableau 4.3 indique donc la précision et le rappel pour chacune des méthodes utilisées afin de faciliter la compréhension de leurs résultats. Ces résultats sont également comparés à une méthode de référence, nommée TousInstru, qui annote tous les morceaux comme étant des instrumentaux. De même, une autre méthode de référence nommée TousChan annote tous les morceaux comme étant des chansons. Cette comparaison permet d'évaluer la capacité des méthodes de classification à traiter une base de données déséquilibrée.

Tableau 4.3 – Précision et rappel pour la détection des instrumentaux et des chansons par sept algorithmes de classification. La base d'apprentissage est constituée de 186 morceaux équitablement répartis entre des instrumentaux et des chansons. SATIN constitue la base de test utilisée et comporte 37 035 chansons (89%) et 4 456 instrumentaux (11%). Les nombres en gras indiquent les meilleurs résultats obtenus pour chaque métrique.

	Instrumentaux		Chansons	
Algorithmes	Précision	Rappel	Précision	Rappel
TousInstru	0.110	**1.000**	0	0
TousChan	0	0	0.889	**1.000**
Aléatoire	0.110	0.500	0.889	0.500
GA	0.173	0.307	0.908	0.824
SVMBFF	0.125	0.803	0.932	0.324
VQMM	0.298	0.706	0.956	0.794
PEA	**0.825**	0.200	**0.959**	0.844

Lorsque tous les morceaux de la base de données musicales sont annotés en tant que chansons, la précision obtenue par la méthode TousChan ne diffère de celles de GA, SVMBFF et VQMM que de 6,7 points au maximum. Cette faible différence de précision est due au fait qu'il est complexe de proposer une méthode de détection de thème musical performante dans un contexte où le thème considéré est en surnombre dans la base de données.

La différence principale entre les méthodes provient du fait que la précision de détection des instrumentaux affichée par la méthode PEA est presque trois fois plus importante que celle de la meilleure méthode dans cet état de l'art. Si les méthodes GA, SVMBFF ou VQMM étaient donc utilisées afin de classer les instrumentaux et les chansons, elles généreraient des listes de lecture musicale instrumentales contenant au plus 30% d'instrumentaux. La méthode PEA amène ce nombre au-delà de 80%. Ceci implique que plus de quatre instrumentaux sur cinq sont correctement annotés automatiquement et une telle différence de classification avec les méthodes de l'état de l'art serait sûrement perçue par un auditeur. La grande précision atteinte par la méthode PEA ne peut en outre pas être due à un effet de sur-apprentissage puisque la base d'apprentissage est 223 fois plus petite que la base de test.

Par ailleurs, la méthode de classification aléatoire prédit qu'un morceau sur deux est un instrumental et elle affiche une précision de 11% dans la détection des instrumentaux. Or, la prévalence des instrumentaux dans la base est de 11%, ce qui explique pourquoi cette méthode affiche en moyenne 11 prédictions correctes sur 100. Ceci explique également que la méthode qui annote tous les morceaux comme des instrumentaux affiche 11% de précision.

De plus, la méthode de classification aléatoire affiche un rappel de 50% puisque la moitié des instrumentaux a correctement été annotée. En revanche, la mé-

thode annotant tous les morceaux comme étant des instrumentaux affiche un rappel de 100% puisque tous les instrumentaux ont correctement été annotés.

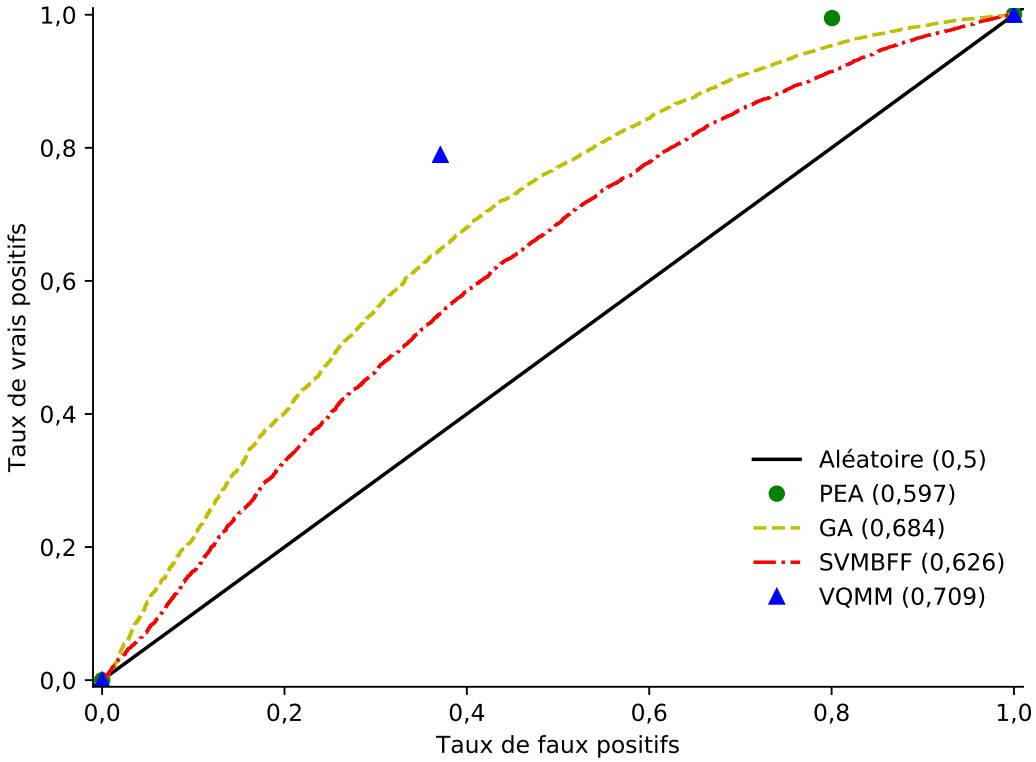


FIGURE 4.6 – Taux de vrais positifs en fonction du taux de faux positifs obtenus par les méthodes de classification GA, SVMBFF, VQMM, PEA et une méthode de classification aléatoire. L'Area Under the Curve (AUC) de chaque méthode est affichée entre parenthèses. La base d'apprentissage est constituée de 186 morceaux équitablement répartis entre des instrumentaux et des chansons. SATIN constitue la base de test utilisée et comporte 37 035 chansons (89%) et 4 456 instrumentaux (11%).

Les précisions obtenues par GA, SVMBFF et VQMM sont supérieures à celle obtenue par la méthode de classification aléatoire. Le gain de performance affiché par VQMM comparé à ceux affichés par GA et par SVMBFF semble donc bien provenir de la méthode d'agrégation de caractéristiques audio locales. Les précisions de classification des instrumentaux affichées par GA, SVMBFF et VQMM demeurent toutefois très faibles puisqu'elles ne dépassent pas 30%. La méthode PEA obtient quant à elle une importante précision, qui est toutefois au détriment du rappel. La figure 4.6 détaille en partie la raison de cette diminution du rappel en affichant le taux de vrais positifs en fonction du nombre de faux positifs obtenus par GA, SVMBFF, VQMM et PEA. Les méthodes VQMM et PEA n'affichent qu'un seul taux de vrais positifs différent

de 0 ou de 1 car, par construction, ces méthodes de classification ne permettent pas la variation du seuil de probabilité de classement d'un morceau.

La figure 4.6 indique que 100% de vrais positifs ne sont obtenus que lorsque les méthodes affichent également un taux de faux positifs de 100%. Or, l'objectif est de garantir un taux de vrais positifs de 100% tout en minimisant le taux de faux positifs. Une liste de lecture thématique sans faux positifs et donc sans erreurs ne peut donc pas être garantie par GA, SVMBFF, VQMM ni PEA.

4.3.3 Discussion

> *« The anonymous programmers who write the algorithms that control the series of songs in these streaming services may end up having a huge effect on the way that people think of musical narrative – what follows what, and who sounds best with whom. Sometimes we will be the D.J.s, and sometimes the machines will be, and we may be surprised by which we prefer. »*
>
> – *You, the D.J. – Online music moves to the cloud*, article du 14 Juin 2010 du quotidien *The New Yorker*, Sasha Frere-Jones

Les méthodes évaluées améliorent la détection des chansons de 7 points au maximum par rapport à une méthode de classification aléatoire sur la base de données déséquilibrée proposée. Un auditeur pourrait par conséquent ne pas percevoir de différences significatives entre les listes de lecture thématiques autour de chansons générées par GA, SVMBFF, VQMM, PEA ou par une méthode de classification aléatoire.

Le fait de garantir une liste de lecture instrumentale demeure toutefois un défi. Si GA, SVMBFF et VQMM généraient en effet une liste de lecture instrumentale, alors celle-ci contiendrait au mieux 30% d'instrumentaux, une proportion qui serait certainement détectée par un auditeur. La précision des méthodes de l'état de l'art ne semble donc pas suffisante pour garantir une liste de lecture musicale thématique homogène à partir d'une base de données déséquilibrée. La méthode PEA proposée obtient en revanche une précision 2,7 fois supérieure aux précisions affichées par les méthodes évaluées et constitue donc la meilleure méthode actuelle dans la détection instrumentale. La méthode PEA semble donc être pertinente afin de générer une liste de lecture thématique à partir d'une base de données musicales qui affiche un déséquilibre de représentation entre les thèmes. Par ailleurs, le gain de précision de PEA par rapport à GA, SVMBFF et VQMM semble provenir de la nouvelle méthode d'agrégation des caractéristiques audio. Les méthodes GA et SVMBFF utilisent en effet la moyenne et l'écart type des caractéristiques audio, ce qui semble réduire l'information permettant de différencier les instrumentaux des chansons. La méthode VQMM utilise quant à elle un modèle de Markov qui infère la présence de chant pour une trame en utilisant uniquement

la trame précédente. Cette procédure semble limiter les capacités d'inférence à long terme de cette méthode. L'agrégation des caractéristiques audio proposée dans la méthode PEA semble au contraire afficher une meilleure modélisation de la présence de trames vocales et de leur l'évolution au sein des morceaux. Il pourrait être pertinent d'évaluer l'agrégation proposée sur des tâches musicales différentes de celle de reconnaissance des instrumentaux et des chansons. La précision affichée par la méthode PEA est supérieure à celle des méthodes de l'état de l'art mais son rappel est inférieur à celui affiché par les méthodes existantes. L'importante précision affichée par la méthode PEA est donc au détriment d'un rappel plus faible, ce qui limite son utilisation dans d'autres contextes.

Cette expérience a également permis d'évaluer les performances des méthodes existantes et de la méthode PEA dans le cadre, d'une part d'un passage à l'échelle et d'autre part en conditions réelles étant donné le déséquilibre thématique dans la base utilisée. Dans ce contexte, la méthode PEA semble donc particulièrement pertinente pour la classification thématique musicale ainsi que pour la génération de listes de lecture thématiques.

La phase d'agrégation des caractéristiques audio locales proposée semble être l'élément décisif dans l'importante précision atteinte au cours de cette expérience. De plus, l'annotation à l'échelle de la trame d'une centaine de morceaux est suffisante pour que la méthode PEA affiche une importante précision sur une base de test qui est beaucoup plus grande que la base d'apprentissage. Le rapport entre la taille de la base d'apprentissage (186 morceaux) et la taille de la base de test (41 491 morceaux) est en effet de 223 alors que ce rapport est dans la littérature généralement inférieur à 2. La méthode PEA semble donc posséder des avantages majeurs qui lui permettent de passer à l'échelle pour la génération de listes de lecture thématiques à partir d'un grand nombre de données non annotées. La méthode PEA atteint en revanche 82,5% de précision de classification des instrumentaux alors que la valeur de 100% est souhaitée afin d'obtenir la génération de listes de lecture musicale sans erreurs, c'est-à-dire qui ne contiennent aucun faux négatif. Dans ce cadre, l'utilisation d'AdaBoost semble être limitant puisque cette méthode ne permet pas de modifier les coûts associés à chaque thème. Or, comme indiqué dans la section 4.1.3, il doit être possible de modifier le coût associé à chaque thème afin de diminuer le nombre de faux négatifs. La section suivante indique plus en détails pourquoi la modification des coûts ne fonctionne néanmoins pas avec les méthodes considérées dans cette expérience.

Modifications des coûts thématiques

La modification des coûts thématiques au sein d'une méthode de classification peut permettre de rééquilibrer artificiellement la proportion des éléments dans chaque thème et donc d'améliorer la détection de ces derniers.

GA utilise RANSAC qui ne peut pas, par constitution, prendre en compte de coût de classification. Il n'a pas non plus été possible d'affiner suffisamment les autres paramètres de la méthode RANSAC afin obtenir plus de 1% de variation de précision. Ceci est dû au fait que la méthode est optimisée pour maximiser l'*accuracy* et non la précision.

VQMM utilise un modèle de Markov qui ne peut pas prendre en compte de coût de classification. Il a cependant été possible d'affiner d'autres paramètres de VQMM mais cela n'a pas suffisamment modifié la précision pour qu'un impact soit perçu quant à la détection des instrumentaux.

Il est possible d'affiner les résultats de la méthode SVMBFF puisque celle-ci utilise SVM qui accepte des coûts de classification. Cependant et après avoir testé plusieurs coûts de classification, la précision n'a au maximum été modifiée que de 1%. Cette limitation semble provenir des caractéristiques audio ainsi que de la méthode d'agrégation utilisée qui ne semblent pas être assez performantes pour détecter la présence de chant.

La méthode PEA ne peut enfin pas être affinée en utilisant une modification des coûts thématiques puisque la méthode de classification AdaBoost n'est pas conçue à cet effet. Il n'est donc pas possible d'affiner les paramètres des méthodes évaluées afin d'obtenir une meilleure précision de classification. Les différentes méthodes évaluées semblent de plus être limitées par leur apprentissage automatique quant à l'atteinte d'une meilleure précision. Les méthodes d'apprentissage automatique sont elles-mêmes restreintes par les caractéristiques audio considérées qui ne différencient pas suffisamment les thèmes des morceaux. Puisque les méthodes d'agrégation de caractéristiques ne semblent pas être en cause dans ce problème, il semble pertinent de rechercher de nouveaux descripteurs du chant à l'échelle de la trame afin d'améliorer la précision de classification.

Limitations de la méthode PEA proposée

De même que pour la méthode VQMM et à cause de l'utilisation d'AdaBoost, il n'est pas possible de trouver des paramètres pour la méthode PEA garantissant 100% de précision sur une classe donnée. AdaBoost atteint une précision de 82,5% sur une base de données de test de 41 491 morceaux et permet une meilleure détection des instrumentaux que les autres méthodes évaluées. Il n'est toutefois pas possible de modifier suffisamment les paramètres internes d'AdaBoost pour améliorer la précision obtenue. Les méthodes de classification SVM et *Random Forest* ont également été testées et leurs paramètres modifiés mais ces méthodes n'atteignent pas de précision aussi importante que celle atteinte lors de l'utilisation d'AdaBoost.

La méthode PEA dans son état actuel atteint la meilleure précision quant à la classification des instrumentaux par rapport aux méthodes de l'état de l'art. Elle ne permet pourtant pas de garantir une liste de lecture musicale

thématique sans erreurs. L'objectif de cette section est de réduire le nombre de faux positifs à zéro. Or, la méthode d'agrégation proposée semble être plus performante que celles de l'état de l'art afin de réaliser cette tâche. Il semble donc que les caractéristiques audio locales soient l'élément limitant de cette réduction de faux positifs. Il serait par conséquent pertinent de tester de nouvelles caractéristiques audio locales provenant notamment des algorithmes de détection de la voix à l'échelle de la trame [Nwe *et al.*, 2004; Lukashevich *et al.*, 2007; Ramona *et al.*, 2008; Regnier et Peeters, 2009; Lehner *et al.*, 2014; Leglaive *et al.*, 2015; Lehner *et al.*, 2015; Schlüter et Grill, 2015; Schlüter, 2016]. La Variance Vocale [Lehner *et al.*, 2014], le Vibrato Vocal [Regnier et Peeters, 2009], l'Atténuation Harmonique [Nwe *et al.*, 2004] ou le filtrage de la moyenne flottante auto-régressive [Lukashevich *et al.*, 2007] constituent des caractéristiques audio dont l'évaluation serait pertinente. Une approche par apprentissage profond a également été proposée dans le cadre de la détection de la voix à l'échelle de la trame [Kereliuk *et al.*, 2015; Leglaive *et al.*, 2015; Lehner *et al.*, 2015; Schlüter et Grill, 2015; Lidy et Schindler, 2016; Pons *et al.*, 2016]. Cependant, l'apprentissage profond constitue une approche récente [1] et peu étudiée en classification musicale [Bayle *et al.*, 2018a,b]. L'impact de l'apprentissage profond sur la classification musicale est donc analysé dans la section suivante.

1. https://github.com/ybayle/awesome-deep-learning-music, consulté le 19 Avril 2018.

4.4 Apprentissage profond

« Any sufficiently advanced technology is indistinguishable from magic. »
– Hazards of prophecy: The failure of imagination, 1962, Arthur Charles Clarke

Les méthodes d'apprentissage profond, ou de *deep learning*, sont qualifiées de méthodes d'apprentissage de bout-en-bout ou de *end-to-end learning* lorsqu'elles prennent des données brutes en entrée et prédisent des annotations pour chaque élément de la base de données. La qualification de bout-en-bout se justifie en effet par le fait qu'une seule méthode remplace plusieurs méthodes dédiées à différentes tâches telles que l'extraction de caractéristiques audio et l'apprentissage machine. La section 4.4.1 introduit les principes de l'apprentissage profond ainsi que le contexte historique qui entoure leur conception. La section 4.4.2 évalue enfin l'impact de l'application des méthodes d'apprentissage profond appliquées à la musique.

4.4.1 Principes de l'apprentissage profond

« Deep learning is not magic. »
– Comments on deep learning[1], 26 Août 2014, Yoshua Bengio

L'apprentissage profond s'est notamment fait connaître du grand public en 2017 grâce à l'algorithme AlphaGo de Google qui a battu un être humain au jeu de Go[2]. Les méthodes d'apprentissage profond connaissent actuellement un regain d'intérêt si l'on considère le nombre d'articles relatifs publiés mais également les applications qui utilisent ce type d'approche [Schmidhuber, 2015]. Les mécanismes de base de l'apprentissage profond ont toutefois été proposés il y a plus de 70 ans et consistent à simuler un réseau simplifié de neurones humains [McCulloch et Pitts, 1943]. En biologie, une cellule nerveuse, appelée neurone, possède un péricaryon en son centre qui reçoit un influx nerveux des dendrites et qui transmet une réponse à des neurones adjacents *via* son axone. McCulloch et Pitts [1943] ont proposé de simuler le fonctionnement d'un neurone en utilisant un perceptron, comme défini dans l'équation 4.2. Le signal transmis aux dendrites, donc à l'extrémité réceptrice du neurone, correspond à la variable x appliquée à l'entrée du perceptron et la réponse transmise par l'axone d'un neurone correspond au signal y en sortie du perceptron. Tous les neurones possèdent une caractéristique commune qui est modélisée par la fonction non-linéaire σ et qui est appliquée à une matrice W. Cette matrice

1. https://plus.google.com/+YoshuaBengio/posts/GJY53aahqS8, consulté le 19 Avril 2018.
2. https://deepmind.com/research/alphago/, consulté le 19 Avril 2018.

correspond aux caractéristiques intrinsèques de chaque péricaryon. En biologie, un cerveau est composé d'un réseau de neurones qui peut être simulé en informatique en cumulant un ensemble de perceptrons. Le fait de cumuler plusieurs perceptrons permet de modéliser mathématiquement des relations plus complexes entre un signal d'entrée x et un signal de sortie y telle que l'opération *OU exclusif*, notée *XOR* ou $\oplus$ [Riedmiller, 1994]. L'équation 4.3 exprime l'empilement de trois couches de perceptrons, appelé Multi Layer Perceptron (MLP), qui justifie la profondeur de l'apprentissage et le nom de ce type de méthode. Il est à noter que le nombre de perceptrons par couche est variable et dépend des données traitées ainsi que du type d'architecture utilisée.

$$y = \sigma(W_3\sigma(W_2\sigma(W_1x))) \tag{4.3}$$

où

- y est le résultat en sortie du système et représente la probabilité que x appartienne à un thème donné,
- σ est une fonction non-linéaire qui peut être définie par l'une des courbes affichées dans la figure 4.3,
- W_1, W_2 et W_3 constituent les matrices de pondération de chacune des trois couches,
- x est une matrice représentant un élément tel qu'un morceau ou une image.

L'équation 4.3 peut être représentée suivant le schéma affiché dans la figure 4.7, où la taille de W_1 est de 4 et celles de W_2 et de W_3 est de 5.

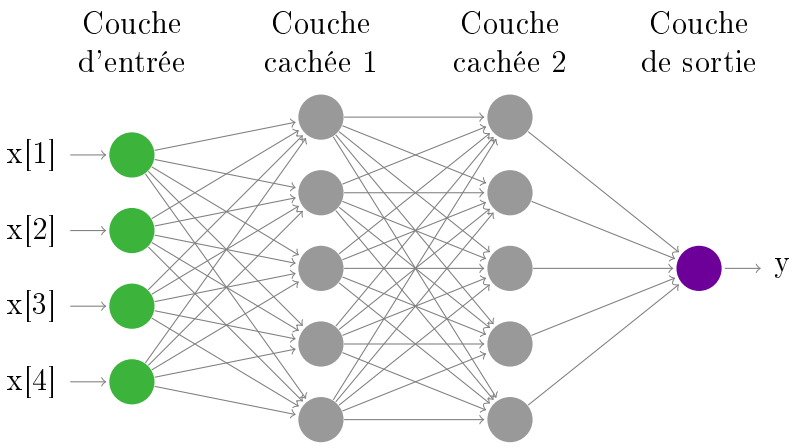


FIGURE 4.7 – Schéma d'un réseau de neurones constituant l'apprentissage profond d'un Multi Layer Perceptron (MLP). L'architecture correspondante est définie dans l'équation 4.3.

L'objectif d'une méthode d'apprentissage profond est de trouver une matrice de pondération W qui permette d'obtenir les meilleures prédictions possibles de y. Cela revient à minimiser le *loss* défini dans l'équation 1.5.

Un MLP constitue néanmoins un réseau de neurones particulier puisqu'il représente un graphe acyclique au sein duquel les calculs s'effectuent depuis l'entrée vers la sortie. Une couche définie par un MLP est qualifiée de « dense », de couche d'aplanissement ou de *flattening layer*.

La figure 4.8 reproduit le type de représentation couramment utilisé dans la littérature pour illustrer une architecture de réseaux de neurones.

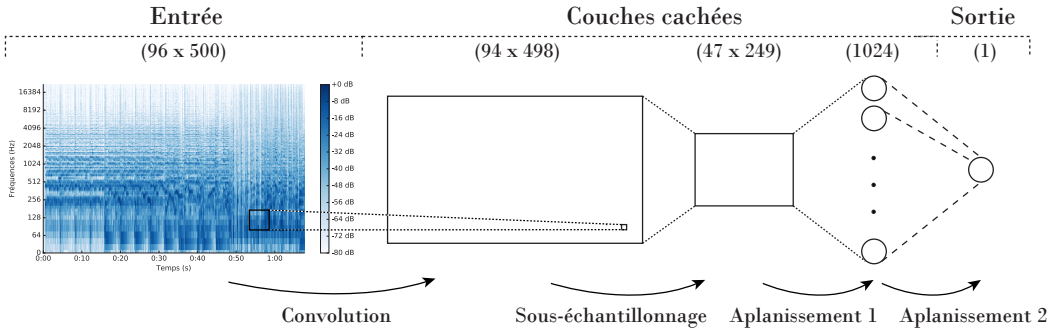


FIGURE 4.8 – Schéma d'un réseau de neurones. Voir également la figure 4 de l'architecture proposée par Murauer et Specht [2018].

Une couche convolutive ou Convolutional Neural Networks (CNN) constitue un autre type populaire de couche acyclique des réseaux de neurones. Un CNN utilise une corrélation croisée sous la forme d'une matrice de convolution de taille constante, appelée kernel, au lieu des multiplications matricielles du MLP. LeCun *et al.* [1995] ont été les premiers à démontrer l'utilité des CNNs et ce dans un cadre de reconnaissance des chiffres manuscrits pour la poste américaine. Il est possible d'appliquer une convolution en deux dimensions à une image (tel qu'un spectrogramme) et en une dimension à un signal temporel (tel qu'un enregistrement audio). Un réseau de neurones peut être constitué d'une succession de couches convolutives définies par le CNN et de couches denses définies par un MLP. Toutefois, l'architecture couramment utilisée concatène plusieurs CNNs ainsi qu'une couche dense censée résumer les informations extraites pour produire la sortie y. En utilisant cette architecture, Krizhevsky *et al.* [2012] ont été les pionniers en terme de reconnaissance d'objets dans des images [1] et les capacités de détection de leur algorithme ont dépassé les capacités humaines. De plus, un réseau de neurones entièrement convolutif permet de traiter des données en entrée et en sortie qui présentent des tailles variables, ce qui n'est pas possible avec une couche dense. Un réseau de neurones entièrement convolutif est notamment utilisé dans la segmentation

1. http://www.image-net.org/challenges/LSVRC/, consulté le 19 Avril 2018.

sémantique d'images [Long *et al.*, 2015]. Velarde [2017] détaille davantage les mécanismes régissant les CNNs.

Afin de prendre en compte des séquences temporelles, des réseaux de neurones présentant des connections cycliques ont été conçus. De tels réseaux sont appelés récurrents ou Recurrent Neural Networks (RNN) et permettent la modélisation de données séquentielles. Un RNN utilise par ailleurs des neurones capables de stocker l'événement précédant un élément donné afin de traiter cet élément courant. Pour cela, deux types principaux de neurones ont été proposés à travers les Long Short-Term Memory (LSTM) [Hochreiter et Schmidhuber, 1997] et les Gated Recurrent Unit (GRU) [Cho *et al.*, 2014].

L'architecture séquence-à-séquence d'un RNN [Sutskever *et al.*, 2014] peut modéliser une relation entre des données de tailles variables en entrée et en sortie. Cette architecture est notamment utilisée par l'outil de traduction automatique de Google [Wu *et al.*, 2016].

Les recherches sur les réseaux de neurones ont commencé il y a plus de 60 ans [Kleene, 1951] mais n'ont pris leur essor que récemment grâce notamment à l'augmentation des capacités de calcul. Les réseaux de neurones utilisent généralement quatre ordres de grandeur supplémentaires de temps de calcul par rapport à des méthodes traditionnelles d'apprentissage automatique [Bayle *et al.*, 2018a] et nécessitent en moyenne 20 millions d'opérations matricielles par instance traitée [1]. Or, il est possible de réduire le temps de calcul d'une méthode d'apprentissage profond d'un ordre de grandeur en effectuant les calculs sur une carte graphique plutôt que sur le processeur principal d'un ordinateur [Raina *et al.*, 2009; Bayle *et al.*, 2018a]. Il est également possible de diviser linéairement le temps de calcul d'une méthode d'apprentissage profond en parallélisant autant de cartes graphiques. Par ailleurs, plus l'architecture est profonde et meilleure est la généralisation, cependant cet approfondissement entraîne une augmentation du temps de calcul [Choi *et al.*, 2016b]. Pour des tâches liées à la musique néanmoins, des architectures trop profondes n'améliorent pas les performances [Choi *et al.*, 2016b] et un compromis autour de quatre couches est généralement mentionné dans la littérature [Jeon *et al.*, 2017; Valin, 2017]. Ioffe et Szegedy [2015] ont quant à eux proposé une couche de normalisation des données à positionner après chaque couche convolutive qui permet également de diminuer le temps de calcul tout en augmentant l'*accuracy*. Glorot *et al.* [2011] ont enfin proposé une fonction non-linéaire appelée Rectified Linear Unit (ReLU) qui accélère le calcul de la convergence d'un réseau de neurones par rapport à une fonction sigmoïde [Bengio *et al.*, 1994] présentée dans la figure 4.3. L'équation 4.4 définit les propriétés d'un ReLU.

1. https://www.tensorflow.org/tutorials/deep_cnn, consulté le 19 Avril 2019

$$f(x) = \begin{cases} 0 & \text{pour } x < 0 \\ x & \text{pour } x \geq 0 \end{cases} \tag{4.4}$$

où x est une matrice représentant un élément tel qu'un morceau ou une image.

Afin de diminuer l'effet de sur-apprentissage des méthodes utilisant des réseaux de neurones, Hinton *et al.* [2012] et Srivastava *et al.* [2014] proposent de réinitialiser aléatoirement un certain nombre de coefficients de chaque couche. Cette opération est qualifiée de *dropout*. Cette méthode s'est révélée être efficace et est maintenant utilisée dans les architectures de réseaux de neurones avec un taux de réinitialisation, ou *dropout rate*, compris entre 5 et 20% [Dieleman *et al.*, 2011; Schlüter et Grill, 2015; Kum et Nam, 2017]. Toutefois, cette méthode augmente le temps de convergence d'un algorithme d'apprentissage profond puisque la mise à jour des paramètres est bruitée par les réinitialisations aléatoires.

De plus, il a été montré que les perceptrons qui constituent le composant de base d'un réseau de neurones sont sensibles au déséquilibre d'une base de données [Branco *et al.*, 2016]. Cette constatation rend d'autant plus pertinent le rééquilibrage de la base de données, comme décrit dans la section 2.4.2.

Les méthodes d'apprentissage profond ont contribué à atteindre des performances constituant l'état de l'art dans plusieurs disciplines, notamment en analyse d'images [Litjens *et al.*, 2017] ou de vidéos [Wu *et al.*, 2017]. La section suivante décrit comment les méthodes d'apprentissage profond peuvent également être appliquées au domaine musical.

4.4.2 L'apprentissage profond appliqué à la musique

« Deep learning takes at least 100,000 examples. »
– Interview lors du projet *Google Brain*[1], 23 Octobre 2017, Jeff Dean

État de l'art de l'apprentissage profond appliqué aux signaux musicaux

Un travail collégial[2] destiné à recenser les travaux en apprentissage profond appliqués à la musique montre que ce domaine est en pleine expansion. L'état actuel du recensement de ces travaux est représenté dans la figure 4.9.

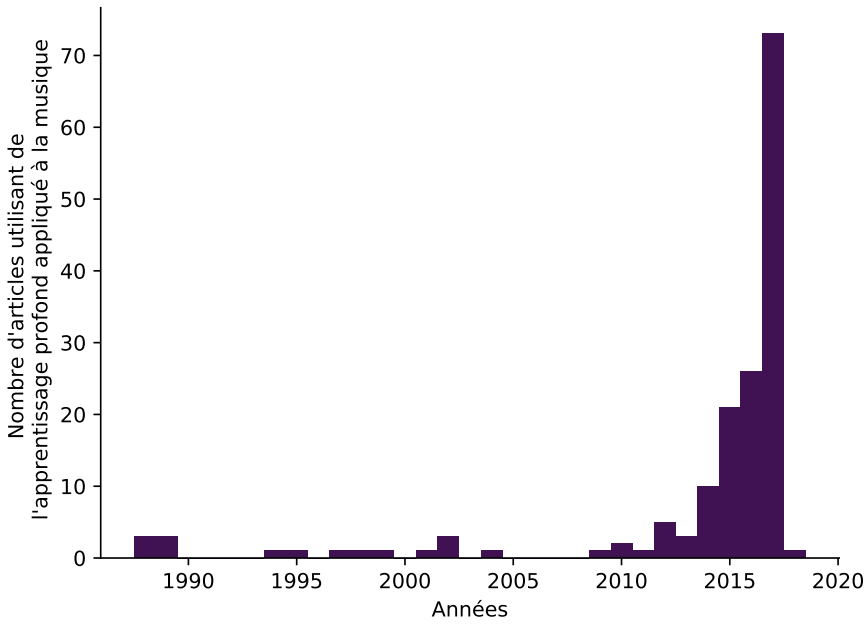


FIGURE 4.9 – Évolution annuelle du nombre d'articles appliquant l'apprentissage profond à des tâches musicales.

Ce recensement identifie trois types de représentations musicales[3] utilisés en entrée d'un réseau de neurones [Choi *et al.*, 2017a,b]. Il est tout d'abord

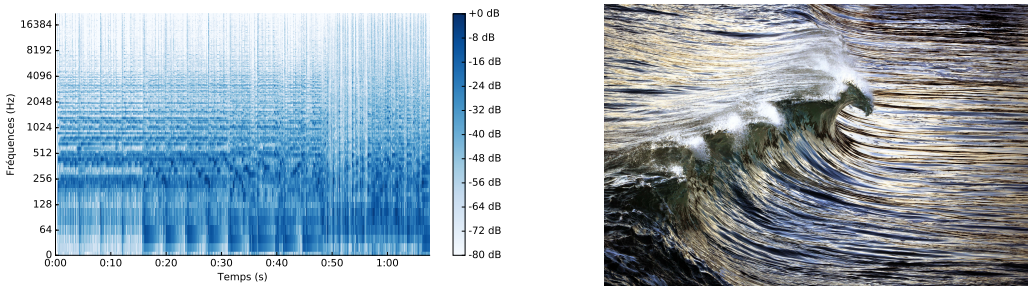
1. https://venturebeat.com/2017/10/23/google-brain-chief-says-100000-examples-is-enough-data-for-deep-learning/, consulté le 19 Avril 2018.
2. https://github.com/ybayle/awesome-deep-learning-music, consulté le 19 Avril 2018.
3. http://www.jordipons.me/deep-learning-architectures-for-audio-classification-a-personal-review/, consulté le 19 Avril 2018.

possible d'utiliser le signal audio brut, qui affiche une unique dimension [Dieleman et Schrauwen, 2014]. Le type de données le plus utilisé est toutefois le spectrogramme, dont un exemple est représenté dans la figure 4.10, qui est issu de la transformation d'un signal audio brut en une représentation en deux dimensions [Lee *et al.*, 2009; Pons *et al.*, 2017b; Choi *et al.*, 2016a]. Schlüter et Böck [2014] ont également proposé d'utiliser une représentation en trois dimensions correspondant à des échelles d'analyse temporelle différentes.

Lors de l'utilisation d'un CNN avec une entrée à une dimension, la taille des fenêtres d'analyse du signal varie de 3 à 512 échantillons [Lee *et al.*, 2009; Dieleman et Schrauwen, 2014]. En revanche, lors de l'utilisation d'un CNN avec une entrée à deux ou à trois dimensions, des filtres à deux dimensions sont utilisés. Dans ce cas, chacune des dimensions de la couche convolutive est comprise entre 1 et 90.

Contraintes techniques quant à l'utilisation de l'apprentissage profond pour l'audio

Les images constituent historiquement l'objet d'étude principal de l'apprentissage profond. Le spectrogramme est le type d'entrée le plus utilisé par les expériences de classification musicale qui utilisent des réseaux de neurones car il s'agit de la représentation qui se rapproche le plus d'une image. Un exemple de spectrogramme et d'image est présenté dans la figure 4.10.



(a) Spectrogramme (b) Image

FIGURE 4.10 – Exemples (a) d'un spectrogramme réalisé à partir du logiciel *librosa* [McFee *et al.*, 2015b] et des premières secondes du morceau *Elysium* d'Al Di Meola et (b) d'une image qui est une photographie personnelle.

Les architectures de neurones utilisées aujourd'hui dans la classification musicale ont été conçues pour de l'analyse d'image, or un spectrogramme ne peut pas être considéré comme une image et ce pour trois raisons. La première raison concerne le fait qu'un pixel correspond à un objet dans une image alors que dans le cas d'un spectrogramme un pixel représente une somme d'énergie fréquentielle provenant de plusieurs instruments. Une analogie consiste à utiliser un réseau de neurones afin d'analyser une image contenant plusieurs

fenêtres dont les vitres sont teintées différemment et superposées dans le but d'identifier à quelles fenêtres appartient un pixel donné.

La deuxième raison pour laquelle un spectrogramme ne peut pas être considéré comme une image concerne la présence de relations harmoniques sonores. Ces relations harmoniques impliquent que des pixels non-contigus proviennent de la même source audio alors que ce n'est pas le cas pour une image.

La troisième raison implique la signification des dimensions d'une image comparée à celles d'un spectrogramme [1]. Dans une image, il n'existe pas de signification associée à l'axe des abscisses et des ordonnées et le fait d'appliquer un effet miroir ou une rotation à l'image n'a pas d'impact sur le contenu de celle-ci. Dans le cas d'un spectrogramme en revanche, l'axe des abscisses représente le temps et l'axe des ordonnées les fréquences. L'application d'un effet miroir sur un spectrogramme a donc un impact sur le contenu audio du signal correspondant.

Les différences qui existent entre une image et un spectrogramme pourraient expliquer pourquoi les méthodes d'apprentissage profond proposées dans la littérature ne dépassent généralement pas les performances des algorithmes existants dans le domaine de la classification thématique musicale [Medhat *et al.*, 2017; Senac *et al.*, 2017; Choi *et al.*, 2017c]. Dans le cas contraire, l'amélioration est inférieure à 1% [Pons *et al.*, 2017a]. Les réseaux de neurones constituent donc un nouvel outil pour l'analyse musicale bien que leur impact sur les performances obtenues est à ce jour moindre par rapport à celui relatif à l'analyse d'image par exemple. Les méthodes de classification musicale utilisant des réseaux de neurones confirment toutefois que les approches proposées par l'état de l'art sont pertinentes puisque les structures musicales sous-jacentes qu'elles retrouvent dans leurs calculs autonomes sont les mêmes que celles décelées par les méthodes de l'état de l'art.

L'apprentissage profond qui utilise des réseaux de neurones est toutefois considéré comme une boîte noire puisque ses paramètres sont complexes à interpréter [Shwartz-Ziv et Tishby, 2017]. L'impact de ses paramètres sur l'analyse des données en entrée et sur la production des résultats en sortie est également difficile à interpréter [Coates *et al.*, 2011]. L'un des Horses les plus connus dans le cadre de l'apprentissage profond concerne le détecteur de tanks proposé en 1987 par une équipe de recherche de la DARPA [2]. Ce détecteur constitué de réseaux de neurones artificiels prédisait avec 100% de succès la présence d'un tank camouflé dans une image. L'apprentissage avait été réalisé à partir de photographies. Or, lors de la phase de test et de mise en situation, il a été constaté que le détecteur se trompait une fois sur deux. Il s'est avéré que les photographies qui avaient été utilisées lors de la phase d'apprentissage et qui contenaient réellement un tank camouflé avait été prises au cours d'une

1. https://towardsdatascience.com/whats-wrong-with-spectrograms-and-cnns-for-audio-processing-311377d7ccd, consulté le 19 Avril 2018.
2. https://neil.fraser.name/writing/tank, consulté le 19 Avril 2018.

même journée nuageuse, alors que celles ne contenant pas de tank avaient été prises par beau temps. La méthode de détection prédisait donc la couleur du ciel, qui s'avérait être corrélée dans ces photographies à la présence d'un tank. Puisque la méthode proposée prédisait des résultats corrects pour les mauvaises raisons, il s'agissait d'un Horse. Lors de l'utilisation d'un réseau de neurones, il est par conséquent complexe d'expliquer quelles sont les combinaisons de neurones et d'activations ayant mené à un tel résultat et donc de résoudre un problème de détection. Cependant, Choi *et al.* [2015] ont proposé un exemple de visualisation des activations des neurones pour une couche donnée. Cette visualisation permet donc d'appréhender les filtres audio qui sont appliqués au signal en entrée bien que l'utilisation qui est faite de ces filtres par le réseau de neurones demeure incertaine.

Limitations actuelles de l'apprentissage profond dans la génération automatique de listes de lecture

La génération d'une liste de lecture thématique parfaite revient à garantir 100% de précision sur un thème donné. Or, aucun élément dans la littérature ne semble indiquer l'existence d'expériences d'apprentissage profond qui atteignent cet objectif, ce qui peut s'expliquer comme suit. Les réseaux de neurones utilisés dans la littérature doivent en effet optimiser la recherche de paramètres, notamment pour la matrice W de l'équation 4.3. La fonction d'optimisation doit être évaluée à partir d'une fonction dérivable [Schulman *et al.*, 2015], comme l'*accuracy*, alors que dans le cas de la génération de listes de lecture musicale, elle devrait être optimisée vis-à-vis de la précision. Or, la précision ne constitue pas une fonction dérivable, ce qui limite l'utilisation des réseaux de neurones pour la génération automatique de listes de lecture musicale.
Afin de garantir une précision de 100% en utilisant un réseau de neurones, il serait également possible de modifier les coûts de classification de chaque thème, comme vu dans la section 4.3.3. Toutefois, ce type de méthode demeure difficilement interprétable et réalisable dans le cadre d'un réseau de neurones [Woods et Bowyer, 1997; Zhao *et al.*, 2011] puisqu'il répète plusieurs fois des opérations matricielles chronophages ayant différents coûts, sans garantie de convergence.

En outre, lors de l'utilisation des méthodes d'apprentissage machine présentées dans la section 4.1.2, les caractéristiques audio extraites des morceaux sont chargées dans la Random-Access Memory (RAM) puisque ce stockage accélère le temps de calcul. Ce chargement est rendu possible puisque la quantité de données à traiter est inférieure à la quantité de RAM disponible sur les ordinateurs grand public actuels, qui est de l'ordre de 4 Go. Afin de connaître l'ordre de grandeur de la quantité de RAM utilisée, il faut multiplier le nombre de caractéristiques audio par le nombre de morceaux ainsi que par le nombre

d'octets adressables par le processeur. Par exemple, la quantité de RAM utilisée pour l'expérience de Defferrard *et al.* [2017] dans la classification des genres inclus dans la FMA s'élève à 420 Mo de données chargées en mémoire (106 574 morceaux $*$ 518 caractéristiques $*$ 8 octets pour un processeur de 64 bits). En apprentissage profond, il n'est toutefois pas possible de charger l'intégralité des morceaux en mémoire et il est donc préférable de les charger par lots, ou *batches*, de 32 à 512 éléments [Keskar *et al.*, 2016]. Le chargement de lots de cette taille est couramment pratiqué puisque l'utilisation de lots d'une taille supérieure diminue les performances du modèle d'apprentissage profond. Ce phénomène demeure incompris [Keskar *et al.*, 2016]. L'utilisation d'un lot de moins de 32 données est par ailleurs néfaste pour les capacités de généralisation du réseau de neurones [Miron *et al.*, 2017; Oramas *et al.*, 2017]. Les réseaux de neurones n'utilisent en effet pas de caractéristiques audio extraites et doivent donc charger en mémoire l'intégralité du signal audio de chaque morceau. Or, un morceau dure en moyenne quatre minutes et est enregistré avec une fréquence d'échantillonnage de 44 100 Hz, ce qui correspond à 4 à 40 Mo de données à stocker en mémoire pour chaque morceau, en fonction du format de fichier d'enregistrement utilisé. La quantité de RAM utilisée lors du chargement en mémoire du signal audio intégral d'un morceau est donc environ 1 000 fois plus grande que lors du chargement en mémoire des caractéristiques audio de ce morceau. De plus, lors de l'utilisation d'une méthode d'apprentissage profond il faut également charger en mémoire une version du réseau de neurones pour chaque morceau, comme cela est défini par l'équation 4.5.

$$Q_{RAM} = L * T_{RN} * VF \tag{4.5}$$

où :

- Q_{RAM} est la quantité minimum de RAM requise,
- L est la taille du lot de données qui est généralement comprise entre 32 et 512,
- T_{RN} est la taille du réseau de neurones qui est généralement comprise entre 50 Mo et 1 Go (240 Mo pour Schlüter et Grill [2015] par exemple),
- VF est le nombre d'octets utilisés pour coder la virgule flottante et qui est généralement compris entre 4 et 8 octets respectivement sur des processeurs 32 et 64 bits.

L'ensemble des contraintes formellement décrites dans l'équation 4.5 implique l'utilisation de compromis lors de la conception et de l'utilisation d'un réseau de neurones. La quantité de RAM généralement disponible ainsi que le nombre d'octets utilisés pour les nombres à virgule flottante sont fixés. Il est donc possible de faire varier uniquement la taille des lots de données considérés ou la taille du réseau de neurones. Il en résulte que l'utilisation d'une architecture plus profonde nécessite de diminuer la taille du lot de données et

risque donc d'amoindrir les capacités de généralisation du réseau de neurones. La diminution de la taille du lot entraîne également une augmentation du temps de calcul. L'augmentation de la taille du lot permet en revanche d'accélérer le temps de calcul et d'augmenter les capacités de généralisation du modèle fourni par le réseau de neurones. Cela implique néanmoins de diminuer la profondeur du réseau et donc la complexité des modèles qu'il peut générer. Dans ce contexte de compromis complexes, les études sur l'identification du genre musical ont proposé de ne pas utiliser l'intégralité des morceaux de musique mais uniquement les 30 premières secondes qui les composent [Kereliuk *et al.*, 2015; Arumugam et Kaliappan, 2016]. L'hypothèse envisagée consiste à considérer un morceau comme une entité ne présentant qu'un seul genre de manière cohérente sur toute sa durée et que les 30 premières secondes de ce morceau sont suffisantes pour étudier son genre. Cette approximation ne se révèle toutefois pas toujours vérifiée [Sturm, 2014b] mais l'analyse des 30 premières secondes ou moins des morceaux est dorénavant admise lors d'études utilisant des réseaux de neurones [Zhang *et al.*, 2015; Hua, 2018].

Les cartes graphiques disponibles pour le grand public disposent généralement de 1 Go de RAM alors que celles disponibles sur les centres de calcul intensifs affichent entre 5 et 12 Go de RAM. Cette quantité de RAM est insuffisante pour charger les centaines de To du réseau de neurones d'analyse des 106 574 morceaux de la FMA. Il est néanmoins possible de doubler la taille du lot de données lorsque l'on double le nombre de cartes graphiques sur un même ordinateur. Il n'est toutefois pas possible de paralléliser plus de 8 cartes graphiques sur le même ordinateur. Une autre solution afin de disposer de plus de RAM consiste à utiliser celle des processeurs plutôt que celle des cartes graphiques. Chacun des processeurs de calcul du centre IT4Innovations utilisés dans l'expérience de classification musicale par Bayle *et al.* [2018a] dispose de 128 Go de RAM [1]. Il a néanmoins été précisé dans la section 4.4.1 qu'un processeur utilise deux ordres de grandeur de temps de calcul supplémentaires par rapport à une carte graphique lors de l'utilisation d'un réseau de neurones. Le temps de calcul requis se compte alors en mois voire en année, ce qui se révèle peu applicable en réalité. L'expérience présentée dans la section qui suit détaille cette différence de temps de calcul.

1. https://docs.it4i.cz/salomon/hardware-overview/, consulté le 19 Avril 2018.

4.4.3 Classification des instrumentaux et des chansons avec de l'apprentissage profond

Matériels et méthodes

> *« La notion de genre musical a traversé l'histoire de la musique et de la musicologie sous le couvert de définitions souvent aussi inconséquentes que capricieuses. »*
>
> – *Comprendre et identifier les genres musicaux*, 1997, Gérard Denizeau

Il n'existe pas de consensus quant à la définition du genre musical, cette section traite donc de la classification des instrumentaux et des chansons. Cette section décrit une expérience de comparaison de l'*accuracy* et du temps de calcul sur la base SATIN par deux méthodes de classification supervisée [Bayle *et al.*, 2018a].

La première méthode utilise les 13 MFCC proposés dans SOFT1 en tant que caractéristiques audio. La méthode des cinq plus proches voisins [Cover et Hart, 1967] implémentée dans le logiciel scikit-learn [Pedregosa *et al.*, 2011] est utilisée afin d'effectuer la classification supervisée des MFCC.

La seconde méthode utilise de l'apprentissage profond. Pour chaque morceau, un mel-spectrogramme est calculé puisqu'il s'agit d'une représentation en 2D communément utilisée pour la classification supervisée en musique [Choi *et al.*, 2016b] et notamment pour la détection du chant [Schlüter et Grill, 2015]. L'échelle mel est préférée à une échelle linéaire en tant que représentation en entrée d'un réseau de neurones et ce pour plusieurs raisons. L'échelle mel permet d'abord une représentation compacte de l'information fréquentielle et affiche une meilleure *accuracy* que l'utilisation d'une échelle linéaire tout en permettant une diminution du temps de calcul [Choi *et al.*, 2017a]. L'utilisation de l'échelle mel est également pertinente puisqu'elle reproduit la perception psychoacoustique des fréquences par l'oreille humaine [Moore, 2012]. Le logiciel *librosa* [McFee *et al.*, 2015b] est utilisé afin d'extraire les spectrogrammes avec une échelle mel sur 128 valeurs et afin de garantir la réplicabilité des résultats. La longueur de la fenêtre de la transformée de Fourier rapide est fixée à 4 096 échantillons avec un taux de recouvrement de 50%.

Le réseau de neurones est composé de trois couches de convolution [Velarde, 2017] et la méthode est nommée CNN. Une couche de normalisation des lots est utilisée après chaque couche de convolution afin d'accélérer le temps de calcul et d'améliorer l'*accuracy* [Ioffe et Szegedy, 2015]. La fonction d'activation retenue est une ReLU puisque celle-ci diminue le temps de calcul sans que l'*accuracy* n'en soit affectée [Glorot *et al.*, 2011]. Après chaque fonction d'activation, un sous-échantillonnage est effectué afin de récupérer la valeur du maximum local des activations et de détecter la présence de chant [Schlüter,

2016]. De plus, un taux de réinitialisation de 20% est fixé afin d'éviter un sur-apprentissage du réseau de neurones sur les données, [Srivastava *et al.*, 2014]. Finalement, une couche dense est utilisée afin de prédire si le morceau analysé est un instrumental ou une chanson. Les mises à jour du réseau de neurones sont assurées par un algorithme stochastique d'optimisation différentiable du gradient. Le chargement en mémoire vive est réalisé par lots de 32 morceaux.

Le réseau de neurones est implémenté en Python 3.6.1 avec la version 2.0.5 du logiciel Keras [Chollet, 2015] et la version 1.1.0 de TensorFlow [Abadi *et al.*, 2016]. Pour cette seconde méthode, le temps de calcul est mesuré pour le lancement du programme sur un processeur et une carte graphique. La carte graphique est une NVIDIA Kepler K20 avec 5 GB de RAM et le processeur est un Intel Sandy Bridge E5-2665 2.4 GHz avec 64 GB de RAM. Une validation croisée n'est pas utilisée dans cette expérience car une méthode d'apprentissage profond garantit une faible variance [Choi *et al.*, 2017a].

Les calculs de cette expérience ainsi que ceux des expériences préliminaires ont été réalisés sur l'infrastructure fournie par le mésocentre de calcul intensif aquitain de l'Université de Bordeaux et l'Université de Pau et des Pays de l'Adour ainsi que sur l'infrastructure du centre IT4Innovations en République Tchèque.

Résultats

Le tableau 4.4 détaille l'*accuracy* atteinte par chacune des méthodes de classification considérées. Le temps de calcul est également affiché et inclut tous les calculs, les transferts des données en mémoire et l'affichage du programme.

Tableau 4.4 – Comparaison de l'*accuracy* et du temps de calcul pour la classification des instrumentaux et des chansons de la base SATIN par deux méthodes de classification supervisée.

Méthodes	*Accuracy* (%)	Temps de calcul
5-NN et MFCC sur processeur	89,0	5 s
CNN sur processeur	81,4	13 h 26 m 6 s
CNN sur carte graphique	81,4	1 h 30 m 49 s

La méthode CNN affiche une *accuracy* inférieure à celle de la méthode d'apprentissage automatique traditionnelle. La méthode CNN effectuant les calculs sur une carte graphique nécessite trois ordres de grandeur supplémentaires de temps de calcul que la méthode d'apprentissage automatique traditionnelle. Pour des calculs effectués sur le processeur, la méthode CNN nécessite neuf fois plus de temps de calcul que sur une carte graphique.

Discussion et Conclusion

L'expérience décrite dans la section précédente compare l'*accuracy* et le temps de calcul de deux méthodes de classification supervisée lors de l'annotation automatique des morceaux de la base de données SATIN. La comparaison indique que la méthode d'apprentissage profond considérée obtient une *accuracy* inférieure à celle de la méthode des plus proches voisins, tout en nécessitant davantage de temps de calcul. Les résultats obtenus ne permettent pas de conclure quant à la supériorité systématique d'une famille de méthodes par rapport à une autre. Ces résultats illustrent cependant qu'une méthode d'apprentissage profond possédant une architecture type peut rencontrer des difficultés de modélisation de caractéristiques musicales pertinentes lors d'une classification supervisée. La complexité de la modélisation des mécanismes sous-jacents peut donc entraîner des résultats qui sont surpassés par une méthode plus simple et plus rapide. En outre, la méthode d'apprentissage profond utilisée ne constitue pas l'état de l'art dans le domaine de la classification musicale mais une architecture type couramment utilisée. L'objectif de son utilisation n'est pas de proposer une méthode affichant de meilleurs résultats que l'état de l'art mais de fournir des tendances quant aux performances et temps de calcul qu'il est actuellement possible d'obtenir. Il faut souligner que les méthodes d'apprentissage profond appliquées à la classification supervisée des thèmes musicaux utilisent des architectures et représentations de données en entrée qui ont été développées pour le traitement de l'image. Les méthodes d'apprentissage profond ont en effet récemment été utilisées dans le domaine de l'analyse musicale et il semble par conséquent que de nouvelles architectures plus spécialisés au traitement de l'audio devraient améliorer les performances obtenues par ces méthodes. La prise en considération, lors de la conception d'une méthode d'apprentissage profond, du nombre de paramètres pouvant varier laisse présumer que plusieurs expériences restent encore à mener afin d'évaluer les avancées et limites de ces méthodes dans le traitement automatique de la musique.

L'exécution plus rapide des calculs d'un réseau de neurones sur une carte graphique par rapport à un processeur central est un fait bien connu [Raina *et al.*, 2009] qui est confirmé par nos résultats. De plus, ces résultats permettent de connaître, pour une même tâche, l'ordre de grandeur du temps de calcul nécessaire pour une méthode d'apprentissage profond par rapport à une méthode d'apprentissage machine traditionnelle. L'architecture de réseau de neurones proposée contient trois couches de convolution. Or, l'utilisation de davantage de couches peut améliorer les performances de classification bien que cela augmente le temps de calcul requis [Choi *et al.*, 2016b]. Toutefois, en classification thématique musicale, des architecture trop profondes, c'est-à-dire contenant plus de six couches, n'améliorent pas significativement les performances [Choi *et al.*, 2016b] et le nombre de couches utilisé dans cette expérience est en ac-

cord avec celui de la littérature [Jeon *et al.*, 2017; Valin, 2017]. Murauer et Specht [2018] ont également réalisé une expérience de comparaison entre deux méthodes d'apprentissage profond et des méthodes de classification supervisée traditionnelles. La tâche de comparaison a consisté à effectuer de la classification des genres de la base FMA. Ces auteurs ont montré que les résultats obtenus avec des méthodes traditionnelles étaient supérieurs à ceux obtenus avec des méthodes d'apprentissage profond et ce pour différentes architectures de réseaux de neurones. Les résultats de classification semblent être limités par l'architecture utilisée et il serait pertinent, afin d'améliorer les performances des méthodes, de chercher de nouvelles représentations des morceaux à utiliser en entrée d'un réseau de neurones.

4.4.4 Conclusion sur les algorithmes d'apprentissage profond

> *« Sometimes it's better to leave something alone, to pause, and that's very true of programming. »*
>
> – Interview de la BBC [1], 15 Janvier 2017, Joyce Wheeler

Les réseaux de neurones existent en informatique depuis plusieurs dizaines d'années mais leur utilisation en tant que méthodes de classification supervisée n'a eu de très fort impact positif sur les résultats d'analyse automatique d'images que très récemment. Les résultats de classification musicale par ces méthodes n'ont toutefois pas encore dépassé significativement ceux de l'état de l'art. Les méthodes d'apprentissage profond revêtent en effet un intérêt particulier lorsqu'il existe peu de connaissances *a priori* quant aux données traitées. Leur impact sur les performances de classification est en revanche limité lorsque, comme dans le cas de la musique, il existe un état de l'art relativement conséquent quant à la compréhension du processus d'analyse de ces données. Les études les plus récentes en classification musicale utilisent les réseaux de neurones en complément d'autres méthodes plutôt qu'en remplacement intégral de celles-ci [Eghbal-Zadeh *et al.*, 2016]. Il n'en demeure pas moins que l'apprentissage profond constitue un nouvel outil dont le potentiel et les limites en musique devraient être mieux perçus avec le développement de dispositifs de calcul plus puissants et mieux distribués.

1. http://www.bbc.com/news/technology-38103893, consulté le 19 Avril 2018.

4.5 Conclusion sur les algorithmes d'apprentissage automatique

« On ne devrait même pas appeler ça intelligence, c'est de l'algorithmie sophistiquée. »

– Interview sur Europe 1, 25 Février 2018, Cédric Villani

La section 4.1 de ce chapitre a décrit les méthodes de classification supervisée utilisées pour traiter les morceaux et leurs caractéristiques afin de générer des listes de lecture musicale thématiques. L'expérience présentée dans la section 4.3 suit les étapes décrites dans la figure 4.11 qui résume la méthodologie permettant de constituer des listes de lecture musicale. La section 4.3 a en outre détaillé les avantages et inconvénients des méthodes de classification existantes dans le cadre de la détection de thèmes musicaux dans de grandes bases de données musicales. L'expérience proposée dans cette section a utilisé avec succès une nouvelle méthode de détection du chant dont les performances surpassent une modélisation par chaîne de Markov. La méthode proposée fournit à ce jour la meilleure précision pour la génération de listes de lecture instrumentale. Le traitement d'une base de données déséquilibrée demeure néanmoins toujours un défi. La section 4.4 a enfin souligné l'intérêt grandissant des méthodes d'apprentissage profond ainsi que les nombreuses interrogations restantes quant à leur application à des tâches de classification musicale.

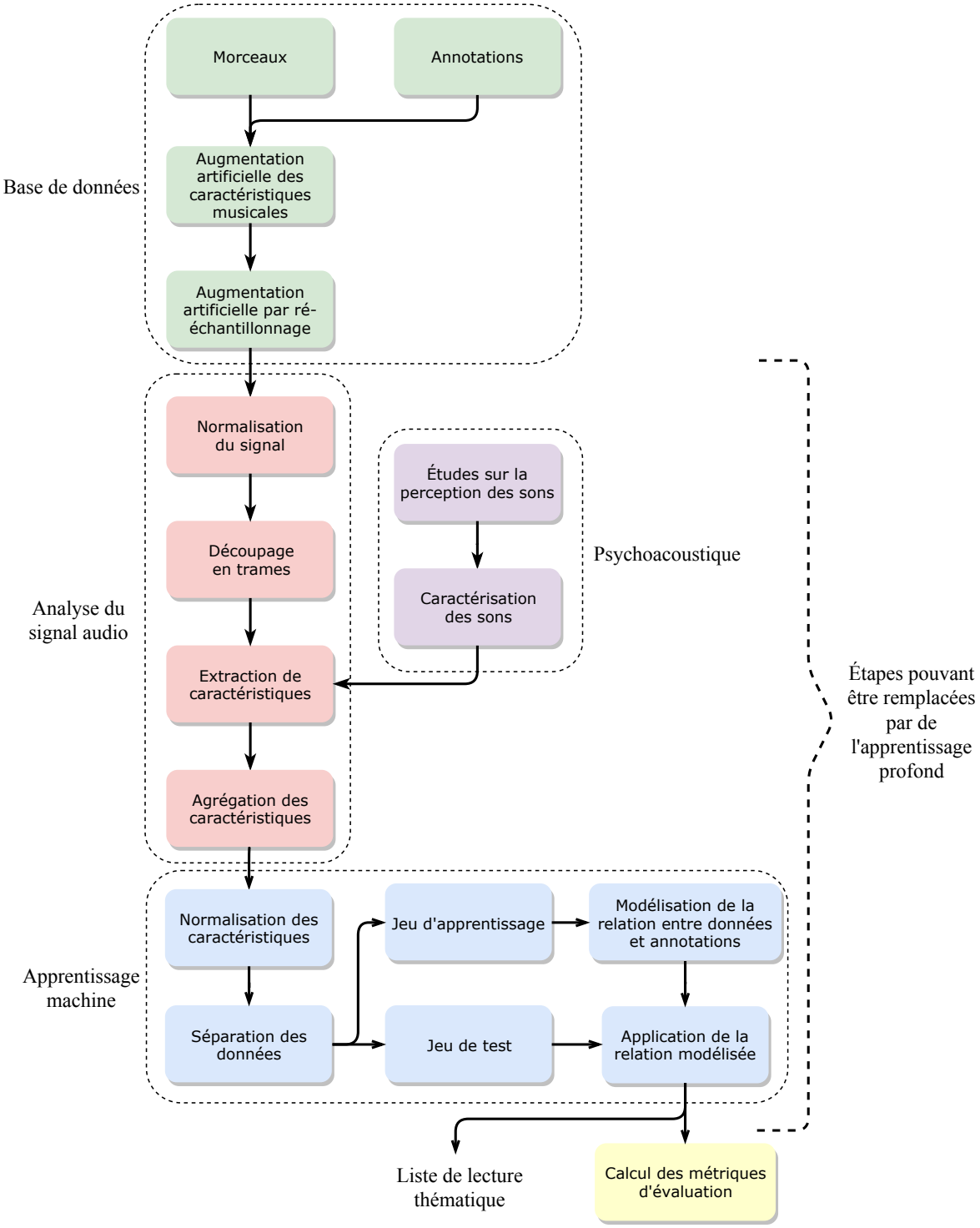


FIGURE 4.11 – Détails de l'ensemble des étapes de la chaîne de traitement de morceaux permettant de constituer des listes de lecture musicale.

5

Apports et Perspectives

« I believe that every scientist who studies music has a duty to keep his or her love to music alive. »
– *Exploring the musical mind*, 2005, John Sloboda

L'un des objectifs principaux de ce mémoire a été d'expliquer quelles sont les étapes clés qui constituent la chaîne de traitement automatique des morceaux afin de classer ces derniers par thèmes. Ce mémoire a particulièrement détaillé quels sont l'état de l'art et les solutions envisagées afin d'appréhender le contexte musical actuel et de faire face à l'expansion des données musicales disponibles. La section 5.1 résume les contributions apportées par cette thèse à la classification supervisée musicale et à la génération automatique de listes de lecture à partir de caractéristiques audio. La section 5.2 fait état des vérifications à effectuer afin d'éviter de produire une méthode qui ne soit un Horse. La section 5.3 indique finalement les perspectives d'exploration de recherche en analyse thématique musicale.

5.1 Contributions

« Un chercheur doit avoir conscience du peu de ce qu'il a trouvé, mais il a droit d'estimer que ce peu est immense. »
– *Inquiétudes d'un biologiste*, 1967, Jean Rostand

Le premier chapitre a introduit les différents contextes et concepts autour des données musicales et de leur consommation.

Le deuxième chapitre a décrit les enjeux autour des bases de données musicales et contient un état de l'art des bases de données existantes dans le cadre de la génération de listes de lecture musicale thématiques. Les problématiques entourant les bases de données musicales de recherche ont été soulignées, notamment en ce qui concerne la faible quantité de données disponibles dans le domaine académique. Afin de pallier cette faible quantité de données musicales dédiées à la recherche, des méthodes d'augmentation artificielle de base de données existent et ont été évaluées. Ce chapitre a proposé deux nouvelles bases

de données, qui sont accompagnées d'une API de gestion des morceaux et de leur annotations. La base de données SATIN a fait l'objet d'une publication [Bayle *et al.*, 2017a] et d'une présentation lors du 15ème *International Workshop on Content-Based Multimedia Indexing* à Florence en Italie. La base de données Kara1K a été réalisée en partenariat avec des chercheurs de République Tchèque et la publication correspondante [Bayle *et al.*, 2017b] a été présentée lors du 19ème *IEEE International Symposium on Multimedia Content-Based Multimedia Indexing* à Taichung à Taiwan, au cours duquel elle a reçu une mention honorable de l'IEEE.

Le troisième chapitre a permis d'expliquer l'importance de l'extraction et du développement de caractéristiques musicales pertinentes afin de décrire le contenu des bases de données musicales. Le développement de caractéristiques musicales est en partie possible grâce à l'étude de certains phénomènes psychoacoustiques qui entrent en jeu dans la perception des sons. Ce chapitre a présenté de nouvelles expériences de psychoacoustique qui ont été publiées dans la revue *Hearing Research* [Demany *et al.*, 2017]. La compréhension de certains phénomènes psychoacoustiques favorise la constitution de nouvelles techniques de traitement du signal sonore qui extraient des caractéristiques audio des morceaux afin de décrire ces derniers. Les caractéristiques audio permettent de décrire différentes représentations musicales appartenant aux catégories définies dans la section 1.1.2, tels que le genre ou l'émotion. Ce chapitre s'est concentré sur le thème de la présence de chant dans les morceaux, grâce auquel il est possible de distinguer les chansons des instrumentaux. La méthode PEA proposée a été évaluée comme étant plus performante que les méthodes de l'état de l'art et ce dans le contexte d'un passage à l'échelle. Les méthodes d'agrégation utilisées par PEA permettent de passer de l'analyse locale d'un morceau à une description globale pertinente de celui-ci. Les méthodes d'agrégation proposées ont notamment fait l'objet d'un article [Bayle *et al.*, 2016] présenté lors des 23èmes Journées d'Informatique Musicale à Albi en France. Enfin, sur les expériences menées, la nouvelle méthode PEA qui utilise ces méthodes d'agrégation affiche moins de faux positifs que les méthodes de l'état de l'art dans la classification des instrumentaux et des chansons.

Le quatrième chapitre a décrit l'utilisation des caractéristiques musicales extraites afin de trier les morceaux par thème et donc de faciliter les recommandations musicales et la proposition de listes de lecture. Ce chapitre a introduit les algorithmes d'apprentissage automatique et plus particulièrement ceux de classification utilisés en musique. L'objectif de l'utilisation de ces méthodes est de proposer des listes de lecture musicale thématiques qui minimisent le nombre de faux positifs, c'est-à-dire le nombre de morceaux hors thème. Ce chapitre, qui a fait l'objet d'un article [Bayle *et al.*, 2018b], a démontré que de meilleures listes de lecture peuvent être générées lorsque l'algorithme d'apprentissage machine se concentre uniquement sur un thème parmi l'ensemble de ceux présents dans une base donnée. Par ailleurs, cette amélioration est

d'autant plus importante que la représentation du thème est minoritaire dans une base de donnée. La plupart des thèmes musicaux sont en effet minoritaires au sein des bases de données industrielles. Ce chapitre a également consigné des conclusions quant aux méthodes d'apprentissages machine et profond et qui sont exposées dans un article [Bayle *et al.*, 2018a] publié dans la revue *Springer Multimedia Tools and Applications*.

Ce mémoire a enfin soulevé l'existence de méthodes d'analyse musicale biaisées que sont les Horses, qui sont néfastes à la reconnaissance thématique musicale. De par leur occurrence souvent sous-estimée, il apparaît pertinent de vérifier méthodiquement plusieurs points permettant d'éviter de créer une méthode de type Horse. Ces points sont résumés dans la section suivante.

5.2 *Horses*

La section 1.1.4 a permis de définir la notion de Horse en tant que méthode qui semble capable de résoudre une tâche mais qui utilise en réalité des caractéristiques non pertinentes pour cette résolution. Afin d'éviter de produire un algorithme de type Horse, plusieurs vérifications ont été étudiées dans les différents chapitres ainsi que dans les différentes étapes de génération d'une liste de lecture musicale et sont résumées dans la liste ci-après.

- Comparer les résultats produits par une méthode donnée à ceux produits par des prédictions aléatoires,
- Comparer les résultats produits par une méthode lors de l'utilisation des données de la base initiale à ceux produits par cette même méthode utilisant des données générées aléatoirement,
- Permettre la reproduction de la méthode proposée en fournissant le code source, ce qui facilite le travail de recherche d'erreurs ou de bugs par la communauté scientifique [Drummond, 2009],
- Utiliser plusieurs bases de données musicales afin de garantir la diversité des méthodes de constitution et d'annotation de données ainsi que la diversité des morceaux représentés,
- Utiliser des méthodes d'augmentation de données,
- Utiliser des méthodes de validation croisée sur une base de données et sur plusieurs bases de données,
- Normaliser le signal audio ou les caractéristiques audio à traiter,
- Favoriser la méthode la plus proche de la loi de parcimonie,
- Vérifier l'égalité des performances sur les jeux d'apprentissage et de test afin de réduire le phénomène de sur-apprentissage.

5.3 Perspectives

> *« Un type comme vous... avec de l'instruction et tout... Vous pourriez contribuer à un grand projet scientifique, mais vous préférez jouer au gratteur de guitare. Est-ce que ça vous amuse ? Est-ce que vous y trouvez votre compte en termes de satisfaction personnelle ? »*
>
> – *Pacific Park*, 1987, Philip K. Dick

Ce mémoire a soulevé les problématiques concernant l'accès à plusieurs grandes bases de données musicales non structurées afin de permettre une écoute musicale sélective aux auditeurs du XXI^{ème} siècle. Afin de faire face à cette problématique, les études existantes ont été détaillées dans les différents chapitres en ce qui concerne l'extraction de caractéristiques audio et l'utilisation de méthodes de classification supervisées pour l'identification des thèmes musicaux. Il serait par la suite pertinent d'évaluer l'explication accordée à chacun de ces algorithmes et plus généralement aux algorithmes de traitement de données [Ribeiro *et al.*, 2016]. À long terme, la capacité d'un chercheur à expliquer et à rendre compte du fonctionnement d'un algorithme et de son impact sur la société est à prévoir [Doshi-Velez et Kim, 2017]. Tout en tâchant d'éviter les Horses et d'assurer un cadre propice aux sciences ouvertes [1], une telle démarche d'interprétation des algorithmes est de plus souhaitable.

Sur le plan technique, plusieurs expériences pourraient être menées afin d'approcher la précision de 100% de classification sur un thème donné. Elles permettraient donc de générer une liste de lecture musicale minimisant le nombre de morceaux ne correspondant pas au thème choisi. Par ailleurs, le fait d'assurer 100% de précision de classification sur un thème aurait un impact sur d'autres disciplines. En médecine par exemple, cela permettrait d'améliorer la qualité de détection et de classification des pathologies au sein d'une base de données médicales. Pour cela, il serait judicieux que des études futures se concentrent sur les problématiques suivantes.

Les bases de données affichent généralement un déséquilibre thématique des données naturelles ou synthétiques qu'elles contiennent. Cette problématique du déséquilibre de représentation des thèmes existant dans une base de données industrielle ou académique est bien connue. Les performances de classification obtenues à l'aide d'une base de données déséquilibrée sont en effet inférieures à celles obtenues pour une base de données équilibrée [Kotsiantis *et al.*, 2006; Yang et Wu, 2006; Longadge et Dongre, 2013]. L'impact du déséquilibre thématique d'une base de données musicale sur les méthodes d'apprentissage profond n'a toutefois pas encore été étudié. Ce déséquilibre a en revanche été étudié par des méthodes d'apprentissage peu profond [Erhan

1. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access, consulté le 19 Avril 2018.

et al., 2010; Huang *et al.*, 2016] en ce qui concerne l'analyse d'image mais les conclusions obtenues ne peuvent pas être transposées au domaine de l'analyse musicale. L'impact du déséquilibre de représentation des thèmes sur les résultats d'algorithmes d'apprentissage profond pourrait être étudié grâce à une expérience à deux conditions. Dans la première condition, la base de données serait par exemple constituée de 10 000 instrumentaux et d'autant de chansons et dans la seconde condition le nombre de chansons serait dix fois supérieur au nombre d'instrumentaux. Dans les deux conditions, la mesure de la précision et du rappel des résultats d'une méthode d'apprentissage profond s'effectuerait dans le cadre d'une validation croisée. Cette expérience permettrait de quantifier dans quelle mesure les réseaux de neurones sont robustes à une égalité de présence de thèmes dans une base. Dans une telle expérience, SATIN pourrait être utilisé en tant que base de données musicales de test. Outre le domaine musical, les disciplines traitant de données sensibles au sein de bases de données déséquilibrées, telle que la médecine [Wan *et al.*, 2014], pourraient bénéficier des conclusions d'une telle expérience.

L'expérience proposée dans le paragraphe précédent a permis d'étudier l'impact de la disproportion entre deux thèmes au sein d'une base de données sur la classification de ces données. Il est également possible de généraliser cette étude à une disproportion entre plus de deux thèmes. L'expérience de classification thématique proposée par Pons *et al.* [2017a] montre notamment que les capacités de généralisation d'un réseau de neurones s'améliorent lorsque davantage de thèmes sont ajoutés à la base et présentés au réseau durant la phase d'apprentissage. Les résultats de l'expérience de classification des instrumentaux et des chansons proposée dans la section 4.3 indiquent au contraire que pour des méthodes d'apprentissage traditionnelles il est préférable de se concentrer sur un thème unique [Bayle *et al.*, 2018b]. Une expérience pourrait être menée afin d'établir l'influence du nombre de thèmes utilisés en tant qu'annotations lors de la phase d'apprentissage sur les résultats de la méthode d'apprentissage automatique. Pour cela, la classification thématique musicale devrait être réalisée en prenant en compte deux conditions pour lesquelles une même base de données serait néanmoins utilisée. Dans la première condition, le réseau de neurones aurait accès aux thèmes de tous les morceaux et la précision de classification serait calculée pour chaque thème. Dans la seconde condition, le réseau de neurones aurait accès uniquement aux annotations concernant un thème donné et tous les autres morceaux n'appartenant pas à ce thème seraient rassemblés et annotés comme appartenant à une même classe. La précision obtenue pour un thème dans le cadre de ce dernier choix binaire pourrait être différente de celle obtenue dans la première condition, qui affiche un choix multiple. Le résultat de cette expérience permettrait de conclure quant à une éventuelle différence de performance dans la considération simultanée d'un ou de plusieurs thèmes par une méthode d'apprentissage automatique.

La recherche d'une fonction de coût dérivable permettant de maximiser la

précision plutôt que l'*accuracy* dans le cadre de l'apprentissage profond est un problème qui a été soulevé dans la section 4.4.2. La recherche de nouvelles métriques d'optimisation des méthodes d'apprentissage profond semble donc être pertinente puisqu'elle permettrait de mieux comprendre et appréhender les limites de l'apprentissage profond sans néanmoins porter préjudice à la qualité des prédictions sur un thème. Une solution à ce problème permettrait de générer une liste de lecture musicale thématique minimisant le nombre de morceaux hors thème et donc de faux positifs mais pourrait également être utilisée dans d'autres domaines telle que l'analyse automatique des images.

Concernant les réseaux de neurones, il semble pertinent de proposer de nouvelles méthodes d'augmentation de données qui permettraient de mieux généraliser la relation entre les caractéristiques audio et un thème donné dans une base, tout en diminuant les biais de représentativité thématiques. Une méthode d'augmentation à partir d'une modification des phases des fréquences du signal serait envisageable par les méthodes d'apprentissage profond qui utilisent directement le signal audio en entrée. La modification de phase pourrait notamment être implémentée dans le logiciel [1] open-source proposé par McFee *et al.* [2015a]. La modification des phases des fréquences du signal doit cependant être effectuée avec parcimonie et n'est pas compatible avec toutes les tâches de reconnaissance musicale car elle modifie le signal sur le plan perceptif. Une méthode d'augmentation des phases aurait donc un impact négatif sur une tâche de reconnaissance des émotions dans le chant mais pourrait par exemple se révéler pertinente dans une tâche de classification des instrumentaux et des chansons.

Afin d'améliorer les résultats actuels de la classification des instrumentaux et des chansons, il est possible d'améliorer la méthode de détection du chant à l'échelle de la trame. Pour cela, il est tout d'abord préférable de comparer les différentes méthodes existantes entre elles [Nwe *et al.*, 2004; Lukashevich *et al.*, 2007; Ramona *et al.*, 2008; Regnier et Peeters, 2009; Lehner *et al.*, 2014; Leglaive *et al.*, 2015; Lehner *et al.*, 2015; Schlüter et Grill, 2015; Schlüter, 2016]. Ces méthodes de détection du chant pourraient de plus bénéficier d'un traitement préalable qui estimerait quelles sont les pistes chantées et instrumentales pour chaque morceau [Roma *et al.*, 2016; McVicar *et al.*, 2016; Fan *et al.*, 2017; Jansson *et al.*, 2017; Mimilakis *et al.*, 2017; Stoller *et al.*, 2018]. L'étude de l'impact d'un tel pré-traitement et de la gestion de deux signaux sur la précision de classification reste donc à évaluer.

En outre, l'utilisation de deux signaux ou bien d'un signal musical stéréophonique en entrée n'a pas été considéré dans le cadre de l'analyse musicale par une méthode d'apprentissage profond. Une méthode d'apprentissage profond pourrait en effet dans ce contexte tirer profit des informations de panoramique

1. https://github.com/bmcfee/muda, consulté le 19 Avril 2018.

sonore provenant des deux canaux audio. Une étude pourrait être menée afin de quantifier le gain en performances de l'utilisation d'un tel type d'entrée. La source stéréophonique pourrait provenir d'un morceau enregistré en stéréophonie ou bien des deux pistes extraites (chantée et instrumentale) par les méthodes de séparation de sources.

Toutefois, ce doublement de la taille des données en entrée rejoint la problématique soulignée dans la section 4.4.2 concernant le chargement en mémoire vive d'une quantité limitée d'informations. Il a en effet été démontré dans la section 4.4.2 qu'en raison de limites matérielles actuelles, les études utilisant des méthodes d'apprentissage profond ne considèrent que les 30 premières secondes d'audio de chaque morceau. La recherche de solutions afin d'analyser l'intégralité d'un morceau se révèle donc pertinente dans ce cadre. L'idée suggérée consisterait à paralléliser une méthode d'apprentissage profond afin de distribuer les calculs requis sur plusieurs processeurs et cartes graphiques [Dean *et al.*, 2012; Sridharan *et al.*, 2018]. Cette distribution des calculs devrait plus précisément considérer des stratégies d'optimisation asynchrone, de répartition des données et des parties du réseau de neurones sur plusieurs processeurs et cartes graphiques [Dean *et al.*, 2012; Recht *et al.*, 2011]. Une répartition des tâches efficiente demeure toutefois un *challenge* lors de l'utilisation de plus de 12 machines [Keuper et Preundt, 2016] et de nombreux travaux restent à mener dans une telle distribution efficiente de ces calculs.

5.4 Conclusion générale

> *« Until I die there will be sounds. And they will continue following my death. One need not fear about the future of music. »*
> – *Experimental Music*, 1957, John Cage

Ce mémoire de thèse a consigné et a détaillé les travaux ainsi que les réflexions développés lors de mon doctorat. Le premier chapitre a tout d'abord dépeint le contexte musical actuel ainsi que les problématiques de recherche qui en découlent. Les chapitres suivants ont exposé les méthodes existantes de classification supervisée de mégadonnées musicales et ont été structurés autour de trois concepts fondamentaux. Le premier concept concernait les bases de données musicales, le deuxième concept a traité de la description audio des morceaux et le troisième concept a gravité autour des algorithmes d'apprentissage machine. Ces trois concepts ont notamment été appliqués à la classification supervisée thématique instrumentale. Toutefois, d'autres tâches incluses dans diverses disciplines auraient pu servir de support et bénéficient des conclusions des travaux proposés. Pour chacun de ces trois concepts, les propositions de méthodes ont permis d'améliorer l'état de l'art dans l'analyse et la compréhension de l'apprentissage automatique sur des caractéristiques audio multi-échelles.

Annexe A

Liste de lecture musicale d'accompagnement du mémoire

« Life has a soundtrack. »
– Festivais de Verão: Para Além da Música, Uma Experiência, 6 Août 2005, A. R. Gomes

La lecture de ce mémoire peut s'effectuer en écoutant une liste de lecture musicale disponible sur Deezer [1] ou sur Spotify [2]. Les morceaux sont présentés ci-après et ont été choisis parce qu'ils correspondent à des exemples précis d'illustration d'éléments décrits et utilisés dans cette thèse ou bien pour des raisons personnelles.

- *Fearless* - Pink Floyd
- *My Research* - Stand High Patrol
- *Hymne d'Ugarit* - Joan Borrell [3]
- *Crowd Chant* - Joe Satriani
- *Pigs (Three Different Ones)* - Pink Floyd [4]
- *Summertime* - Janis Joplin
- *Boom Boom* - John Lee Hooker
- *Love Is Like A Violin* - Barclay James Harvest
- *Trains* - Porcupine Tree
- *House Of The Rising Sun* - Chanson populaire
- *Le Pénitencier* - Johnny Hallyday

1. http://www.deezer.com/playlist/4046702022, consulté le 19 Avril 2018.
2. https://open.spotify.com/user/zikbone/playlist/4ytJ3BrPf3iTeJE16hIrEj, consulté le 19 Avril 2018.
3. Plus vieille chanson connue à ce jour, à écouter sur https://lejournal.cnrs.fr/articles/sur-un-air-de-musique-antique, consulté le 19 Avril 2018.
4. Le son de la *talking box* est à écouter sur https://youtu.be/4DmsvdCPD2Y?t=2m11s. D'autres exemples de *talking box* sont à écouter sur http://www.deezer.com/fr/playlist/16461734, consulté le 19 Avril 2018.

- *Little Wing* - Jimi Hendrix
- *Little Wing* - Stevie Ray Vaughan
- *Smells Like Teen Spirits* - Nirvana
- *Teen Sprite* - Nirvirna [1]
- *Face À La Mer* - Calogero et Passi
- *Calojira* - Ultra Vomit
- *The Sound Of Silence* - Simon and Garfunkel
- *The Sound Of Silence* - Disturbed
- *Time* - Pink Floyd
- *Time* - Easy Star All-Stars
- *Sandiala* - Panda Dub
- *Smile Is The Key* - Panda Dub
- *Where Do I Belong* - Infected Mushroom
- *I Want To Break Free* - Queen
- *Space Cadet* - Kyuss
- *Serendipity* - St.Tropez Lounge Music
- *Elysium* - Al Di Meola

1. Version majeure de la chanson mineure *Smells Like Teen Spirits* par Nirvana à écouter sur https://subtletv.com/baajUJY/Nirvirna_-_Teen_Sprite, consulté le 19 Avril 2018.

Annexe B

Liste des publications

Bayle Y., Hanna P. et Robine M., 2016. Classification à grande échelle de morceaux de musique en fonction de la présence de chant. Dans *Actes des 23èmes Journées d'Informatique Musicale*, pages 144–152. Albi, France. http://jim2016.gmea.net/?ddownload=450

Bayle Y., Hanna P. et Robine M., 2017. SATIN: A persistent musical database for music information retrieval. Dans *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, pages 1–5. Florence, Italy. https://www.researchgate.net/publication/317824533_SATIN_A_Persistent_Musical_Database_for_Music_Information_Retrieval

Bayle Y., Maršík L., Rusek M., Robine M., Hanna P., Slaninová K., Martinovič J. et Pokorný J., 2017. Kara1k: A karaoke dataset for cover song identification and singing voice analysis. Dans *Proceedings of the 19th IEEE International Symposium on Multimedia Content-Based Multimedia Indexing*, pages 1–8. Taichung, Taiwan. http://ieeexplore.ieee.org/document/8241597/

Demany L., **Bayle Y.**, Puginier E. et Semal C., 2017. Detecting temporal changes in acoustic scenes: The variable benefit of selective attention. *Hearing Research*, 353, pages 17–25. https://doi.org/10.1016/j.heares.2017.07.013

Bayle Y., Robine M. et Hanna P., 2018. SATIN: A persistent musical database for music information retrieval and a supporting deep learning experiment on song instrumental classification. *Springer Multimedia Tools and Applications*. Sous presse. https://link.springer.com/article/10.1007%2Fs11042-018-5797-8

Bayle Y., Hanna P. et Robine M., 2018. Toward faultless content-based playlists generation for instrumentals. Soumis à une revue. https://arxiv.org/abs/1706.07613

Bibliographie

Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., Kudlur M., Levenberg J., Monga R., Moore S., Murray D. G., Steiner B., Tucker P., Vasudevan V., Warden P., Wicke M., Yu Y. et Zheng X., 2016. Tensorflow: A system for large-scale machine learning. Dans *Proceedings of the 12th USENIX Symposium on Operating System Design Implementation*, pages 265–283.

Aizenberg N., Koren Y. et Somekh O., 2012. Build your own music recommender by modeling internet radio streams. Dans *Proceedings of the 21st International Conference on World Wide Web*, pages 1–10.

Aljanaki A., 2016. *Emotion in music: Representation and computational modeling*. Thèse de doctorat, Univ. Utrecht, Netherlands.

Andreou L.-V., Kashino M. et Chait M., 2011. The role of temporal regularity in auditory segregation. *Hearing research*, 280(1), pages 228–235.

Arumugam M. et Kaliappan M., 2016. An efficient approach for segmentation, feature extraction and classification of audio signals. *Circuits and Systems*, 7(4), pages 255–279.

Aucouturier J.-J. et Pachet F., 2003. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1), pages 83–93.

Balkwill L.-L. et Thompson W. F., 1999. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music perception: An interdisciplinary journal*, 17(1), pages 43–64.

Balkwill L.-L., Thompson W. F. et Matsunaga R., 2004. Recognition of emotion in japanese, western, and hindustani music by japanese listeners. *Japanese Psychological Research*, 46(4), pages 337–349.

Barone M., Dacosta K., Viglensoni G. et Woolhouse M., 2015. GRAIL: A music identity space collection and API. Dans *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 4–5. Málaga, Spain.

Barone M. D., Dacosta K., Viglensoni G. et Woolhouse M. H., 2016. GRAIL: A music metadata identity API. Dans *Proceedings of the 17th*

International Society for Music Information Retrieval Conference, pages 1–3. New York, NY, USA.

Batista G. E. A. P. A., Bazzan A. L. C. et Monard M. C., 2003. Balancing training data for automated annotation of keywords: A case study. Dans *Proceedings of the 2nd Brazilian Workshop on Bioinformatics*, pages 10–18. Rio de Janeiro, Brazil.

Batista G. E. A. P. A., Prati R. C. et Monard M. C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1), pages 20–29.

Bayle Y., Hanna P. et Robine M., 2016. Classification à grande échelle de morceaux de musique en fonction de la présence de chant. Dans *Actes des 23èmes Journées d'Informatique Musicale*, pages 144–152. Albi, France.

Bayle Y., Hanna P. et Robine M., 2017a. SATIN: A persistent musical database for music information retrieval. Dans *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, pages 1–5. Florence, Italy.

Bayle Y., Maršík L., Rusek M., Robine M., Hanna P., Slaninová K., Martinovič J. et Pokorný J., 2017b. Kara1k: A karaoke dataset for cover song identification and singing voice analysis. Dans *Proceedings of the 19th IEEE International Symposium on Multimedia Content-Based Multimedia Indexing*, pages 1–8. Taichung, Taiwan.

Bayle Y., Robine M. et Hanna P., 2018a. SATIN: A persistent musical database for music information retrieval and a supporting deep learning experiment on song instrumental classification. *Multimedia Tools and Applications*. Sous presse.

Bayle Y., Robine M. et Hanna P., 2018b. Toward faultless content-based playlists generation for instrumentals. Soumis à une revue.

Bekios-Calfa J., Buenaposada J. M. et Baumela L., 2011. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4), pages 858–864.

Bengio Y., Simard P. et Frasconi P., 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), pages 157–166.

Berenzweig A., Logan B., Ellis D. P. W. et Whitman B., 2004. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2), pages 63–76.

Bergstra J., Casagrande N., Erhan D., Eck D. et Kégl B., 2006a. Aggregate features and AdaBoost for music classification. *Springer Journal on Machine learning*, 65(2-3), pages 473–484.

Bergstra J., Casagrande N., Erhan D., Eck D. et Kégl B., 2006b. Meta-features and AdaBoost for music classification. *Machine Learning Journal: Special Issue on Machine Learning in Music*, pages 1–28.

Bertin-Mahieux T., Eck D. et Mandel M. I., 2010. Automatic tagging of audio: The state-of-the-art. Dans *Machine Audition: Principles Algorithms and Systems*, chapitre 14, pages 334–352. Information Science Reference, IGI Global.

Bertin-Mahieux T., Ellis D. P. W., Whitman B. et Lamere P., 2011. The million song dataset. Dans *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 591–596. Miami, FL, USA.

Bispham J., 2006. Rhythm in music: What is it ? Who has it ? And why ? *Music Perception: An Interdisciplinary Journal*, 24(2), pages 125–134.

Bittner R. M., Salamon J., Tierney M., Mauch M., Cannam C. et Bello J. P., 2014. MedleyDB: A multitrack dataset for annotation-intensive MIR research. Dans *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 155–160. Taipei, Taiwan.

Bogdanov D., Serrà J., Wack N., Herrera P. et Serra X., 2011. Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, 13(4), pages 687–701.

Bogdanov D., Wack N., Gómez E., Gulati S., Herrera P., Mayor O., Roma G., Salomon J., Zapata J. R. et Serra X., 2013. Essentia: An audio analysis library for music information retrieval. Dans *Proceedings of the 14th International Society for Music Information Retrieval Conference*, pages 493–498. Curitiba, Brazil.

Bonnet C., 1986. *Manuel pratique de psychophysique*. A. Colin.

Bonnin G. et Jannach D., 2014. Automated generation of music playlists: Survey and experiments. *ACM Computing Surveys*, 47(2), pages 1–35.

Branco P., Torgo L. et Ribeiro R. P., 2016. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), pages 1–50.

Breese J. S., Heckerman D. et Kadie C., 1998. Empirical analysis of predictive algorithms for collaborative filtering. Dans *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52. Madison, WI, USA.

Breiman L., 2001. Random forests. *Machine learning*, 45(1), pages 5–32.

Breiman L., Friedman J., Stone C. J. et Olshen R. A., 1984. *Classification and regression trees*. CRC press.

Bu J., Tan S., Chen C., Wang C., Wu H., Zhang L. et He X., 2010. Music recommendation by unified hypergraph: Combining social media information and music content. Dans *Proceedings of the 18th ACM International Conference on Multimedia*, pages 391–400. New York, NY, USA.

Carcagno S., Semal C. et Demany L., 2011. Frequency-shift detectors bind binaural as well as monaural frequency representations. *Journal of Experimental Psychology: Human Perception and Performance*, 37(6), pages 1976–1987.

Casey M. A., Veltkamp R., Goto M., Leman M., Rhodes C. et Slaney M., 2008. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4), pages 668–696.

Celma Ò., Ramírez M. et Herrera P., 2005. Foafing the music: A music recommendation system based on RSS feeds and user preferences. Dans *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 457–464. London, UK.

Chan T.-S., Yeh T.-C., Fan Z.-C., Chen H.-W., Su L., Yang Y.-H. et Jang R., 2015. Vocal activity informed singing voice separation with the iKala dataset. Dans *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 718–722.

Chau P. Y. K., Ho S. Y., Ho K. K. W. et Yao Y., 2013. Examining the effects of malfunctioning personalized services on online users' distrust and behaviors. *Decision Support Systems*, 56, pages 180–191.

Chawla N. V., Bowyer K. W., Hall L. O. et Kegelmeyer W. P., 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pages 321–357.

Chen S., Moore J. L., Turnbull D. et Joachims T., 2012. Playlist prediction via metric embedding. Dans *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*, pages 714–722.

Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. et Bengio Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Dans *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734. Doha, Qatar.

Choi K., Fazekas G., Sandler M. et Cho K., 2016a. Convolutional Recurrent Neural Networks for Music Classification. Rapport technique.

Choi K., Fazekas G., Cho K. et Sandler M. B., 2017a. A comparison on audio signal preprocessing methods for deep neural networks on music tagging. Rapport technique.

Choi K., Fazekas G., Cho K. et Sandler M. B., 2017b. A tutorial on deep learning for music information retrieval. Rapport technique.

Choi K., Fazekas G. et Sandler M. B., 2016b. Automatic tagging using deep convolutional neural networks. Dans *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 805–811. New York, NY, USA.

Choi K., Fazekas G., Sandler M. B. et Cho K., 2017c. Transfer learning for music classification and regression tasks. Dans *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pages 141–149. Suzhou, China.

Choi K., Kingdom T. U., Sandler M. et Kingdom T. U., 2015. Understanding Music Playlists. Rapport technique.

Chollet F., 2015. Keras: Deep learning library for theano and tensorflow. Rapport technique.

Chudáček V., Georgoulas G., Lhotská L., Stylios C., Petrík M. et Čepek M., 2009. Examining cross-database global training to evaluate five different methods for ventricular beat classification. *Physiological measurement*, 30(7), pages 661–677.

Coates A., Ng A. et Lee H., 2011. An analysis of single-layer networks in unsupervised feature learning. Dans *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 215–223. Fort Lauderdale, FL, USA.

Conard N. J., Malina M. et Münzel S. C., 2009. New flutes document the earliest musical tradition in southwestern germany. *Nature*, 460(7256), pages 737–740.

Constantino F. C., Pinggera L., Paranamana S., Kashino M. et Chait M., 2012. Detection of appearing and disappearing objects in complex acoustic scenes. *PLoS One*, 7(9), pages e46167.

Cortes C. et Vapnik V., 1995. Support-vector networks. *Machine learning*, 20(3), pages 273–297.

Cover T. et Hart P. E., 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), pages 21–27.

Craft A. J. D., Wiggins G. A. et Crawford T., 2007. How many beans make five ? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. Dans *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 73–76. Vienna, Austria.

De Cheveigné A., 1997. Harmonic fusion and pitch shifts of mistuned partials. *The Journal of the Acoustical Society of America*, 102(2), pages 1083–1087.

Dean J., Corrado G. S., Monga R., Chen K., Devin M., Le Q. V., Mao M. Z., Ranzato M., Senior A., Tucker P., Yang K. et Ng A. Y., 2012. Large scale distributed deep networks. Dans *Proceedings of the 26th Conference on the Advances in Neural Information Processing Systems*, pages 1223–1231. Lake Tahoe, NV, USA.

Defferrard M., Benzi K., Vandergheynst P. et Bresson X., 2017. FMA: A dataset for music analysis. Dans *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pages 316–323. Suzhou, China.

Demany L., Bayle Y., Puginier E. et Semal C., 2017. Detecting temporal changes in acoustic scenes: The variable benefit of selective attention. *Hearing Research*, 353, pages 17–25.

Demany L. et Ramos C., 2005. On the binding of successive sounds: Perceiving shifts in nonperceived pitches. *The Journal of the Acoustical Society of America*, 117(2), pages 833–841.

Demany L., Semal C. et Pressnitzer D., 2011. Implicit versus explicit frequency comparisons: Two mechanisms of auditory change detection. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2), pages 597–606.

Deng J., Dong W., Socher R., Li L.-J., Li K. et Fei-Fei L., 2009. Imagenet: A large-scale hierarchical image database. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Miami, FL, USA.

Dieleman S., Brakel P. et Schrauwen B., 2011. Audio-based music classification with a pretrained convolutional network. Dans *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 669–674. Miami, FL, USA.

Dieleman S. et Schrauwen B., 2014. End-to-end learning for music audio.

Dietterich T. G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7), pages 1895–1923.

Doshi-Velez F. et Kim B., 2017. Towards a rigorous science of interpretable machine learning. Rapport technique, Google Research.

Drummond C., 2009. Replicability is not reproducibility: nor is it good science. Dans *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML Conference*, pages 1–4. Montréal, Canada.

Eck D., Lamere P., Bertin-Mahieux T. et Green S., 2007. Automatic generation of social tags for music recommendation. Dans *Proceedings of the 21st Conference on Advances in Neural Information Processing Systems*, pages 385–392. Vancouver, BC, Canada.

Efron B., 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, pages 1–26.

Eghbal-Zadeh H., Lehner B., Dorfer M. et Widmer G., 2016. Cp-jku submissions for dcase-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks. *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*.

Eghbal-zadeh H. et Widmer G., 2016. Noise robust music artist recognition using I-Vector features. Dans *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 709–715. New York, NY, USA.

Ehmer R. H., 1959. Masking by tones vs noise bands. *The Journal of the Acoustical Society of America*, 31(9), pages 1253–1256.

Elkan C., 2001. The foundations of cost-sensitive learning. Dans *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978. Lawrence Erlbaum Associates Ltd, Seattle, WA, USA.

Ellis D. P. W. et Cotton C. V., 2007. The 2007 LabROSA cover song detection system. Dans *Music Information Retrieval Evaluation eXchange*.

Erhan D., Bengio Y., Courville A., Manzagol P.-A., Vincent P. et Bengio S., 2010. Why does unsupervised pre-training help deep learning ? *Journal of Machine Learning Research*, 11, pages 625–660.

Eronen A. et Klapuri A., 2000. Musical instrument recognition using cepstral coefficients and temporal features. Dans *Proceedings of the 24th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 753–756. Istanbul, Turkey.

Fan Z.-C., Chan T. S., Yang Y.-H. et Jang J.-S. R., 2017. Music signal processing using vector product neural networks. Dans *Proceedings of the 1st International Workshop on Deep Learning for Music*, pages 26–30. Anchorage, AK, USA.

Fernández C., Huerta I. et Prati A., 2015. A comparative evaluation of regression learning algorithms for facial age estimation. Dans *Face and Facial Expression Recognition from Real World Videos*, pages 133–144. Springer, Cham.

Fields B., 2011. *Contextualize your listening: The playlist as recommendation engine*. Thèse de doctorat, Goldsmiths College, Univ. of London.

Fischler M. A. et Bolles R. C., 1981. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Communications of the ACM*, 24(6), pages 381–395.

Flexer A., 2007. A closer look on artist filters for musical genre classification. Dans *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 341–344. Vienna, Austria.

Flexer A. et Schnitzer D., 2009. Album and artist effects for audio similarity at the scale of the web. Dans *Proceedings of 6th Sound and Music Computing*, pages 59–64. Porto, Portugal.

Flexer A. et Schnitzer D., 2010. Effects of album and artist filters in audio similarity computed for very large music databases. *Computer Music Journal*, 34(3), pages 20–28.

Foote J. T., 1997. Content-based retrieval of music and audio. Dans *Multimedia Storage and Archiving Systems II*, tome 3229, pages 138–148. International Society for Optics and Photonics.

Formby C., 1985. Differential sensitivity to tonal frequency and to the rate of amplitude modulation of broadband noise by normally hearing listeners. *The Journal of the Acoustical Society of America*, 78(1), pages 70–77.

Frank E., Hall M. A. et Witten I. H., 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann. Fourth Edition.

Freund Y. et Schapire R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), pages 119–139.

Fritz T., Jentschke S., Gosselin N., Sammler D., Peretz I., Turner R., Friederici A. D. et Koelsch S., 2009. Universal recognition of three basic emotions in music. *Current biology*, 19(7), pages 573–576.

Fu Z., Lu G., Ting K. M. et Zhang D., 2011. A Survey of Audio-Based Music Classification and Annotation. *IEEE Transactions on Multimedia*, 13(2), pages 303–319.

Furui S., 1986. Speaker-independent isolated word recognition based on emphasized spectral dynamics. Dans *Proceedings of the 11th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1991–1994. Tokyo, Japan.

Futrelle J. et Downie J. S., 2002. Interdisciplinary communities and research issues in music information retrieval. Dans *Proceedings of the 3rd International Society for Music Information Retrieval Conference*, pages 215–221. Paris, France.

Futrelle J. et Downie J. S., 2003. Interdisciplinary research issues in music information retrieval: Ismir 2000–2002. *Journal of New Music Research*, 32(2), pages 121–131.

Gabrielsson A. et Lindström E., 2001. *Series in affective science. Music and emotion: Theory and research.* Juslin, P. N. and Sloboda, J. A. at Oxford Univ. Press.

Gaikwad S. K., Gawali B. W. et Yannawar P., 2010. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3), pages 16–24.

Gemmeke J. F., Ellis D. P. W., Freedman D., Jansen A., Lawrence W., Moore C., Plakal M. et Ritter M., 2017. Audio Set: An ontology and human-labeled dataset for audio events. Dans *Proceedings of the 42nd IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1–5. New Orleans, LA, USA.

Ghosal A., Chakraborty R., Dhara B. C. et Saha S. K., 2013. A hierarchical approach for speech-instrumental-song classification. *SpringerPlus*, 2(526), pages 1–11.

Giraud A.-L., Lorenzi C., Ashburner J., Wable J., Johnsrude I., Frackowiak R. et Kleinschmidt A., 2000. Representation of the temporal

envelope of sounds in the human brain. *Journal of Neurophysiology*, 84(3), pages 1588–1598.

Glorot X., Bordes A. et Bengio Y., 2011. Deep sparse rectifier neural networks. Dans *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 315–323. Fort Lauderdale, FL, USA.

Goto M., Hashiguchi H., Nishimura T. et Oka R., 2002. RWC music database: Popular, classical and jazz music databases. Dans *Proceedings of the 3rd International Conference on Music Information Retrieval*, pages 287–288. Paris, France.

Gouyon F. et Dixon S., 2004. Dance music classification: A tempo-based approach. Dans *Proceedings of the 5th International Society for Music Information Retrieval Conference*, pages 501–504. Barcelona, Spain.

Gouyon F., Sturm B. L., Oliveira J. L., Hespanhol N. et Langlois T., 2014. On evaluation validity in music autotagging. Rapport technique.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. et Witten I. H., 2009. The WEKA data mining software. 11(1), pages 10–18.

Han H., Wang W.-Y. et Mao B.-H., 2005. Borderline-smote: A new over-sampling method in imbalanced data sets learning. Dans *Proceedings of the 1st International Conference on Intelligent Computing*, pages 878–887. Hefei, China.

Hart P. E., 1968. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3), pages 515–516.

Hastie T. et Tibshirani R., 1996. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6), pages 607–616.

He H., Bai Y., Garcia E. A. et Li S., 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. Dans *Proceedings of the 5th IEEE International Joint Conference on Neural Networks*, pages 1322–1328. Hong Kong, China.

Herrera P., Dehamel A. et Gouyon F., 2003. Automatic labeling of unpitched percussion sounds. Dans *Proceedings of the 114th Audio Engineering Society Convention*, pages 1–14. Amsterdam, The Netherlands.

Hespanhol N., 2013. *Using Autotagging for Classification of Vocals in Music Signals*. Thèse de doctorat, Univ. Porto, Portugal.

Hidasi B., Karatzoglou A., Baltrunas L. et Tikk D., 2016. Session-based recommendations with recurrent neural networks. Dans *Proceedings of the 4th International Conference on Learning Representations*, pages 1–10. San Juan, Puerto Rico.

Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I. et Salakhutdinov R. R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. Rapport technique.

Hoashi K., Matsumoto K. et Inoue N., 2003. Personalization of user profiles for content-based music retrieval based on relevance feedback. Dans *Proceedings of the 11th ACM International Conference on Multimedia*, pages 110–119. Berkeley, CA, USA.

Hochreiter S. et Schmidhuber J., 1997. Long short-term memory. *Neural Computation*, 9(8), pages 1735–1780.

Hsu C.-L. et Jang J.-S. R., 2010. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2), pages 310–319.

Hu X., Downie J. S., Laurier C., Bay M. et Ehmann A. F., 2008. The 2007 mirex audio mood classification task: Lessons learned. Dans *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 462–467. Philadelphia, PA, USA.

Hua K., 2018. Modeling singing F0 with neural network driven transition-sustain models. Rapport technique.

Huang C., Li Y., Change Loy C. et Tang X., 2016. Learning deep representation for imbalanced classification. Dans *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384. Las Vegas, NV, USA.

Huron D., 2000. Perceptual and cognitive applications in music information retrieval. Dans *Proceedings of the 1st International Symposium on Music Information Retrieval*, pages 1–2. Plymouth, USA.

Huron D., 2001. Is music an evolutionary adaptation ? *Annals of the New York Academy of sciences*, 930(1), pages 43–61.

Ikeda S., Oku K. et Kawagoe K., 2017. Music playlist recommendation using acoustic-feature transition inside the songs. Dans *Proceedings of the 15th International Conference on Advances in Mobile Computing and Multimedia*, pages 216–219. New York, NY, USA.

Inskip C., Macfarlane A. et Rafferty P., 2012. Towards the disintermediation of creative music search: Analysing queries to determine important facets. *International Journal on Digital Libraries*, 12(2), pages 137–147.

Ioffe S. et Szegedy C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Dans *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456. Lille, France.

Jansson A., Humphrey E. J., Montecchio N., Bittner R., Kumar A. et Weyde T., 2017. Singing voice separation with deep u-net convolutional networks. Dans *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pages 745–751. Suzhou, China.

Jäschke R., Marinho L., Hotho A., Schmidt-Thieme L. et Stumme G., 2007. Tag recommendations in folksonomies. Dans *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 506–514. Warsaw, Poland.

Jeon B., Kim C., Kim A., Kim D., Park J. et Ha J.-W., 2017. Music emotion recognition via end-to-end multimodal neural networks. Dans *Proceedings of the 11th ACM Conference on Recommender Systems*, pages 1–2. Como, Italy.

Jones M. R., Kidd G. et Wetzel R., 1981. Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), pages 1059–1073.

Jones M. R., Moynihan H., MacKenzie N. et Puente J., 2002. Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological science*, 13(4), pages 313–319.

Kereliuk C., Sturm B. L. et Larsen J., 2015. Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17(11), pages 2059–2071.

Keskar N. S., Mudigere D., Nocedal J., Smelyanskiy M. et Tang P. T. P., 2016. On large-batch training for deep learning: Generalization gap and sharp minima. Dans *Proceedings of the 5th International Conference on Learning Representations*, pages 1–16. Toulon, France.

Keuper J. et Preundt F.-J., 2016. Distributed training of deep neural networks: Theoretical and practical limits of parallel scalability. Dans *Proceedings of the 1st Workshop on Machine Learning in High Performance Computing Environments*, pages 19–26. Salt Lake City, UT, USA.

Kim S., Unal E. et Narayanan S., 2008. Music fingerprint extraction for classical music cover song identification. Dans *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1261–1264. Hannover, Germany.

Kleene S. C., 1951. Representation of events in nerve nets and finite automata. Rapport technique, Rand project air force, Santa Monica, CA, USA.

Knees P., Pohle T., Schedl M. et Widmer G., 2007. A music search engine built upon audio-based and web-based similarity measures. Dans *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 447–454. Amsterdam, The Netherlands.

Kotsiantis S., Kanellopoulos D. et Pintelas P., 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1), pages 25–36.

Kotsiantis S. B., Zaharakis I. et Pintelas P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, pages 3–24.

Krizhevsky A. et Hinton G., 2009. Learning multiple layers of features from tiny images. Rapport technique.

Krizhevsky A., Sutskever I. et Hinton G. E., 2012. Imagenet classification with deep convolutional neural networks. Dans *Proceedings of the 26th Annual Conference on Advances in Neural Information Processing Systems*, pages 1097–1105. Lake Tahoe, NV, USA.

Kubat M. et Matwin S., 1997. Addressing the curse of imbalanced training sets: One-sided selection. Dans *Proceedings of the 14th International Conference on Machine Learning*, pages 179–186. Nashville, Tennessee, USA.

Kum S. et Nam J., 2017. Classification-based singing melody extraction using deep convolutional neural networks. Rapport technique, Music and Audio Computing Lab, Korea Advanced Institute of Science and Technology, Republic of Korea.

Langlois T. et Marques G., 2009. A music classification method based on timbral features. Dans *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 81–86. Kobe, Japan.

Laurikkala J., 2001. Improving identification of difficult small classes by balancing class distribution. Dans *Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe*, pages 63–66. Springer, Cascais, Portugal.

Law E., West K., Mandel M. I., Bay M. et Downie J. S., 2009. Evaluation of algorithms using games: The case of music tagging. Dans *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 387–392. Kobe, Japan.

Law E. L. M., von Ahn L., Dannenberg R. B. et Crawford M., 2007. Tagatune : A game for music and sound annotation. Dans *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 361–364. Vienna, Austria.

LeCun Y., Bottou L., Bengio Y. et Haffner P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), pages 2278–2324.

LeCun Y., Jackel L. D., Bottou L., Brunot A., Cortes C., Denker J. S., Drucker H., Guyon I., Muller U. A., Sackinger E., Simard P. et Vapnik V., 1995. Comparison of learning algorithms for handwritten digit recognition. Dans *Proceedings of the 1st International Conference on Artificial Neural Networks*, pages 53–60. Paris, France.

Lee H., Pham P., Largman Y. et Ng A. Y., 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. Dans *Proceedings of the 23rd Conference on Advances in Neural Information Processing Systems*, pages 1096–1104. Vancouver, BC, Canada.

Leglaive S., Hennequin R. et Badeau R., 2015. Singing voice detection with deep recurrent neural networks. Dans *Proceedings of the 40th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 121–125. Brisbane, Australia.

Lehner B. et Widmer G., 2015. Monaural blind source separation in the context of vocal detection. Dans *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 309–315.

Lehner B., Widmer G. et Sonnleitner R., 2014. On the reduction of false positives in singing voice detection. Dans *Proceedings of the 39th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 7480–7484. Florence, Italy.

Lehner B., Widmer G. et Böck S., 2015. A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks. Dans *Proceedings ot the 23rd European Signal Processing Conference*, pages 21–25. Nice, France.

Lemaître G., Nogueira F. et Aridas C. K., 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), pages 1–5.

Lerch A., 2012. *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons.

Lesimple C., Sankey C., Richard M.-A. et Hausberger M., 2012. Do horses expect humans to solve their problems ? *Frontiers in psychology*, 3, pages 306–309.

Levitan C. A., Ban Y.-H. A., Stiles N. R. B. et Shimojo S., 2015. Rate perception adapts across the senses: Evidence for a unified timing mechanism. *Scientific reports*, 5, pages 8857–8862.

Levy M. et Sandler M. B., 2007. A semantic space for music derived from social tags. Dans *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 411–416. Vienna, Austria.

Li T. et Ogihara M., 2003. Detecting emotion in music. Dans *Proceedings of the 4th International Conference on Music Information Retrieval*, pages 1–2. Baltimore, Maryland, USA.

Liddell H. G. et Scott R., 1996. *A greek-english lexicon*. New York: American Book Company.

Lidy T. et Schindler A., 2016. CQT-based convolutional neural networks for audio scene classification and Domestic Audio Tagging. Dans *Proceedings of the IEEE Audio and Acoustic Signal Processing Challenge Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 60–64. Budapest, Hungary.

Lin Y.-C., Yang Y.-H. et Chen H. H., 2011. Exploiting online music tags for music emotion classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 7(1), pages 26–40.

Litjens G., Kooi T., Bejnordi B. E., Setio A. A. A., Ciompi F., Ghafoorian M., van der Laak J. A., van Ginneken B. et Sánchez C. I., 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42, pages 60–88.

Liutkus A., Fitzgerald D., Rafii Z., Pardo B. et Daudet L., 2014. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16), pages 4298–4310.

Livshin A. et Rodet X., 2003. The importance of cross database evaluation in sound classification. Dans *Proceedings of the 4th International Conference On Music Information Retrieval*, pages 1–2. Baltimore, MD, USA.

Livshin A. et Rodet X., 2009. Purging musical instrument sample databases using automatic musical instrument recognition methods. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5), pages 1046–1051.

Llamedo M., Khawaja A. et Martinez J. P., 2012. Cross-database evaluation of a multilead heartbeat classifier. *IEEE Transactions on Information Technology in Biomedecine*, 16(4), pages 658–664.

Logan B., 2002. Content-based playlist generation: Exploratory experiments. Dans *Proceedings of the 3rd International Conference on Music Information Retrieval*, pages 6–7. Paris, France.

Long J., Shelhamer E. et Darrell T., 2015. Fully convolutional networks for semantic segmentation. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440. Boston, MA, USA.

Longadge R. et Dongre S., 2013. Class imbalance problem in data mining review. *International Journal of Computer Science and Network*, 2, pages 1–6.

Lu L., Liu D. et Zhang H. J., 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio Speech and Language Processing*, 14(1), pages 5–18.

Lukashevich H., Gruhne M. et Dittmar C., 2007. Effective singing voice detection in popular music using arma filtering. Dans *Proceedings of the 10th International Workshop on Digital Audio Effects*, pages 165–168. Bordeaux, France.

Macmillan N. A. et Creelman C. D., 2004. *Detection theory: A user's guide*. Psychology press.

Mandel M. I. et Ellis D. P. W., 2008. Multiple-instance learning for music information retrieval. Dans *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 577–582. Philadelphia, PA, USA.

Mani I. et Zhang J., 2003. knn approach to unbalanced data distributions: A case study involving information extraction. Dans *Proceedings of the Workshop on Learning from Imbalanced Data Sets*, pages 1–7. Washington, DC, USA.

Markaki M., Holzapfel A. et Stylianou Y., 2008. Singing voice detection using modulation frequency features. Dans *Proceedings of the Workshop on Statistical and Perceptual Audition at the 9th Annual Conference of the International Speech Communication Association*, pages 7–10. Brisbane, Australia.

Marques G., Domingues M. A., Langlois T. et Gouyon F., 2011. Three current issues in music autotagging. Dans *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 795–800. Miami, FL, USA.

Martel Baro H., 2017. *A deep learning approach to source separation and remixing of hiphop music*. Thèse de maîtrise, Escola Superior Politècnica UPF, Spain.

Martin R., Mollineda R. A. et Garcia V., 2009. Melodic track identification in midi files considering the imbalanced context. Dans *Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis*, pages 489–496. Springer, Póvoa de Varzim, Portugal.

Maršík L., 2016. harmony-analyser.org - Java Library and Tools for Chordal Analysis. Dans *Proceedings of 2016 Joint WOCMAT-IRCAM Forum Conference*, pages 38–43. Taoyuan City, Taiwan.

Mathieu B., Essid S., Fillon T., Prado J. et Richard G., 2010. YAAFE, an easy to use and efficient audio feature extraction software. Dans *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 441–446. Utrecht, Netherlands.

Mauch M. et Dixon S., 2010. Approximate Note Transcription for the Improved Identification of Difficult Chords. Dans *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 135–140. Utrecht, Netherlands.

Mauch M., Cannam C., Bittner R., Fazekas G., Salamon J., Dai J., Bello J. et Dixon S., 2015. Computer-aided melody note transcription using the tony software: Accuracy and efficiency. Dans *Proceedings of the 1st International Conference on Technologies for Music Notation and Representation*, pages 1–8. Paris, France.

Mayer A. M., 1894. Researches in acoustics. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 37(226), pages 259–288.

McAfee A., Brynjolfsson E., Davenport T. H., Patil D. J. et Barton D., 2012. Big data: The management revolution. *Harvard business review*, 90(10), pages 60–68.

McCulloch W. S. et Pitts W. H., 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), pages 115–133.

McFee B., Humphrey E. J. et Bello J. P., 2015a. A software framework for musical data augmentation. Dans *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 248–254. Málaga, Spain.

McFee B., Raffel C., Liang D., Ellis D. P. W., McVicar M., Battenberg E. et Nieto O., 2015b. Librosa: Audio and music signal analysis in python. Dans *Proceedings of the 14th Python in Science Conference*, pages 18–25. Austin, TX, USA.

McVicar M., Santos-Rodriguez R. et De Bie T., 2016. Learning to separate vocals from polyphonic mixtures via ensemble methods and structured output prediction. Dans *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 450–454. Shanghai, China.

Medhat F., Chesmore D. et Robinson J., 2017. Music genre classification using masked conditional neural networks. Dans *Proceedings of the 1st International Conference on Neural Information Processing*, pages 470–481. Springer International Publishing, Guangzhou, China.

Mencattini A., Martinelli E., Costantini G., Todisco M., Basile B., Bozzali M. et Di Natale C., 2014. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems*, 63, pages 68–81.

Meyers O. C., 2007. *A mood-based music classification and exploration system.* Thèse de doctorat, Massachusetts Institute of Technology, USA.

Mimilakis S. I., Drossos K., Santos J. F., Schuller G., Virtanen T. et Bengio Y., 2017. Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask. Rapport technique.

Miron M., Janer Mestres J. et Gómez Gutiérrez E., 2017. Generating data to train convolutional neural networks for classical music source separation. Dans *Proceedings of the 14th Sound and Music Computing Conference*, pages 227–234. Espoo, Finland.

Moffat D., Ronan D. et Reiss J. D., 2015. An evaluation of audio feature extraction toolboxes. Dans *Proc of the 18th Int. Conference on Digital Audio Effects*, pages 1–7. Trondheim, Norway.

Moore B. C. J., 2012. *An introduction to the psychology of hearing.* Brill.

Moore B. C. J. et Shailer M. J., 1992. Modulation discrimination interference and auditory grouping. *Philosophical Transactions of the Royal Society of London for Biology*, 336(1278), pages 339–346.

Munkong R. et Juang B.-H., 2008. Auditory perception and cognition. *IEEE Signal Processing Magazine*, 25(3), pages 98–117.

Murauer B. et Specht G., 2018. Detecting music genre using extreme gradient boosting. Dans *Proceedings of the Companion of the The Web Conference 2018*, pages 1923–1927. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.

Nanayakkara C. V. et Caldera H. A., 2016. Music emotion recognition with audio and lyrics features. *International Journal of Digital Information and Wireless Communications*, 6(4), pages 260–273.

Ng A. Y., 1997. Preventing "overfitting" of cross-validation data. Dans *Proceedings of the 14th International Conference on Machine Learning*, pages 245–253. Nashville, TN, USA.

Nguyen H. M., Cooper E. W. et Kamei K., 2009. Borderline over-sampling for imbalanced data classification. Dans *Proceedings of the 5th International Workshop on Computational Intelligence and Applications*, pages 24–29. Hiroshima, Japan.

Nilsson N. J., 1965. *Learning machines: Foundations of trainable pattern-classifying systems.* McGraw-Hill.

Nwe T. L., Shenoy A. et Wang Y., 2004. Singing voice detection in popular music. Dans *Proceedings of the 12th annual ACM International Conference on Multimedia*, pages 324–327. New York, NY, USA.

Oramas S., Nieto O., Sordo M. et Serra X., 2017. A deep multimodal approach for cold-start music recommendation. Dans *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, pages 32–37. Como, Italy.

Orio N., 2006. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1), pages 1–90.

Osmalskyj J., Piérard S., Van Droogenbroeck M. et Embrechts J.-J., 2013. Efficient database pruning for large-scale cover song recognition. Dans *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, pages 714–718. Vancouver, BC, Canada.

Oyamada K., Kameoka H., Kaneko T., Tanaka K., Hojo N. et Ando H., 2018. Generative adversarial network-based approach to signal reconstruction from magnitude spectrograms. Rapport technique.

Pachet F. et Roy P., 1999. Automatic generation of music programs. Dans *Proceedings of the 5th International Conference on Constraint Programming*, pages 331–345. Alexandria, VA, USA.

Paiva R. P., 2013. Moodetector: Automatic music emotion recognition. Rapport technique.

Pampalk E., Flexer A. et Widmer G., 2005. Improvements of audio-based music similarity and genre classificaton. Dans *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 634–637. London, UK.

Panda R. et Paiva R. P., 2011. Using support vector machines for automatic mood tracking in audio music. Dans *Proceedings of the 130th Audio Engineering Society Convention*, pages 1–8. Audio Engineering Society, London, UK.

Payton K. L. et Braida L. D., 1999. A method to determine the speech transmission index from speech waveforms. *The Journal of the Acoustical Society of America*, 106(6), pages 3637–3648.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. et Duchesnay É., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pages 2825–2830.

Peng B., Guan K., Chen M., Lawrence D. M., Pokhrel Y., Suyker A., Arkebauer T. et Lu Y., 2018. Improving maize growth processes in the community land model: Implementation and evaluation. *Agricultural and Forest Meteorology*, 250-251, pages 64–89.

Pfungst O., 1911. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston.

Pikrakis A., 2013. Audio latin music genre classification: A MIREX 2013 submission based on a deep learning approach to rhythm modelling. Dans *Proceedings of the 9th Music Information Retrieval Evaluation eXchange*, pages 1–2. Curitiba, Brazil.

Pohle T., Knees P., Schedl M., Pampalk E. et Widmer G., 2007. "reinventing the wheel": A novel approach to music player interfaces. *IEEE Transactions on Multimedia*, 9(3), pages 567–575.

Pohle T., Pampalk E. et Widmer G., 2005. Evaluation of frequently used audio features for classification of music into perceptual categories. Dans *Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing*, pages 1–8. Riga, Lituana.

Pons J., Lidy T. et Serra X., 2016. Experimenting with musically motivated convolutional neural networks. Dans *Proceedings of the 14th International Workshop on Content-Based Multimedia Indexing*, pages 1–6. Bucharest, Romania.

Pons J., Nieto O., Prockup M., Schmidt E. M., Ehmann A. F. et Serra X., 2017a. End-to-end learning for music audio tagging at scale. Dans *Proceedings of the 31st Conference on the Advances in Neural Information Processing Systems*, pages 1–5. Long Beach, CA, USA.

Pons J., Slizovskaia O., Gong R., Gómez E. et Serra X., 2017b. Timbre Analysis of Music Audio Signals with Convolutional Neural Networks. Rapport technique.

Porter A., Bogdanov D., Kaye R., Tsukanov R. et Serra X., 2015. Acousticbrainz: A community platform for gathering music information obtained from audio. Dans *Proceedings of the 16th International Conference on Music Information Retrieval*, pages 786–792. Málaga, Spain.

Prockup M., Ehmann A. F., Gouyon F., Schmidt E. M., Celma O. et Kim Y. E., 2015. Modeling genre with the Music Genome Project: Comparing human-labeled attributes and audio features. Dans *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 31–37. Málaga, Spain.

Průša Z. et Rajmic P., 2017. Toward high-quality real-time signal reconstruction from stft magnitude. *IEEE Signal Processing Letters*, 24(6), pages 892–896.

Quinlan J. R., 1993. C4.5: Programming for machine learning. *Morgan Kauffmann*, 38, pages 48.

Rabiner L. R. et Juang B.-H., 1993. *Fundamentals of speech recognition*. PTR Prentice Hall.

Raina R., Madhavan A. et Ng A. Y., 2009. Large-scale deep unsupervised learning using graphics processors. Dans *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 873–880. Montreal, Canada.

Ramona M., Richard G. et David B., 2008. Vocal detection in music with support vector machines. Dans *Proceedings of the 32nd IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1885–1888. Las Vegas, NV, USA.

Recht B., Re C., Wright S. et Niu F., 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. Dans *Proceedings of the 25th Conference on the Advances in Neural Information Processing Systems*, pages 693–701. Granada, Spain.

Regnier L. et Peeters G., 2009. Singing voice detection in music tracks using direct voice vibrato detection. Dans *Proceedings of the 33rd IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1685–1688. Taipei, Taiwan.

Ribeiro M. T., Singh S. et Guestrin C., 2016. Model-agnostic interpretability of machine learning. Dans *Proceedings of the Workshop on Human Interpretability in Machine Learning at the 33rd International Conference on Machine Learning*, pages 91–95. New-York, NY, USA.

Riedmiller M., 1994. Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces*, 16(3), pages 265–278.

Rocamora M. et Herrera P., 2007. Comparing audio descriptors for singing voice detection in music audio files. Dans *Proceedings of the 11th Brazilian Symposium on Computer Music*, pages 187–196. San Pablo, Brazil.

Roma G., Grais E. M., Simpson A. J. et Plumbley M. D., 2016. Singing voice separation using deep neural networks and F0 estimation. Dans *Proceedings of the 9th Music Information Retrieval Evaluation eXchange*. New York, NY, USA.

Rosenblatt F., 1957. *The perceptron, a perceiving and recognizing automaton Project Para.* Cornell Aeronautical Laboratory.

Sadjadi S. O., Ahadi S. M. et Hazrati O., 2007. Unsupervised speech/music classification using one-class support vector machines. Dans *Proceedings of the 6th International Conference on Information, Communications and Signal Processing*, pages 1–5. Singapore, Singapore.

Sanden C. et Zhang J. Z., 2011. Enhancing multi-label music genre classification through ensemble techniques. Dans *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 705–714. Beijing, China.

Schlüter J., 2016. Learning to pinpoint singing voice from weakly labeled examples. Dans *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 44–50. New York, NY, USA.

Schlüter J. et Grill T., 2015. Exploring data augmentation for improved singing voice detection with neural networks. Dans *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 121–126. Málaga, Spain.

Schlüter J. et Böck S., 2014. Improved musical onset detection with convolutional neural networks. Dans *Proceedings of the 39rd IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6979–6983. Florence, Italy.

Schmidhuber J., 2015. Deep learning in neural networks: An overview. *Neural networks*, 61, pages 85–117.

Schreiner C. et Langner G., 1988. Periodicity coding in the inferior colliculus of the cat. ii. topographical organization. *Journal of Neurophysiology*, 60(6), pages 1823–1840.

Schröger E., Bendixen A., Denham S. L., Mill R. W., Bőhm T. M. et Winkler I., 2014. Predictive regularity representations in violation detection and auditory stream segregation: From conceptual to computational models. *Brain topography*, 27(4), pages 565–577.

Schulman J., Heess N., Weber T. et Abbeel P., 2015. Gradient estimation using stochastic computation graphs. Dans *Proceedings of the 29th Conference on the Advances in Neural Information Processing Systems*, pages 3528–3536. Montréal, Canada.

Senac C., Pellegrini T., Mouret F. et Pinquier J., 2017. Music feature maps with convolutional neural networks for music genre classification. Dans *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, pages 19–23. Florence, Italy.

Serrà J., Serra X. et Andrzejak R. G., 2009. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9), pages 093017.

Shardanand U. et Maes P., 1995. Social information filtering: Algorithms for automating "word of mouth". Dans *Proceedings of the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems*, pages 210–217. Denver, CO, USA.

Shepitsen A., Gemmell J., Mobasher B. et Burke R., 2008. Personalized recommendation in social tagging systems using hierarchical clustering. Dans *Proceedings of the Association for Computing Machinery 2nd Conference on Recommender Systems*, pages 259–266. Lausanne, Switzerland.

Shwartz-Ziv R. et Tishby N., 2017. Opening the black box of deep neural networks via information. Rapport technique.

Skowronek J., McKinney M. F. et van de Par S., 2006. Ground truth for automatic music mood classification. Dans *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 395–396. Victoria, BC, Canada.

Smith J. C., 2013. *Correlation analyses of encoded music performance*. Thèse de doctorat, Stanford Univ., Stanford, CA, USA.

Smith M. R., Martinez T. et Giraud-Carrier C., 2014. An instance level analysis of data complexity. *Machine learning*, 95(2), pages 225–256.

Sordo M., Laurier C. et Celma Ò., 2007. Annotating music collections: How content-based similarity helps to propagate labels. Dans *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 531–534. Vienna, Austria.

Sridharan S., Vaidyanathan K., Kalamkar D., Das D., Smorkalov M. E., Shiryaev M., Mudigere D., Mellempudi N., Avancha S., Kaul B. et Dubey P., 2018. On scale-out deep learning training for cloud and HPC. Rapport technique.

Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I. et Salakhutdinov R., 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), pages 1929–1958.

Stevens S. S., Volkmann J. et Newman E. B., 1937. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3), pages 185–190.

Stolcke A., Kajarekar S. et Ferrer L., 2008. Nonparametric feature normalization for svm-based speaker verification. Dans *Proceedings of the 33rd IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1577–1580. Las Vegas, Nevada, USA.

Stoller D., Ewert S. et Dixon S., 2018. Adversarial semi-supervised audio source separation applied to singing voice extraction. Dans *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. Calgary, Canada.

Sturm B. L., 2013. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. Rapport technique.

Sturm B. L., 2014a. A simple method to determine if a music information retrieval system is a "Horse". *IEEE Transactions on Multimedia*, 16(6), pages 1636–1644.

Sturm B. L., 2014b. The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, 43(2), pages 147–172.

Sturm B. L., 2015. Faults in the latin music database and with its use. Dans *Proceedings of the Late Breaking Demo of the 16th International Society for Music Information Retrieval Conference*, pages 1–2. Málaga, Spain.

Sturm B. L., Kereliuk C. et Pikrakis A., 2014. A closer look at deep learning neural networks with low-level spectral periodicity features. Dans *Proceedings of the 4th International Workshop on Cognitive Information Processing*, pages 1–6. Copenhagen, Denmark.

Sturmel N. et Daudet L., 2011. Signal reconstruction from stft magnitude: A state of the art. Dans *Proceedings of the 14th International Conference on Digital Audio Effects*, pages 375–386. Paris, France.

Sutskever I., Vinyals O. et Le Q. V., 2014. Sequence to sequence learning with neural networks. Dans *Proceedings of the 28th Conference on the Advances in Neural Information Processing Systems*, pages 3104–3112. Montréal, Canada.

Tachibana H., Ono T., Ono N. et Sagayama S., 2010. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. Dans *Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 425–428. Dallas, TX, USA.

Tajadura-Jiménez A., Pantelidou G., Rebacz P., Västfjäll D. et Tsakiris M., 2011. I-space: The effects of emotional valence and source of music on interpersonal distance. *PLoS One*, 6(10), pages e26083.

Takahashi N., Gygli M., Pfister B. et Van Gool L., 2016. Deep convolutional neural networks and data augmentation for acoustic event detection. Dans *Proceedings of the Interspeech conference*, pages 2982–2986. San Francisco, USA.

Terhardt E., Stoll G. et Seewann M., 1982. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, 71(3), pages 679–688.

Thickstun J., Harchaoui Z., Foster D. et Kakade S. M., 2017. Invariances and data augmentation for supervised music transcription. Rapport technique.

Tingle D., Kim Y. E. et Turnbull D., 2010. Exploring automatic music annotation with "acoustically-objective" tags. Dans *Proceedings of the 11th ACM International Conference on Multimedia Information Retrieval*, pages 55–62. Philadelphia, PA, USA.

Tintarev N., Lofi C. et Liem C., 2017. Sequences of diverse song recommendations: An exploratory study in a commercial system. Dans *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 391–392. Bratislava, Slovakia.

Tomek I., 1976a. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 1(6), pages 448–452.

Tomek I., 1976b. Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 6, pages 769–772.

Turnbull D., Barrington L., Torres D. et Lanckriet G., 2007. Towards musical query-by-semantic-description using the CAL500 data set. Dans *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 439–446. Amsterdam, The Netherlands.

Turnbull D., Barrington L., Torres D. et Lanckriet G., 2008. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio Speech and Language Processing*, 16(2), pages 467–476.

Tzanetakis G. et Cook P., 2000. Marsyas: A framework for audio analysis. *Organised sound*, 4(3), pages 169–175.

Tzanetakis G. et Cook P., 2002. Musical genre classification of audio signals. *IEEE Transaction on Speech and Audio Processing*, 10(5), pages 293–302.

Urbano J., Bogdanov D., Herrera P., Gómez E. et Serra X., 2014. What is the effect of audio quality on the robustness of MFCCs and chroma features ? Dans *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 573–578. Taipei, Taiwan.

Valin J.-M., 2017. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. Rapport technique.

Van Noort J., 1969. *The structure and connections of the inferior colliculus: An investigation of the lower auditory system.* Van Gorcum.

Velarde G., 2017. *Convolutional methods for music analysis.* Thèse de doctorat, Aalborg Universitetsforlag.

Von Schroeter T., Doraisamy S. et Rüger S. M., 2000. From raw polyphonic audio to locating recurring themes. Dans *Proceedings of the 1st International Symposium on Music Information Retrieval*, pages 1–11. Plymouth, USA.

Wan X., Liu J., Cheung W. K. et Tong T., 2014. Learning to improve medical decision making from imbalanced data without a priori cost. *BMC Medical Informatics and Decision Making*, 14(1), pages 111.

Weninger F., Durrieu J.-L., Eyben F., Richard G. et Schuller B., 2011. Combining monaural source separation with long short-term memory for increased robustness in vocalist gender recognition. Dans *Proceedings of the 36th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 2196–2199. Prague, Czech Republic.

West K. et Cox S., 2004. Features and classifiers for the automatic classification of musical audio signals. Dans *Proceedings of the 5th International Society for Music Information Retrieval Conference*, pages 1–6. Barcelona, Spain.

Whitman B. et Ellis D. P. W., 2004. Automatic record reviews. Dans *Proceedings of the 5th International Conference on Music Information Retrieval*, pages 470–477. Barcelona, Spain.

Wiggins G. A., 2009. Semantic gap ? ? Schemantic schmap ! ! Methodological considerations in the scientific study of music. Dans *Proceedings of the 11th IEEE International Symposium on Multimedia*, pages 477–482. San Diego, CA, USA.

Wilson D. L., 1972. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 2(3), pages 408–421.

Woods K. et Bowyer K. W., 1997. Generating roc curves for artificial neural networks. *IEEE Transactions on Medical Images*, 16(3), pages 329–337.

Wu D., Sharma N. et Blumenstein M., 2017. Recent advances in video-based human action recognition using deep learning: A review. Dans *Proceedings of the 30th International Joint Conference on Neural Networks*, pages 2865–2872. Anchorage, AK, USA.

Wu X., Kumar V., Ross Quinlan J., Ghosh J., Yang Q., Motoda H., McLachlan G. J., Ng A., Liu B., Yu P. S., Zhou Z.-H., Steinbach M.,

Hand D. J. et Steinberg D., 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), pages 1–37.

Wu Y., Schuster M., Chen Z., Le Q. V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q., Macherey K., Jeff Klingner, Shah A., Johnson M., Liu X., Kaiser L., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M. et Dean J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. Rapport technique.

Yang Q. et Wu X., 2006. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(4), pages 597–604.

Ycart A. et Benetos E., 2017. A study on LSTM networks for polyphonic music sequence modelling. Dans *Proceedings of the 18th International Conference on Music Information Retrieval*, pages 421–427. Suzhou, China.

Yin D., Bond S. D. et Zhang H., 2010. Are bad reviews always stronger than good ? Asymmetric negativity bias in the formation of online consumer trust. Dans *Proceedings of 31st International Conference of Information Systems*, pages 1–18. Saint-Louis, MO, USA.

Zhang P., Zheng X., Zhang W., Li S., Qian S., He W., Zhang S. et Wang Z., 2015. A deep neural network for modeling music. Dans *Proceedings of the 5th Annual ACM International Conference on Multimedia Retrieval*. Shanghai, China.

Zhang T. et Kuo C.-C. J., 2013. *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*. Springer Science & Business Media.

Zhang Y. C., Séaghdha D. Ó., Quercia D. et Jambor T., 2012. Auralist: Introducing serendipity into music recommendation. Dans *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 13–22. Seattle, WA, USA.

Zhao P., Jin R., Yang T. et Hoi S. C., 2011. Online AUC maximization. Dans *Proceedings of the 28th International Conference on Machine Learning*, pages 233–240. Bellevue, WA, USA.

Glossaire

Cent Échelle de représentation des notes musicales qui découpe un demi-ton en cent unités. 66, 68, 71, 77, 81

Chanson Morceau comprenant au moins une piste audio sur laquelle ont été enregistrés des sons provenant directement ou indirectement de la voix humaine. 21, 25, 26, 29–34, 36, 37, 39, 41–43, 47, 48, 83, 86, 87, 89–93, 106, 108–114, 128, 129, 136, 139, 140, 181–183, 185

Colliculus inférieur Élément du cerveau qui fait partie de la voie auditive ascendante entre l'oreille et le cortex auditif [Van Noort, 1969]. 61, 82

Échantillon Représentation numérique de la valeur de la pression acoustique à un instant donné au niveau du micro d'enregistrement. 24, 26, 46, 47, 84, 123, 175, 176

Émotion Réaction affective transitoire d'assez grande intensité, habituellement provoquée par une stimulation venue de l'environnement [1]. 5, 7, 25, 26, 136, 140, 176

Enregistrement Morceau enregistré sur support matérialisé ou dématérialisé fixant une interprétation et composé d'un ensemble d'échantillons. 22–24, 31, 38, 119, 175, 176

Harmonique Son dont la fréquence est un multiple entier d'une fréquence fondamentale propre à un autre son. 62

Heuristique Se dit d'une méthode de calcul qui fournit relativement rapidement, en temps polynomial, une solution réalisable mais pas nécessairement optimale pour un problème d'optimisation NP-difficile. 50, 51

Horse Se dit d'une méthode qui ne traite pas le problème qu'elle prétend résoudre [Sturm, 2014a]. 9–11, 23, 45, 47, 48, 99, 100, 106, 107, 124, 125, 135, 137, 138

Instrumental Morceau ne comprenant aucune piste audio sur laquelle ont été enregistrés des sons provenant directement ou indirectement de la voix humaine. 21, 25, 29–34, 47, 48, 83, 86–93, 106–115, 128, 129, 136, 139, 140, 181, 182, 185

1. http://larousse.fr/dictionnaires/francais/emotion, consulté le 19 Avril 2018.

Morceau Œuvre musicale interprétée. 1, 4–15, 18–49, 52, 54, 55, 57, 58, 61, 64, 83–95, 97, 100, 101, 104–115, 118, 121, 123, 125–129, 131–133, 135–141, 143, 175, 176, 181–186

Mégadonnées Les mégadonnées désignent des ensembles de données croissants devenus si volumineux et complexes qu'ils dépassent l'intuition et les capacités humaines d'analyse ainsi que celles des outils informatiques classiques de gestion de bases de données ou de l'information. L'utilisation de ce mot est recommandée par la délégation générale à la langue française et aux langues de France [1]. L'équivalent anglais des mégadonnées est *big data*. 10

Piste Un enregistrement comprend une ou plusieurs pistes qui correspondent à chaque source unique de sons. 30, 38, 40, 41, 46, 85, 140, 141, 175, 176

Reprise Se dit d'un morceau existant qui est rejoué, de façon similaire ou non, par un interprète différent de celui de la version originale. 21, 38–41, 44, 184

Reproduire Capacité déterministe à générer à nouveau les résultats d'une expérience scientifique, à ne pas confondre avec la réplicabilité [Drummond, 2009]. 176

Réplicabilité Fait de parvenir aux mêmes conclusions qu'une expérience scientifique, à ne pas confondre avec reproduire [Drummond, 2009]. 22, 32, 36, 54, 176

Sentiment État affectif complexe et durable lié à certaines émotions ou représentations [2]. 5, 25, 26

Streaming (anglicisme) Diffusion ou transfert de données en flux continu. 4–8, 20, 34, 93

Sérendipité Capacité, art de faire une découverte, notamment scientifique, par hasard. La sérendipité dénote également la découverte ainsi faite. 8, 9

Tessiture Ensemble de notes pouvant être émises avec homogénéité par un instrument. 23

Timbre Ensemble des caractéristiques sonores qui permettent d'identifier un instrument. 23

Trame Subdivision arbitraire d'un enregistrement regroupant plusieurs échantillons audio. 26, 27, 30, 84–93, 102, 113–116, 140, 185

1. http://www.culturecommunication.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/La-DGLFLF, consulté le 19 Avril 2018.
2. http://larousse.fr/dictionnaires/francais/sentiment, consulté le 19 Avril 2018.

Acronymes

ADASYN ADAptive SYNthetic sampling approach for imbalanced learning. 51

AllKNN All K-Nearest Neighbours. 49, 53

API Application Programming Interface. 32, 33, 35, 36, 136

AUC Area Under the Curve. 16, 17, 112, 183, 185

BPM Battements Par Minute. 63

bSMOTE borderline Synthetic Minority Over-sampling TEchnique. 50, 53

CIFAR Canadian Institute For Advanced Research. 19, 20

CNN Convolutional Neural Networks. 119, 120, 123, 128, 129

Condensed-NN Condensed Nearest Neighbours. 49, 50, 53

CSV Comma Separated Value. 33

DAMP Digital Archive of Mobile Performances. 39

EFF Electronic Frontier Foundation. 33

ENN Edited Nearest Neighbours. 49–51, 53

FMA Free Music Archive. 24, 52, 126, 127, 131

GA Algorithme de Ghosal *et al.* [2013]. 108, 110–113, 115, 185

GRU Gated Recurrent Unit. 120

IEEE Institute of Electrical and Electronics Engineers. 136

IFPI International Federation of the Phonographic Industry. 4, 29, 31, 33, 35, 106, 183

IHT Instance Hardness Threshold. 50, 53

INRIA Institut National de Recherche en Informatique et en Automatique. 30

IRISA Institut de Recherche en Informatique et Systèmes Aléatoires. 30

ISO International Organization for Standardization. 32

ISRC International Standard Recording Code. 31–33, 35–37, 41, 183

k-NN K-Nearest Neighbours. 103

Kara1K Karaoke database of 1,000 tracks. 40–44, 136, 183

LSTM Long Short-Term Memory. 120

MBID MusicBrainz IDentifier. 32, 33

MDI Modulation Discrimination Interference. 62, 76

MFCC Mel-Frequency Cepstral Coefficients. 85, 86, 93, 108, 109, 128

MIREX Music Information Retrieval Evaluation eXchange. 38, 45

MLP Multi Layer Perceptron. 118, 119, 185

MNIST Modified National Institute of Standards and Technology. 19, 20

MP3 Moving Picture Experts Group Phase 1 Audio Layer III. 58

MSD Million Song Dataset. 30

NCR Neighboorhood Cleaning Rule. 50

NearMiss NearMiss using nearest neighbours. 50, 53

OSS One-Sided Selection. 50, 53

PEA ProbaH2-10, ETRAll et AdaBoost. 109–115, 136, 185

QMUL Queen Mary University of London. 10

RAM Random-Access Memory. 125–127, 129

RANSAC RANdom Sample And Consensus. 103, 108, 115

ReLU Rectified Linear Unit. 120, 128

RENN Repeated Edited Nearest Neighbours. 49, 53

RNN Recurrent Neural Networks. 120

ROS Random minority Over-Sampling. 51, 53

RUS Random majority Under-Sampling. 50, 53

SATIN Set of Audio Tags and Identifiers Normalized. 31, 33–37, 41, 43, 44, 109, 111, 112, 128–130, 136, 139, 182, 183, 185

SHS Second Hand Songs. 38

SMOTE Synthetic Minority Over-sampling TEchnique. 50, 51, 53

SMOTEENN Synthetic Minority Over-sampling TEchnique followed by Edited Nearest Neighbours. 51, 53

SMOTETOMEK Synthetic Minority Over-sampling TEchnique followed by Tomek links. 51, 53

SOFT1 first Set Of FeaTures. 33, 34, 128

SPL Sound Pressure Level. 60, 70

SVM Support Vector Machine. 51, 102, 108, 115

SVMBFF Support Vector Machine applied to Bags of Frames of Features. 108, 110–113, 115, 185

SVMSMOTE Support Vector Machine and Synthetic Minority Over-sampling TEchnique. 51, 53

Tomek Extraction of majority-minority Tomek links. 50, 51, 53

VQMM Vector Quantization and Markov Model. 108, 110–113, 115, 185

Liste des tableaux

Liste des figures

"There's no sense in going further –
it's the edge of cultivation,"
So they said, and I believed it –
broke my land and sowed my crop –
Built my barns and strung my fences
in the little border station
Tucked away below the foothills
where the trails run out and stop.

Till a voice, as bad as Conscience,
rang interminable changes
In one everlasting Whisper
day and night repeated – so:
"Something hidden. Go and find it.
Go and look behind the Ranges –
Something lost behind the Ranges.
Lost and waiting for you. Go!"

– *The Explorer*, 1898, Rudyard Kipling

Français - Ce mémoire de thèse de doctorat présente, discute et propose des outils de fouille automatique de mégadonnées dans un contexte de classification supervisée musicale. L'application principale concerne la classification automatique des thèmes musicaux afin de générer des listes de lecture thématiques homogènes.

Le premier chapitre introduit les différents contextes et concepts autour des mégadonnées musicales et de leur consommation. Le deuxième chapitre s'attelle à la description des problématiques concernant la variété et les proportions inégales des thèmes contenus dans les bases de données musicales existantes dans le cadre d'expériences académiques d'analyse audio. Le troisième chapitre détaille le développement et l'extraction de caractéristiques audio afin de décrire le contenu des morceaux. Ce chapitre explique plusieurs phénomènes psychoacoustiques et utilise des techniques de traitement du signal afin de calculer des caractéristiques musicales. De nouvelles méthodes d'agrégation de caractéristiques audio locales sont proposées afin d'améliorer la classification des morceaux. Le quatrième chapitre décrit l'utilisation des caractéristiques musicales extraites afin de trier les morceaux par thèmes et donc de permettre les recommandations musicales et la génération automatique de listes de lecture thématiques homogènes. Cette partie implique l'utilisation d'algorithmes d'apprentissage automatique afin de réaliser des tâches de classification musicale. Les contributions de ce mémoire sont résumées dans le cinquième chapitre qui propose également des perspectives de recherche dans l'apprentissage automatique et l'extraction de caractéristiques audio.

English - This doctoral dissertation presents, discusses and proposes tools for the automatic information retrieval in big musical databases. The main application is the supervised classification of musical themes to generate thematic playlists.

The first chapter introduces the different contexts and concepts around big musical databases and their consumption. The second chapter focuses on issues concerning the variety and unequal proportions of themes contained in existing music databases as part of academic experiments in audio analysis. The third chapter explains the importance of extracting and developing relevant audio features in order to better describe the content of music tracks in these databases. This chapter explains several psychoacoustic phenomena and uses sound signal processing techniques to compute audio features. New methods for aggregating local audio features are proposed to improve song classification. The fourth chapter describes the use of the extracted audio features in order to sort the songs by themes and thus to allow the musical recommendations and the automatic generation of homogeneous thematic playlists. This part involves the use of machine learning algorithms to perform music classification tasks. The contributions of this dissertation are summarized in the fifth chapter which also proposes research perspectives in machine learning and audio features extraction.