Jusqu'où les goûts musicaux sont-ils prédictibles par l'intelligence artificielle ?

Nicolas Dauban

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Nicolas DAUBAN

Le 6 avril 2021

Jusqu'où les goûts musicaux sont-ils prédictibles par l'intelligence artificielle ?

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Julien PINQUIER et Pascal GAILLARD

Jury

M. Hervé GLOTIN, Rapporteur
M. Geoffroy PEETERS, Rapporteur
Mme Myriam DE SAINTE-CATHERINE, Examinatrice
M. Julien PINQUIER, Directeur de thèse
M. Pascal GAILLARD, Co-directeur de thèse
Mme Christine SENAC, Co-encadrante de thèse
M. Gaël RICHARD, Président

Résumé

Cette thèse a pour objet l'utilisation de l'intelligence artificielle (IA) pour la recommandation de musique en tant que prédiction de goûts. Notre recherche s'est focalisée sur l'étude des goûts musicaux, du fonctionnement de la recommandation de musique et de la place de l'humain dans ce processus. Nous nous sommes intéressés à la pertinence des données utilisées pour la recommandation de musique : les données comportementales d'une part, souvent supposées traduire les affinités des utilisateurs ; les descripteurs des morceaux, utilisés pour la recommandation basée sur le contenu d'autre part. La démarche proposée s'est notamment appuyée sur des travaux en sciences humaines et en musicologie.

Les problématiques sont focalisées sur la prédiction des goûts musicaux, dans un domaine où les incertitudes autour de l'explicabilité de ces goûts sont nombreuses :

— Quels éléments influencent les goûts musicaux ? Quels sont les liens entre les genres musicaux et les goûts des auditeurs en matière de musique ?
— La recommandation de musique est-elle une prédiction de goûts ? Comment les données comportementales sont-elles interprétées par les plateformes de streaming ? Le comportement traduit-il vraiment des goûts musicaux ?
— Les paramètres acoustiques appris par un réseau de neurones convolutionnels pour la reconnaissance automatique de genre sont-ils pertinents pour la prédiction de goûts musicaux ? Quels facteurs déterminants dans l'affinité musicale un réseau de neurones convolutionnels est-il capable d'identifier dans un spectrogramme ?

Le premier chapitre présente notre étude des goûts musicaux, leurs liens avec les genres et les paramètres contextuels qui peuvent influencer l'écoute de musique. Cette étude met en relation des conclusions issues de notre bibliographie et des données provenant d'une plateforme de streaming. Dans le second chapitre, nous nous sommes intéressés aux données, aux méthodes, et aux modes d'évaluation employées par les plateformes de streaming pour la recommandation. Nous avons également proposé une méthode afin de lier des paramètres acoustiques à une catégorisation humaine, en vue d'une recommandation de musique personnalisée. Le troisième chapitre présente une méthode de prédiction de goûts mise au point en nous appuyant sur les conclusions des chapitres précédents. Cette méthode est basée sur des réseaux de neurones convolutionnels profonds pré-entraînés pour la prédiction de genres. Les résultats de cette

expérience sont présentés et analysés de manière quantitative.

Remerciements

En premier lieu je tiens à remercier Julien Pinquier, Christine Sénac et Pascal Gaillard de m'avoir donné l'opportunité de mener ces travaux de thèse sur ce sujet passionnant. Merci pour votre encadrement toujours efficace, votre bienveillance et votre patience durant de ces trois années.

Merci à mes deux rapporteurs Hervé Glotin et Geoffroy Peeters, pour leurs remarques pertinentes et constructives ainsi qu'aux membres du jury, Myriam De Sainte-Catherine et Gaël Richard. Les discussions auxquelles ont donné lieu la soutenance de cette thèse m'ont permis de conclure ce chapitre de la meilleure des manières.

Merci à Ludovic Florin, Paul Albenge, Antoine Vervier et Stéphane Escoubet, pour leurs apports en musicologie et pour les découvertes musicales. Ce sont ces échanges qui ont permis d'apporter à cette thèse son caractère singulier et, pour moi, une bonne partie de son intérêt.

Merci à Manuel Moussallam et à Deezer pour m'avoir fourni des données qui m'ont permis de mener des études passionnantes.

Je tiens à exprimer toute ma gratitude envers tous les volontaires pour leur participation aux expériences que j'ai menées. Merci pour votre temps et pour votre aide.

Merci à l'ensemble de l'équipe SAMoVA pour son accueil. J'ai pu profiter en son sein d'un cadre de travail et d'une ambiance agréable, merci pour votre bonne humeur.

Merci à mes amis et à mon entourage pour votre soutien. J'ai toujours passé d'agréables moments de détente en votre compagnie, vous m'avez permis de décompresser lorsque cela était nécessaire. Par ailleurs, j'ai la conviction que ces longues discussions passées à vulgariser mon sujet m'ont permis de mieux le maîtriser. Ainsi, je suis redevable à vous et à votre curiosité, d'une partie de ma réussite.

Je remercie mes parents et ma famille pour m'avoir poussé vers cette voie. Sans vous je ne serais jamais arrivé jusqu'ici.

En dernier lieu, merci aux modèles Nouka et Kobe d'avoir accepté de prendre la pose pour illustrer mon propos en Partie 2.6.

Table des matières

Introduction : la recommandation de musique

Enjeux du streaming de musique et de la recommandation de musique

En France, d'après le rapport du Syndicat National de l'Edition Phonographique (SNEP) 2020 sur les revenus de la musique enregistrée[1], les revenus des ventes de musique au format numérique ont augmenté de près de 18%. Au premier semestre de l'année 2020, le format numérique représente 80% du chiffre d'affaire des ventes : 76% de streaming, 4% de téléchargement, 20% de physique. En France, le nombre d'écoutes en ligne a été multiplié par 5 entre 2014 et 2018 et s'approche à présent de 80 milliards par an, représentant ainsi de loin la première source de revenus issus de la musique enregistrée.

Cette tendance se retrouve ailleurs dans le monde et aux États-Unis notamment, où la RIAA (Recording Industry Association of America) rapporte quant à elle une part de marché du streaming de 85% dans la vente de musique[2].

Sur l'année 2019, la SNEP indiquait dans son rapport[3] que cette tendance ne concerne pas que le jeune public : un quart des utilisateurs des plateformes de streaming a plus de 55 ans.

Bien que l'augmentation de la part d'écoute de musique en streaming suive l'évolution des années précédentes, cette tendance est également imputable à la pandémie mondiale de covid-19. Non seulement cette épidémie a eu un fort impact sur les ventes de musique en format physique, mais elle a mis à l'arrêt l'industrie du spectacle vivant et des concerts[4].

Cette situation est d'autant plus préoccupante pour le secteur que les concerts étaient jusqu'alors la source de revenus principale pour les artistes, bien au delà

1. https://snepmusique.com/non-classe/musique-enregistree-resultats-du-premier-semestre-2020/
2. https://www.riaa.com/wp-content/uploads/2020/09/Mid-Year-2020-RIAA-Revenue-Statistics.pdf
3. https://snepmusique.com/actualites-du-snep/marche-2019-de-la-musique-enregistree-decryptage-des-resultats/
4. https://snepmusique.com/communiques-dossiers-de-presse/tplm-etude-dimpact-du-covid-19-sur-la-filiere-musicale/

des ventes physiques ou du streaming, que ce soit pour les meilleures ventes[5] ou pour les artistes indépendants[6].

Dans ce contexte de crise, le streaming de musique apparaît donc comme l'un des derniers modes de consommation de musique possible, et donc l'une des dernières sources de revenus pour les artistes. Le streaming de musique est donc crucial pour une industrie de la musique plus que jamais en péril. C'est aussi pourquoi une recommandation musicale intelligente permettant la découverte de nouveaux talents artistiques dans un style/genre musical donné est aussi essentielle pour les plateformes de streaming.

Les utilisateurs des plateformes de streaming ont accès à un catalogue immense avec 56 millions de titres disponibles chez Deezer[7] et plus de 60 millions chez Spotify[8]. Le rôle de la recommandation de musique est de guider les utilisateurs de ces plateformes en leur proposant des titres à écouter. La recommandation de musique peut se faire soit via des playlists éditoriales, c'est-à-dire des playlists composées manuellement par des curateurs, soit de manière automatique, c'est ce dernier cas que nous traiterons au cours de cette thèse. Cette recommandation musicale s'appuie sur un ensemble d'outils informatiques dont certains sont issus de l'intelligence artificielle. Lorsque la recommandation automatique est basée sur le signal audio, elle met en œuvre des méthodes appartenants au champ académique très actif appelé *Music Information Retrieval*.

Contexte académique : intelligence artificielle et recherche d'information musicale

La recherche d'information musicale, ou *Music Information Retrieval* (MIR) en anglais est un domaine de recherche interdisciplinaire au croisement - entre autres - du traitement du signal, de la psychoacoustique et de la musicologie. Au delà de la recommandation de musique, les domaines d'application de la MIR concernent également :

— La génération automatique de musique [21].
— La séparation de morceaux en plusieurs pistes instrumentales [112].
— La transcription automatique de musique en partitions [11].
— La classification automatique en genres de musique [50].
— Etc.

Les outils utilisés en MIR (et plus généralement en recommandation) sont informatiques et font de plus en plus appel à des algorithmes d'« Intelligence Artificielle » (IA).

5. https://www.billboard.com/photos/8520668/2018-highest-paid-musicians-money-makers
6. https://thecreativeindependent.com/music-industry-report/
7. https://www.deezer.com/fr/company
8. https://newsroom.spotify.com/company-info/

L'application difficile de l'IA à la modélisation des goûts musicaux

Les méthodes dites d'IA sont ainsi nommées, car elles ont pour but de mimer des fonctions cognitives de haut niveau, telles que la reconnaissance d'objets, de sons, ou encore la prédiction d'évènements. L'intelligence est donc ici considérée comme un système complexe, en sortie duquel est renvoyée une réponse appropriée selon les données qui lui sont transmises en entrée et le contexte. Le terme d'IA désigne un ensemble de méthodes auxquelles appartiennent les méthodes d'apprentissage machine (*machine learning* en anglais). Ces méthodes appartiennent aux sciences de la modélisation. La particularité de l'apprentissage machine est que les modèles sont construits uniquement à partir de données fournies à un algorithme. L'expertise humaine n'est alors plus utile qu'à la sélection des paramètres permettant de représenter de manière pertinente les données à modéliser. L'apprentissage automatique peut se diviser en deux catégories : supervisé et non-supervisé.

— Le mode supervisé désigne la présence d'annotation des données en tant que vérité terrain. Par exemple, pour une tâche de reconnaissance de véhicule, des annotations peuvent être « voiture », « moto » ou encore « bateau ». Un paramètre pertinent pour discriminer ces trois classes peut être le nombre de roues. Avec un modèle construit sans IA, nous écririons un algorithme en tenant compte de ces caractéristiques connues : une voiture a 4 roues, une moto 2 et un bateau 0. En utilisant une IA supervisée, l'algorithme apprend ces caractéristiques seulement à partir de multiples exemples qui lui sont fournis, où chaque donnée est une association de paramètres à une annotation. Si les paramètres choisis sont pertinents et les exemples suffisamment nombreux, l'algorithme parviendra peut-être à établir une règle similaire ou identique à celle donnée par un expert, associant un nombre de roues à un type de véhicule.

— Le mode non-supervisé désigne le cas où les données fournies à l'algorithme ne sont pas annotées. Un algorithme non-supervisé peut séparer les données en plusieurs groupes sur la base seule des paramètres qui lui sont fournis. Ici, l'algorithme n'apprend pas une catégorisation fournie dans des annotations, mais plutôt une structure sous-jacente contenue dans les paramètres. En établissant des liens entre plusieurs variables, ces algorithmes détectent une organisation dans les données et peuvent ainsi être utilisés à des fins de compression de l'information, c'est par exemple le cas des autoencodeurs [6].

Davantage de détails sont donnés sur les méthodes supervisées et non-supervisées dans la partie 2.1.

Les récents progrès dans le domaine de l'informatique ainsi que l'abondance de données à disposition ont permis l'avènement d'une nouvelle discipline d'apprentissage machine : l'apprentissage profond (*deep learning* en anglais). La modélisation a pu alors s'affranchir de l'extraction de paramètres pour représenter les données : nous parlons alors de systèmes *end-to-end*, où des signaux bruts tels que des images ou des sons peuvent être traités directement (sans

paramétrisation au préalable). L'IA a su montrer son efficacité dans un grand nombre de domaines, allant de la finance à la médecine : elle est à présent souvent privilégiée pour la résolution de nouveaux problèmes.

Dans le cas de la recommandation de musique, les algorithmes ont pour rôle de modéliser des comportements ou des goûts musicaux : un utilisateur va-t-il aimer ce morceau ? Va-t-il l'écouter en entier ? Va-t-il le réécouter plus tard ? Les paramètres qui influent sur les affects et les comportements des individus sont aussi nombreux que variés, car les seules propriétés acoustiques des morceaux écoutés ne suffisent souvent pas à expliquer les goûts des utilisateurs. Comme le montre la figure 1, le contexte d'écoute, l'entourage de l'auditeur ainsi que son bagage culturel sont autant d'éléments qui peuvent avoir une influence sur sa perception et ses affinités musicales.

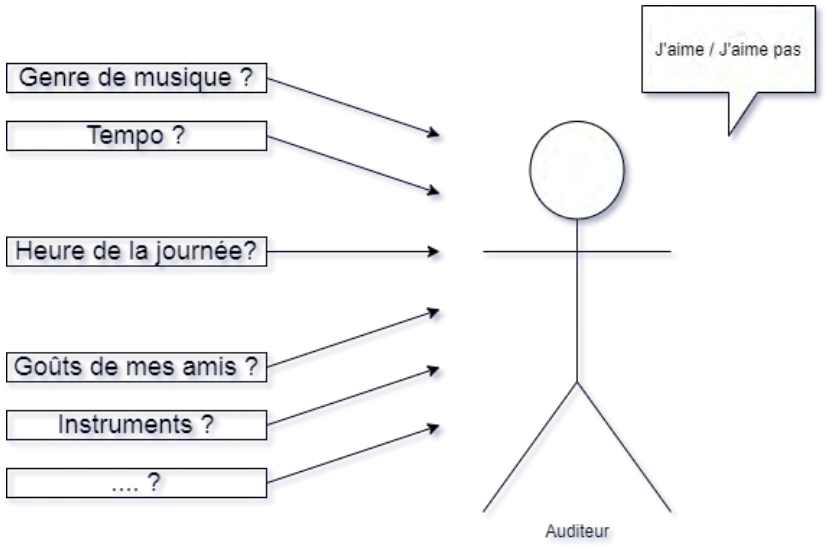

FIGURE 1: Quels éléments influencent les goûts ?

Par ailleurs, si les méthodes d'apprentissage profond présentent parfois de meilleures performances, en termes de scores, elles demeurent souvent trop peu explicables : l'entraînement automatique aboutit généralement à des modèles opaques, où il demeure difficile pour l'utilisateur de savoir quels paramètres influencent le choix des algorithmes.

La figure 2 illustre ce phénomène : même si une IA peut être capable de fournir des recommandations satisfaisantes, elle apparaît souvent comme une « boîte noire ». Les utilisateurs consentent à fournir leurs données aux plateformes afin de recevoir des recommandations, par un processus sur lequel ils n'ont généralement aucune prise.

L'explicabilité des modèles issus de l'IA demeure un challenge important, car elle permet aux utilisateurs de mieux comprendre les choix des algorithmes,

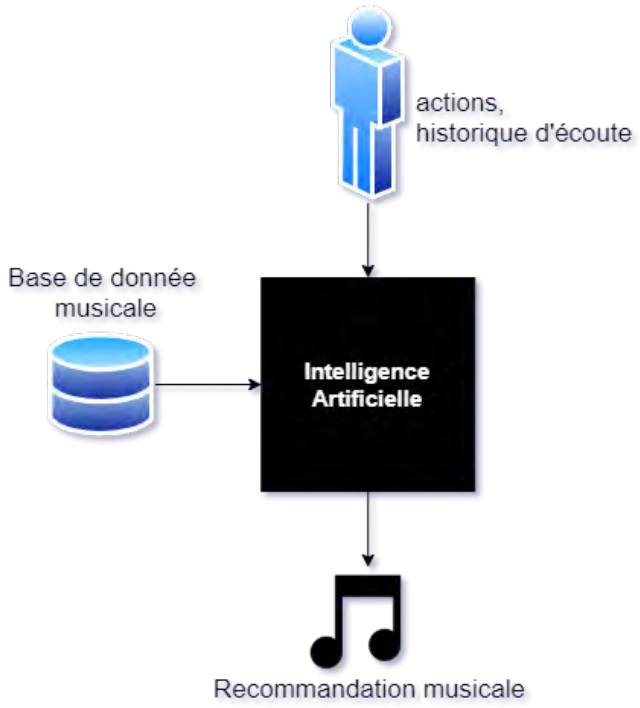

FIGURE 2: Schéma général de la recommandation de musique.

améliorant ainsi la transparence des plateformes et la confiance des utilisateurs vis-à-vis de ces dernières [110]. Par ailleurs, en expliquant le choix qui a motivé l'algorithme à proposer un morceau à un auditeur, nous pouvons améliorer l'efficacité de la recommandation [72] en donnant une information de contexte : *« Si vous avez aimé ce morceau, alors vous aimerez peut-être celui-ci »*. Néanmoins, si comme dit précédemment les goûts musicaux sont dépendants d'un nombre inconnu de facteurs, la difficulté de l'explicabilité est double : comment expliquer le choix d'un modèle, si nous ne connaissons pas suffisamment bien le phénomène à modéliser ? Pour répondre à cette question, notre démarche a été de nous tourner dans un premier temps davantage vers l'humain que vers la machine (IA), dans le but de proposer des pistes d'amélioration de cette dernière.

Problématique : jusqu'où les goûts musicaux sont-ils prédictibles par l'IA ?

La recommandation de musique ayant pour but de prédire les affinités des utilisateurs et compte tenu de la double difficulté entre l'explication des goûts

musicaux et l'explicabilité des modèles, les travaux de cette thèse se sont concentrés sur la problématique suivante : jusqu'où les goûts musicaux sont-ils prédictibles par l'IA ?

Cette problématique peut être décomposée en plusieurs sous-problématiques :

— Quels éléments influencent les goûts musicaux ? Quels sont les liens entre les genres musicaux et les goûts des auditeurs en matière de musique ?

— La recommandation de musique est-elle une prédiction de goûts ? Comment les données comportementales sont-elles interprétées par les plateformes de streaming ? Le comportement traduit-il vraiment des goûts musicaux ?

— Les paramètres acoustiques appris par un réseau de neuronnes convolutionnel pour la reconnaissance automatique de genre sont-ils pertinents pour la prédiction de goûts musicaux ? Quels facteurs déterminants dans l'affinité musicale un réseau de neurones convolutionnel est-il capable d'identifier dans un spectrogramme ?

Ces problématiques peuvent faire appel à un large panel de disciplines académiques. Sur ces questions, les connaissances d'un musicologue sont probablement aussi pertinentes que celles d'un informaticien. Par ailleurs, nous verrons que la sociologie peut également apporter des éléments de réponse, tout comme l'acoustique.

Contributions principales

Bien que le sujet de la recommandation de musique soit pluridisciplinaire, nous avons utilisé les outils qui sont les nôtres, à savoir l'informatique et le traitement du signal. Cependant, loin d'ignorer les autres sciences, nous avons travaillé dans le cadre de cette thèse avec les musicologues Ludovic Florin, Paul Albenge, Stéphane Escoubet ainsi qu'Antoine Vervier[9] afin de profiter de leur expertise. Ainsi, au cours des travaux de cette thèse, nous avons tiré profit des sciences humaines et de la musicologie comme autant d'apports extérieurs nous permettant de positionner et d'enrichir notre réflexion.

Pour répondre aux questions soulevées précédemment, nous avons dans un premier temps interrogé la nature même des goûts musicaux, puis étudié les outils actuels de recommandation de musique, pour enfin proposer une méthode de prédiction de goûts tenant compte de nos résultats.

Les travaux effectués durant cette thèse ont apporté 3 contributions principales :

— Nous avons **caractérisé des comportements d'écoute** à travers l'étude de données d'une plateforme de streaming, en nous focalisant sur l'influence des genres, de l'heure d'écoute ainsi que sur les interactions des utilisateurs avec la plateforme.

— Nous avons mené une expérience de catégorisation avec 20 volontaires et nous avons tenté de donner une **explication de la catégorisation**

9. Université de Toulouse - Jean Jaurès

obtenue par des critères musicologiques ainsi que par des paramètres acoustiques. Pour cela, nous avons au préalable constitué un corpus.

— **Nous avons implémenté une méthode de transfer learning pour la prédiction de goûts basée sur la classification en genres**. Nous avons constitué un corpus d'affinités musicales auprès d'un groupe de volontaires, comparé plusieurs corpus de reconnaissance en genres, et utilisé un réseau de neurones convolutionnels pour la prédiction de goûts.

Plan détaillé

Dans le chapitre 1, nous allons expliciter la notion de goûts, étudier les facteurs sociaux, culturels et temporels qui les influencent. Par ailleurs nous étudierons le lien entre les genres musicaux et les goûts. Nous verrons enfin comment les goûts des utilisateurs des plateformes de streaming tendent vers une forte uniformisation.

Dans la chapitre 2, nous répondrons à la question suivante : la recommandation de musique est-elle une prédiction de goûts ? Après une présentation de l'apprentissage automatique, nous nous intéresserons à la pertinence des données utilisées pour l'entraînement de ces algorithmes. Nous étudierons les 2 approches principalement utilisées en recommandation musicale, le filtrage collaboratif et l'approche basée « contenu », ainsi que leurs avantages et leurs défauts. Nous présenterons une expérience afin de lier des paramètres acoustiques et des critères musicologiques à une catégorisation effectuée par plusieurs volontaires.

Lors du chapitre 3, nous présenterons une méthode de prédiction de goûts puis nous analyserons ses performances et ses limites, à travers une étude qualitative des résultats des volontaires.

Enfin, nous apporterons une conclusion sur les travaux effectués afin de répondre aux problématiques et nous donnerons des pistes de perspectives à court et moyen terme.

Chapitre 1

Les goûts musicaux et les genres

Avant de nous intéresser aux méthodes de prédictions de goût, nous allons nous intéresser au goût musical, à sa définition et aux différents contextes pouvant l'influencer. Ce travail se fera à travers une étude bibliographique que nous avons effectuée ainsi qu'une analyse de données provenant de la plateforme Deezer. Le lien entre les goûts et les genres musicaux sera par ailleurs étudié, ainsi que la concentration des écoutes sur certains morceaux. Nous conclurons alors sur le rôle à jouer des algorithmes de recommandation dans la diversification des horizons musicaux des auditeurs.

1.1 Préambule : enquête de musicologues

Dans le cadre du projet PERMUSES (Processus d'Evaluation des Recommandations de MUSique En Streaming)[1] sur lequel nous avons travaillé durant cette thèse, nous avons collaboré avec deux musicologues, Stephane Escoubet et Antoine Vervier. Ils ont réalisé une enquête sur 10 participants âgés de 20 à 25 ans qui étaient chargés d'écouter et de trier 26 extraits musicaux en 3 catégories : « j'aime », « je n'aime pas », « mitigé ». Le corpus, constitué pour l'occasion, était composé de 26 extraits ($\approx$ 1 min) de morceaux appartenant à 13 genres différents. La liste des morceaux est disponible en annexe E. L'un des buts de cette expérience était de mettre en évidence un éventuel lien entre les goûts musicaux des participants et les genres de morceaux. Après l'écoute des morceaux et la catégorisation, un entretien a été mené avec chaque participant afin d'expliquer davantage sa classification.

1. Appel à projets exploratoires (AAPEX 201) - Maison des Sciences de l'Homme de Toulouse (MSH-T)

1.1.1 Caractéristiques musicales

Pour effectuer leur catégorisation, la logique de genre musical semble avoir joué un rôle important pour les participants : la moitié des extraits proposés ont été appairés par genre dans la même catégorie d'appréciation par les participants. Le critère de genre a par ailleurs été cité régulièrement lors des entretiens.

Pour les appréciations positives des morceaux, la présence d'un genre musical déjà apprécié est apparue comme une condition nécessaire, mais pas suffisante. Les morceaux des genres méconnus des volontaires ont reçu peu d'appréciations positives, et les morceaux des genres connus et appréciés des participants peuvent avoir reçu des appréciations négatives. Par exemple, nous avons observé une discrimination forte des deux morceaux de reggae ou de pop, voire une discrimination au sein du répertoire d'un même artiste. Un participant a notamment cité Orelsan pour lequel il peut apprécier les morceaux « quand il arrive à faire des chansons à texte, et pas du commercial, et puis quand il arrive à faire des trucs très planants », mais n'a pas aimé le morceau présenté, car jugé en l'occurrence trop commercial.

Au delà des genres, les participants ont listé d'autres caractéristiques déterminantes dans leurs appréciations. La qualité de la voix a été régulièrement citée, jugée indépendamment de la partie instrumentale. Outre sa qualité, sa prégnance a été déterminante, ainsi que les paroles si elles peuvent être comprises par l'auditeur.

1.1.2 Une affaire d'usages et de contextes

La compatibilité d'une musique avec des usages (s'isoler dans les transports en commun, faire son footing, musique de fond pour travailler...) est apparue dans les témoignages comme un élément décisif dans l'appréciation de nouveaux horizons musicaux. Un participant a notamment souligné l'intérêt de morceaux sans paroles ou adaptés à une écoute passive, dans un contexte de travail. Par exemple, le morceau *Dog Trot* de Initiative H et Moondog a pour ce participant l'intérêt d'être peu entraînant, limitant ainsi le risque de déconcentration lorsqu'une tâche effectuée en parallèle de l'écoute nécessite une forte attention. Le caractère « peu entraînant » de ce morceau représente donc une qualité dans un contexte de travail, mais pourrait, pour d'autres utilisateurs ou d'autres contextes être un défaut.

Le degré d'implication dans le choix des morceaux a été également cité : des morceaux que nous ne sommes pas prêts à écouter de notre propre initiative, mais que nous sommes toutefois disposés à apprécier dans certaines circonstances. Dans ce cas, les participants ont utilisé la catégorie « mitigé » pour y placer ces morceaux, qu'ils pourraient apprécier s'ils étaient issus d'une recommandation automatique, si un ami leur avait proposé, ou s'ils y étaient exposés par hasard lors d'un concert. Pour ces volontaires, ces morceaux sont plutôt appréciés bien qu'ils ne fassent pas partie de leur univers musical habituel.

Enfin, la fréquence d'écoute a été identifiée comme un paramètre déterminant : certains morceaux sont davantage compatibles avec une écoute ponctuelle

alors que d'autres sont plus aptes à une écoute régulière. Ainsi, certains participants ont utilisé la catégorie « je n'aime pas » pour les morceaux qu'ils ne voudraient jamais écouter, « j'aime » pour des morceaux qu'ils pourraient écouter régulièrement, et « mitigé » pour des morceaux qu'ils pourraient écouter ponctuellement.

Les différents points évoqués nous montrent qu'une question en apparence triviale sur les affinités musicales peut être interprétée différemment selon les usages des personnes interrogées. Si nous désirons obtenir des réponses sur les seules caractéristiques musicales des morceaux de musiques, alors ces usages et contextes d'écoutes doivent être cadrés.

1.1.3 Hypothèses

Cette étude, portant sur seulement 10 participants, ne permet pas de conclure sur les facteurs influençant les goûts des auditeurs. Néanmoins, elle nous a permis d'émettre un certain nombre d'hypothèses et de questions :

— Les genres semblent avoir une influence sur les goûts des personnes interrogées. Peuvent-ils expliquer une catégorisation « j'aime »/« je n'aime pas » sur une plus grande quantité de données ?

— Les témoignages évoquent des contextes d'écoutes plus propices à certaines musiques que d'autres. Est-il possible d'observer des moments plus propices à certains genres dans les données massives des plateformes de streaming ?

Dans ce qui suit, nous allons tenter d'étudier ces questions et hypothèses à travers une étude bibliographique et l'étude de données fournies par le site Deezer.

1.2 Influence du contexte sur les goûts musicaux

1.2.1 Contexte social et culturel

L'influence du contexte social sur les goûts d'une population a été étudiée pour la première fois en 1979 par Bourdieu [17]. En se basant sur un sondage auprès d'échantillons représentatifs, il a montré une distinction entre les goûts musicaux des classes aisées et ceux des classes populaires. La distinction était alors observée entre les genres de musiques savantes (classique et jazz) et les genres de musiques populaires (chanson, rock, etc.). Ce constat a été ultérieurement confirmé à travers d'autres observations [41, 115]. Ainsi, bien avant l'avènement de l'intelligence artificielle, la sociologie proposait de prédire les goûts musicaux des individus selon leur groupe social d'appartenance.

En 1990, Simkus et Peterson ont utilisé une méthode de corrélation linéaire pour noter simultanément le rang professionnel et les goûts musicaux d'un échantillon national de la population américaine [86]. Ils en ont conclu que le clivage culturel entre les classes sociales supérieures et inférieures ne s'observe plus seulement à travers les genres, mais plutôt à travers la variété des

genres écoutés au sein d'un même groupe social : les milieux favorisés auraient tendance à avoir une consommation culturelle plus diversifiée que les milieux moins favorisés.

Il y a toujours débat entre les deux théories dominantes de Bourdieu et Peterson [26, 4, 88, 71]. Les goûts musicaux constituent toujours un champ de recherche en sociologie, avec de nouvelles théories telles que « la tablature des goûts musicaux », proposée par Glevarec et Pinet [53]. La figure 1.1 donne une vision d"ensemble des théories sociologiques des goûts musicaux : quand Bourdieu établissait une hiérarchie entre les genres sous forme de colonne, et Peterson illustrait le degré de diversité sous forme de pyramide inversée, la tablature place les genres côte à côte, la hiérarchisation de la colonne se retrouvant ainsi horizontale. Néanmoins, des hiérarchies verticales subsistent toujours dans cette tablature, mais elles ne sont observables qu'à l'intérieur des genres eux-mêmes. Les goûts sont alors représentés sous la forme d'ensembles : leur étendue verticale définira de degré d'amateurisme ou d'expertise dans un genre donné, et l'étendue horizontale définira l'omnivorité.

FIGURE I. – ***Colonne, pyramide et tablature***

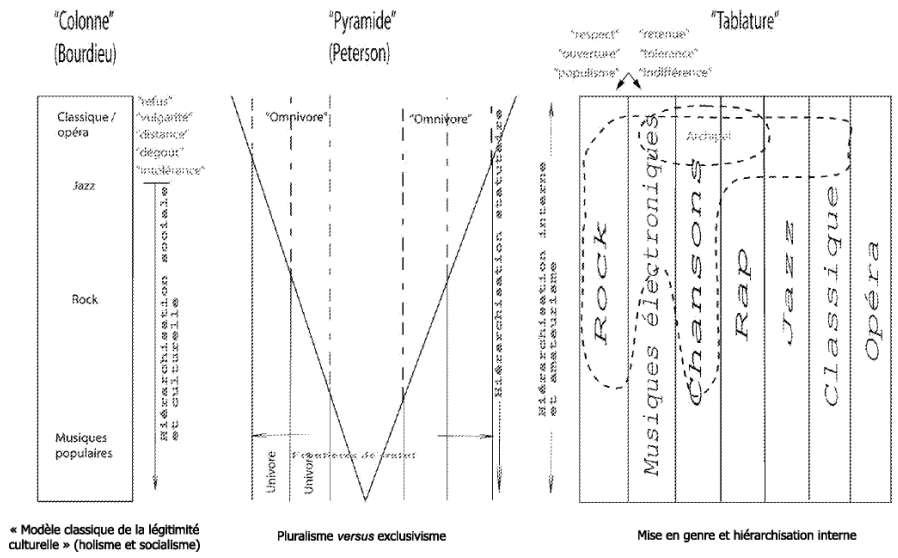

FIGURE 1.1: Les trois grandes théories sociologiques des goûts musicaux. Figure tirée de l'article « La tablature des goûts musicaux : un modèle de structuration des préférences et des jugements » [53].

À une plus petite échelle, Schäfer et. al ont montré en 2016 l'influence de l'opinion de l'entourage social dans les goûts musicaux [94].

Dans le domaine de la recommandation de musique, certaines études considèrent ces facteurs sociaux. Ainsi dans [126] les auteurs ont modélisé le contexte culturel des auditeurs à travers des indicateurs socioculturels de leur pays d'ap-

partenance. Ils ont montré que ces informations utilisées conjointement à des paramètres acoustiques ont permis d'améliorer les modèles de recommandation musicale. Par ailleurs, dans [23], ce sont les relations sociales des utilisateurs du site LastFM[2] qui ont été utilisées en vue d'améliorer les outils de recommandation.

Coulangeon a montré que l'âge et en particulier la génération sont également très importants [33]. Des résultats similaires ont été mis en évidence dans une étude plus récente [45]. L'âge ou la génération des auditeurs peut être considéré comme un facteur social et temporel.

Nous allons, à présent, nous intéresser à des facteurs temporels à plus court terme, tels que l'activité au cours de la journée.

1.2.2 Contexte temporel

Différentes recherches, basées sur des entretiens, ont montré que le type d'activité a un rôle important sur les habitudes d'écoutes [61]. Par exemple, les auditeurs ont davantage tendance à choisir les morceaux écoutés un par un lors d'une écoute attentive, alors qu'ils se contenteront souvent d'une playlist toute faite pendant qu'ils feront du sport (voir figure 1.2). Par ailleurs, les playlists constituées par les utilisateurs de plateformes de streaming sont souvent liées à un contexte d'écoute précis [35]. Nous n'écoutons pas la même musique pour travailler, pour faire du sport ou pour nous reposer.

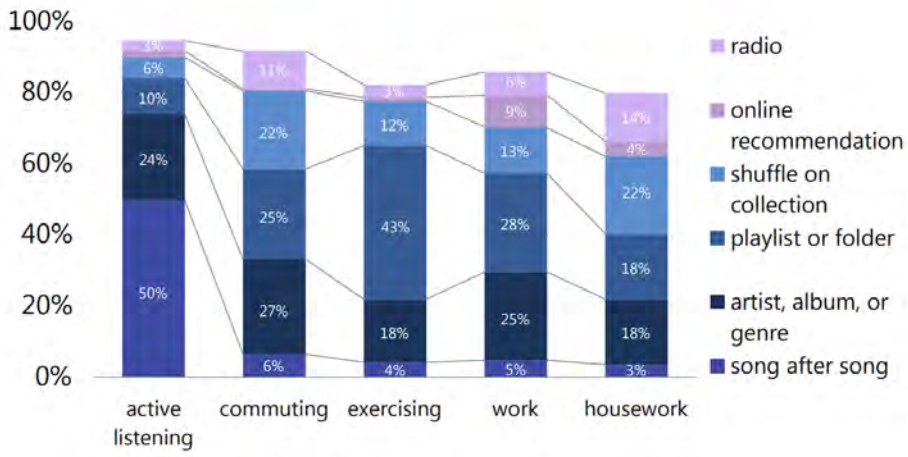

FIGURE 1.2: Média de diffusion utilisé selon l'activité. Figure issue de [61].

Ces pistes ont été étudiées afin de tenir compte des différents contextes d'écoutes dans la recommandation de musique. Dans [7], les auteurs utilisent plusieurs sous profils pour chaque utilisateur. Par exemple, un utilisateur peut être décrit à travers 2 sous profils : matin et week-end. Dans [119], les auteurs

2. https://www.last.fm/fr/

utilisent un ensemble de capteurs de smartphones tels que des gyroscopes, accéléromètres, GPS, microphones et capteurs de lumière afin de déterminer l'activité de l'utilisateur parmi les suivantes : courir, marcher, dormir, travailler, étudier, faire du shopping. Des morceaux sont ensuite proposés aux utilisateurs en fonction de leur activité. [87] s'est appuyé sur le nom des playlists afin d'en extraire des informations sur le contexte d'écoute, comme par exemple « musique de Noël » ou bien « fête ». Dans [40], les auteurs tiennent compte de l'heure, du jour de la semaine, et de la date pour définir le contexte. Par ailleurs, ils caractérisent également une session d'écoute par la diversité des morceaux écoutés par un utilisateur.

1.3 Les genres suffisent-ils à décrire les goûts ?

Nous venons de voir différentes études sur ce qui peut influencer les goûts musicaux. Dans la majorité d'entre elles, les goûts musicaux sont expliqués au travers des genres : certains genres seraient plus écoutés par certaines classes sociales, certains genres sont plus adaptés à certaines activités... Ici, nous allons, à l'aide de données fournies par Deezer, tenter de vérifier l'hypothèse selon laquelle les goûts musicaux sont liés aux genres.

1.3.1 Descriptif des données Deezer

Deezer est une plateforme d'écoute de musique en ligne qui propose un catalogue de plus de 56 millions de titres différents. Nous avions à notre disposition un *log* provenant du site Deezer. En informatique le terme de log désigne un fichier qui contient un historique d'évènements. Dans ce log est inscrite l'activité de 10.000 utilisateurs français âgés de 18 à 24 ans durant une semaine de juin 2018. Chaque ligne de ce log contient un évènement d'écoute d'un morceau par un utilisateur. Le morceau est identifié par un numéro, qui permet de retrouver plus d'informations via l'API de Deezer[3], comme le nom du morceau ou l'artiste, ainsi que son genre. Ici, les genres sont au nombre de 92. Dans le reste de ce chapitre, le genre des morceaux correspond à celui donné par cette API, sauf indication contraire. Chaque utilisateur est identifié par un numéro anonyme, qui permet toutefois de connaître l'ensemble des morceaux écoutés par un utilisateur donné au cours de cette semaine. Enfin, nous avions à disposition la durée de chaque écoute, ainsi que des variables nous indiquant si l'utilisateur a aimé le morceau (**like**), s'il est passé au suivant, ou s'il a appuyé sur le bouton **ban**, indiquant *a priori* une aversion pour le morceau. Un extrait de ce log est disponible en annexe C.

3. `https://developers.deezer.com/api`

1.3.2 Analyse ban/like du log deezer

Expérience 1

Sur les 10.000 utilisateurs, nous avons conservé uniquement ceux qui ont au moins 3 likes et au moins 3 bans dans la semaine. Ce seuil a été fixé afin d'avoir suffisamment de morceaux par utilisateur sélectionné tout en ayant suffisamment d'utilisateurs. Nous avons ainsi obtenu 68 utilisateurs. A l'aide de l'API de Deezer, et pour chaque utilisateur, nous avons observé quel genre est majoritaire parmi les morceaux aimés et parmi les morceaux bannis. Nous avons observé ensuite si le genre majoritaire est le même pour les morceaux « likés » et pour les morceaux « bannis ».

1. Cas 1 : 35 utilisateurs avec le même genre majoritaire pour like et ban.
2. Cas 2 : 33 utilisateurs avec des genres majoritaires différents pour like et ban.

Dans le cas 2, où le genre majoritaire est différent, nous avons observé 13 égalités de genre entre le premier et le second morceau d'une des deux catégories, ou bien entre les seconds morceaux de chaque catégorie.

Par exemple, voici les genres décomptés pour un utilisateur :

— Morceaux like :
alternative : 9, pop : 4, reggae : 1, singer & songwriter : 1, rap/hip-hop : 1, dance : 1, chanson française : 1
Genre Majoritaire like : alternative

— Morceaux ban :
pop : 11, alternative : 7, rock : 4, chanson française : 2, rap/hip-hop : 2, r&b : 2, dance : 1, reggae : 1
Genre majoritaire ban : pop

Pour cet utilisateur, le genre majoritaire est différent pour like et ban. Cependant, il y a une égalité parmi les 2 genres majoritaires, puisque pop est le deuxième genre le plus représenté parmi les likes et le premier représenté parmi les ban, et inversement ave le genre alternative.

Ainsi, sur ces 68 utilisateurs, 35 ont le même genre majoritaire dans la catégorie ban et like et 13 ont au moins un genre identique parmi les 2 les plus aimés et bannis. Pour 48 utilisateurs sur 68 observés (70%), nous pouvons considérer que les genres de Deezer ne permettent pas de discriminer les 2 catégories.

Cette expérience a été basée sur les genres de Deezer, qui sont au nombre de 92 dans les données observées, avec une forte disparité dans leur répartition en termes d'écoute (et probablement de catalogue).

Expérience 2

La même expérience a été réalisée en utilisant cette fois les genres selon l'API de Spotify[4], qui en compte près de 4000 différents, là où Deezer n'en comptait

4. https://developer.spotify.com/documentation/web-api/libraries/

que 92. Ce nombre de genres bien supérieur n'est pas l'indicateur d'une plus grande diversité dans la bibliothèque de Spotify mais seulement d'une précision sur les sous-genres. Par exemple, quand l'API de Deezer nous indique qu'un morceau appartient au genre *metal*, celle de Spotify pourrait nous préciser qu'il s'agit en fait de *grisly death metal*, *japanese black metal* ou bien encore de *progressive technical death metal*. Nous avons par ailleurs distingué le troisième cas évoqué précédemment, où des genres identiques se retrouvent parmi les 2 premiers genres les plus représentés des catégories like et ban.

— Cas 1 : 16 utilisateurs ont le même genre Spotify majoritaire pour like et ban.
— Cas 2 : 52 utilisateurs ont des genres Spotify majoritaires différents pour like et ban.
— Cas 3 : 10 utilisateurs ont des genres Spotify majoritaires égaux parmi les 2 majoritaires.

Le fait d'utiliser une nomenclature plus précise permet donc de mieux distinguer les morceaux entre like et ban. Pour un utilisateur, les genres Deezer majoritaires pour les deux catégories étaient 'Pop'. En refaisant cette analyse avec les genres de Spotify, les 2 genres les plus représentés pour la classe like sont : french pop et french rock. Pour la classe ban, les deux genres les plus présents sont : pop et dance pop. Nous pouvons donc supposer que cet utilisateur aime la musique pop française en particulier. Nous voyons que les genres plus précis de Spotify nous permettent de discriminer les 2 catégories dans plus de cas.

Une analyse des morceaux aimés et bannis par les utilisateurs a montré que les genres sont bien souvent insuffisants pour expliquer leurs goûts. Des sous-genres plus précis comme ceux utilisés par Spotify permettent néanmoins une meilleure distinction. D'autre part, au-delà des sous-genres toujours plus précis et restreints, serait-il pertinent d'utiliser un nombre réduit de critères musicologiques permettant d'expliquer les goûts, indépendamment des genres et sous-genres ? L'explication des goûts par des critères musicologiques sera abordée dans le chapitre suivant de cette thèse (voir partie 2.6).

1.3.3 Conclusion

L'étude bibliographique précédemment effectuée montre que certains contextes socioculturels et l'activité effectuée en parallèle de l'écoute ont une influence sur les genres écoutés. Une expérience simple a montré qu'une centaine de genres différents n'étaient pas suffisamment précis pour expliquer les affinités des utilisateurs. Une nomenclature plus précise, comme celle de Spotify permet d'améliorer ces prédictions. La précision apportée permet de supposer que la différence d'appréciation entre les morceaux d'un même genre porte bien sur des traits musicaux et non pas seulement sur un contexte d'activité différent. Le contexte peut avoir un impact sur les genres écoutés, mais les goûts musicaux ne peuvent pas être simplement réduits à de simples genres ou au contexte d'écoute.

1.4 Caractérisation des genres selon le contexte

Nous étudions dans cette partie différentes manières d'établir des liens entre les genres à travers les statistiques d'écoute des auditeurs présents dans le log de Deezer ainsi que des annotations en genre des morceaux.

1.4.1 Caractérisation des genres selon les concentrations d'écoutes

Nous avons effectué un décompte des genres des morceaux écoutés dans le log Deezer. Il apparaît que les genres les plus écoutés ne sont pas nécessairement les plus représentés en nombre d'occurrences (nombre de morceaux différents). Par exemple, il y a sensiblement autant de morceaux **différents** appartenant aux genres classique et rap français qui ont été écoutés, mais les morceaux de rap français ont été écoutés près de 20 fois plus (voir table 1.1).

Ainsi, pour un genre donné, il peut être intéressant de calculer le ratio :

$$R = \frac{E}{O} \tag{1.1}$$

avec O le nombre de morceaux différents écoutés, et E le nombre total d'écoutes de ces morceaux.

Ce ratio R nous permet de déterminer si un genre a tendance à accumuler des écoutes sur un nombre restreint de morceaux, ou s'il disperse son audience sur un plus grand nombre de titres. Les résultats ont montré que le genre classique affiche un ratio écoutes/occurrence de 2,7 contre 52,5 pour le rap français.

Ce descripteur pourrait permettre de caractériser un genre à travers les habitudes de son auditoire dans sa globalité, que nous pourrions résumer ainsi :

— Les gens qui écoutent ce genre de musique écoutent-ils tous les mêmes morceaux ?

— Le succès de ce genre de musique est il le fait de quelques tubes ?

Le ratio R, précédemment calculé, ne tient pas compte du comportement individuel des auditeurs d'un genre. Toutefois, nous pouvons calculer ce ratio pour chaque auditeur i d'un genre, et ensuite calculer la moyenne R_M et la variance σ_R^2 sur les N_a auditeurs :

$$R_M = \frac{1}{N_a}\sum_{i=1}^{N_a} R_i = \frac{1}{N_a}\sum_{i=1}^{N_a} \frac{E_i}{O_i} \tag{1.2}$$

Un ratio moyen R_M élevé indique qu'individuellement, les auditeurs ont tendance à concentrer leurs écoutes de ce genre sur peu de morceaux. La variance, quant à elle, nous informe sur l'homogénéité du comportement des auditeurs d'un genre. Par exemple, le genre pop a un R_M élevé, mais également une forte variance σ_R^2. Il y a donc une tendance globale des auditeurs à écouter peu de morceaux de pop différents, mais on peut difficilement généraliser cette information à l'ensemble des auditeurs, en raison de la forte variance.

TABLE 1.1: Représentation des genres selon le nombre d'occurrences O, d'écoutes totales E, le ratio écoutes/occurrence R, le ratio moyen par utilisateur R_M et variance du ratio σ_R^2.

Genre	O	E	R	R_M	σ_R^2
Pop	46376	542694	11,70	2,29	8,40
Rap/Hip Hop	35913	1272009	35,42	2,26	2,89
Rock	24905	169384	6,80	2,02	3,95
Electro	17998	135641	7,54	2,00	9,30
Dance	17537	205998	11,75	2,13	6,24
Alternative	15668	128410	8,20	1,85	4,21
Films/Jeux vidéo	8873	60750	6,85	2,03	8,15
Musiques de films	8106	55426	6,84	1,99	6,65
R&B	6883	123516	17,95	2,24	5,73
Reggae	4794	34071	7,11	2,01	4,52
Classique	4453	11889	2,67	1,62	3,07
Rap français	4443	233260	52,50	2,31	4,89
Metal	4289	19369	4,52	1,97	5,42
Chanson française	4181	39781	9,51	1,85	4,03
Pop internationale	4029	44802	11,12	2,14	5,43
Jazz	3706	10726	2,89	1,65	6,15
Variété Internationale	3388	24902	7,35	1,87	5,20
Techno/House	3184	30539	9,59	1,90	4,18
Rock indépendant	2775	19820	7,14	1,68	2,34
Latino	2604	34174	13,12	2,23	8,05
Hard Rock	2337	12730	5,45	1,74	2,94
Singer & Songwriter	1979	24071	12,16	1,89	3,99
Soul	1927	12161	6,31	1,77	4,27
Musique africaine	1831	6766	3,70	1,88	7,59
Jeunesse	1588	7839	4,94	1,57	2,08
Rock Indé/Pop Rock	1586	8280	5,22	1,90	4,78
Pop Indépendant	1529	9612	6,29	1,61	2,10
Folk	1282	6426	5,01	1,73	3,98
Disco	1135	4998	4,40	1,70	9,78
Musique arabe	1024	3582	3,50	1,86	3,70

Nous retrouvons dans la table 1.1 un phénomène déjà observé pour l'ensemble des auditeurs sur le genre classique : les auditeurs de ce genre ont écouté en moyenne 1,6 fois le même morceau, et cette tendance est assez solide puisque la variance est relativement basse. Le genre techno/house a cumulé près de 3 fois plus d'écoutes que le genre jazz, en revanche les auditeurs de ce dernier ont globalement et individuellement écouté plus de morceaux différents.

La tendance observée ici rejoint les deux théories sociologiques de Bourdieu

et Peterson évoquées dans la partie 1.2.1. Quand le premier faisait une distinction entre la musique savante (jazz, classique) et la musique populaire (pop, rap/hip-hop), le second considérait un auditoire aux goûts omnivores en opposition à l'univore. Ici, nous voyons que les auditeurs de musiques savantes ont vraisemblablement des goûts plus omnivores que les auditeurs de musiques populaires.

Les ratios présentés ici permettent de caractériser les genres selon l'ouverture de leurs auditoires, et pourraient être un critère à considérer dans la recommandation de musique. Des travaux similaires ont par ailleurs étés réalisés par M. Schedl dans [99] pour décrire un utilisateur selon ces critères : *« Diversity, mainstreamness, and Novelty »*. La diversité, l'appartenance (ou non-appartenance) à un courant dominant et la nouveauté sont en effet des critères de choix qui peuvent s'avérer déterminants pour les auditeurs, et qui transcendent les genres musicaux.

1.4.2 Co-occurence des genres dans les annotations

Les morceaux sont souvent annotés avec plusieurs genres différents. Il peut donc s'avérer utile d'observer quels genres sont régulièrement annotés conjointement. Nous pouvons notamment construire une matrice de co-occurrence : la valeur (i, j) de cette matrice nous indique combien de fois le genre numéro i a été annoté conjointement avec le genre j. Nous avons construit ette matrice à partir de tous les morceaux différents observés dans le log de Deezer. Elle est carrée, symétrique, et sa diagonale indique combien de fois chaque genre a été observé au total (voir figure 1.3). Pour l'affichage de cette matrice, une échelle logarithmique a été utilisée en raison des forts écarts de valeur entre les différents genres.

Pour des raisons de lisibilité, seulement une partie de la matrice est montrée ici. Nous voyons des co-occurrences triviales, telles que Rap/Hip Hop avec Rap français, ou bien encore Musiques de films apparaît souvent avec Films/Jeux vidéo. Nous pouvons également remarquer que certains genres ont tendance à être affiliés avec un grand nombre d'autres genres : Pop, Rock, Pop internationnale. Le fait que Pop soit affilié avec un grand nombre de genres différents peut être une explication de la valeur de σ_R^2, qui nous indiquait une hétérogénéité du comportement de ses auditeurs.

Pour représenter les similarités des genres selon leurs annotations, cette matrice présente 2 défauts :

— Comme dit précédemment, cette matrice est symétrique. Pourtant, certaines associations ne sont pas réciproques. Par exemple, les morceaux tagués comme Rap français sont très souvent également taggés comme Rap/Hip hop, mais l'inverse n'est pas vrai, car ne nombreux morceaux de rap ne sont pas français.

— Ici est représenté le **nombre** de fois où nous avons observé deux genres conjointement, et non pas la proportion de cette association parmi l'ensemble des apparitions du genre. Les valeurs de cette matrice peuvent

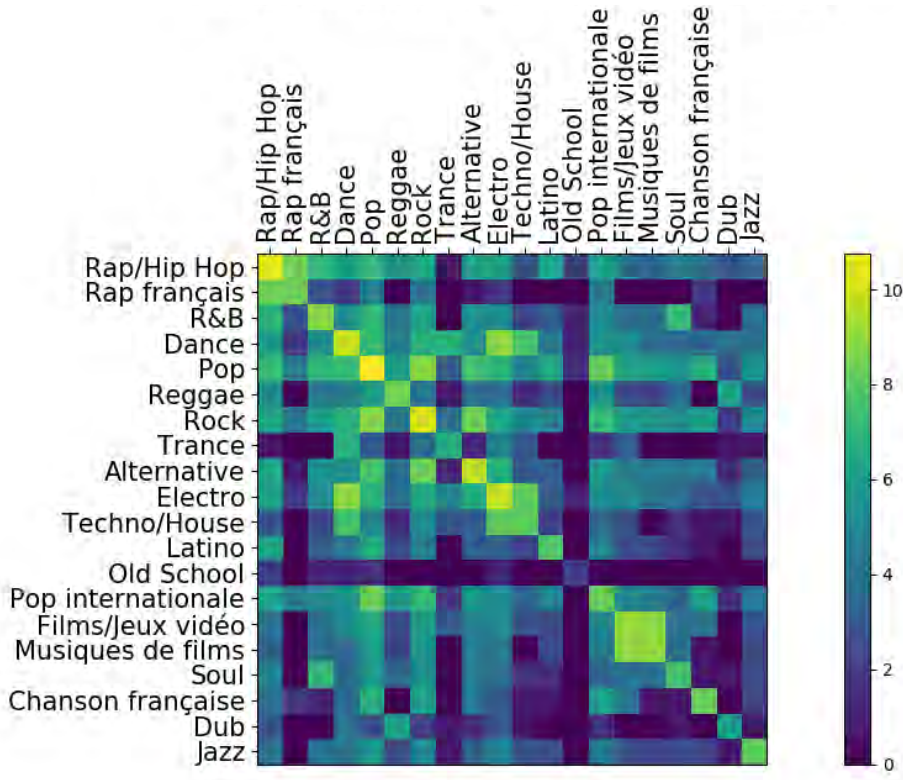

FIGURE 1.3: Co-occurrence des genres dans les annotations (échelle logarithmique, 20 genres).

être donc influencées par le fait que certains genres soient largement surreprésentés dans le log.

Pour rendre cette matrice asymétrique, et pour obtenir une **proportion** d'associations plutôt qu'un nombre, nous divisons chaque ligne de cette matrice par sa somme. Nous obtenons ainsi la matrice asymétrique visible en figure 1.4. Sur cette figure, chaque ligne correspond donc à un genre, et chaque colonne au nombre de fois qu'il a été annoté conjointement à un autre genre, normalisé par le nombre total d'annotations.

Nous voyons que le Rap français est souvent associé au Rap/Hip hop dans les annotations, mais que l'inverse est moins vrai. De même, le Dub est souvent annoté conjointement au Reggae, puisqu'il s'agit d'un sous-genre de ce dernier, mais le Reggae, en proportion, apparaît souvent annoté sans le Dub. Les fortes valeurs dans les cases jaunes (en haut à gauche et en bas à droite de la matrice) nous indiquent que, en proportion, l'annotation Rap/Hip hop apparaît souvent seule, tout comme Jazz (en bas, à droite).

Cette matrice de similarité permet donc de faire apparaître des interdépen-

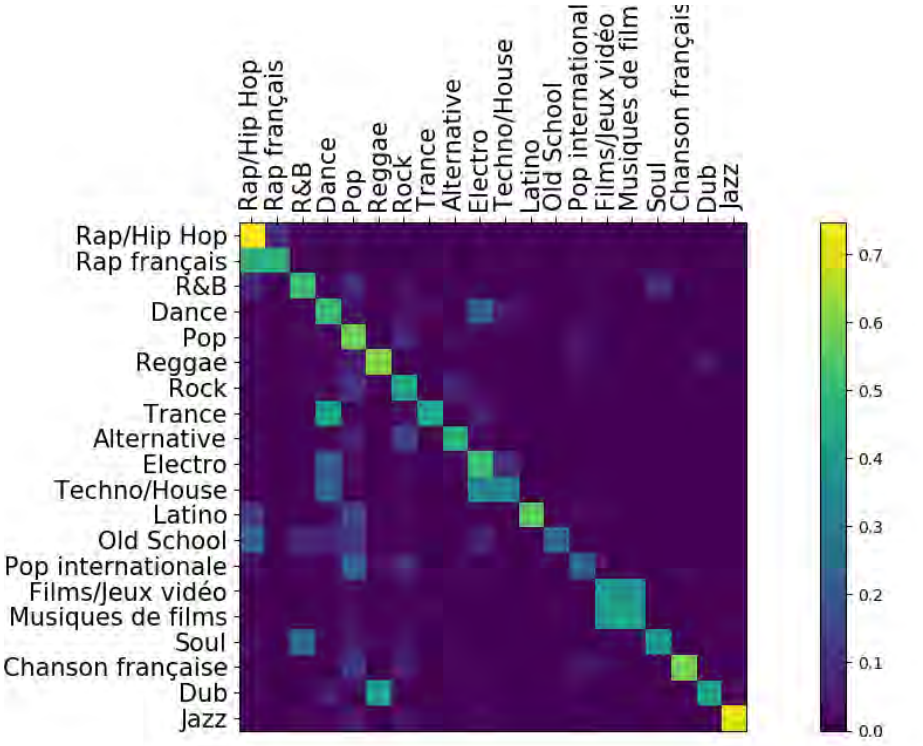

FIGURE 1.4: Co-occurrence des genres dans les annotations (20 genres), distance asymétrique.

dances dans les annotations en genre : des genres parents, comme rap/hip-hop et reggae donnent naissance à des genres enfants, comme rap français ou dub, associés généralement à leurs parents dans les annotations.

1.4.3 Dendrogramme selon les annotations

Chaque ligne de la matrice précédemment calculée peut à présent être considérée comme un ensemble de paramètres décrivant chaque genre, selon la fréquence de ses annotations avec d'autres genres. Cet ensemble de paramètres peut notamment nous permettre d'établir une distance entre 2 genres.

Ces distances nous permettent à présent d'établir une classification ascendante hiérarchique selon la méthode de Ward [120], décrite ci-dessous.

À l'initialisation de l'algorithme, chaque genre de musique i forme une classe indépendante, de centre de gravité g_i.

À chaque itération, nous procédons à une fusion entre 2 classes, de manière à maximiser l'inertie inter-classes I_e (avec n le nombre d'individus total, n_i le nombre d'individus par classe, et g le centre de gravité de l'ensemble des individus) :

$$I_e = \frac{1}{n} \sum_{i=1}^{k} n_i \times d^2(g_i, g) \tag{1.3}$$

Ici d^2 désigne le carré de la distance entre 2 genres de musique. Nous avons choisi d'utiliser une distance euclidienne dans l'espace des paramètres formés à l'aide des annotations. D'autres mesures de distances pourraient être utilisés, comme la similarité cosinus, ou bien la distance de manhattan. Les différentes mesures de distances testées n'ont pas produit de différence significative dans les résultats obtenus.

L'algorithme se poursuit jusqu'à la fusion totale de toutes les classes, et le résultat peut se représenter sous forme d'un arbre (appelé dendrogramme, voir figure 1.5), où l'axe des ordonnées nous indique la distance entre 2 classes : plus 2 branches sont regroupées bas, plus ces classes sont proches.

Nous pouvons définir un seuil de distance à partir duquel les classes sont séparées. Par exemple, en prenant un seuil de 1, nous obtenons 13 classes distinctes. Cela peut être utile pour réduire le nombre de genres à quelques catégories, mais l'intérêt majeur de cette représentation repose dans les différents niveaux de hiérarchies entre les classes, et la représentation des distances entre ces classes. Ici, nous voyons que les genres de musiques Classique sont ceux les plus éloignés des autres, juste derrière les genres affiliés à Rap/Hip hop. Pour rappel, les distances considérées ici reposent sur l'annotation conjointe en genre des morceaux.

Des méthodes dites « basées sur le contenu » (contenu ici au sens « métadonnées ») pourraient s'inspirer de cette méthode de classification : des morceaux appartenant à 2 genres distincts peuvent donc être considérés comme similaires, car appartenant à 2 catégories souvent regroupées.

1.4.4 Dendrogramme selon les écoutes

La même méthode a été employée en s'appuyant cette fois sur la co-occurrence de genres dans les écoutes d'un même auditeur. La matrice de co-occurrence présente donc dans sa case (i, j) le nombre d'occurences des morceaux du genre i qui ont été écoutés par un auditeur ayant également écouté un morceau du genre j. De la même manière que précédemment, nous pouvons ensuite construire une matrice asymétrique, afin d'obtenir une information en proportion et non plus en nombre.

Dans le dendrogramme obtenu (voir figure 1.6), des genres et des groupes de genres proches ont donc été écoutés souvent par les mêmes personnes. Nous pouvons observer que certaines branches de l'arbre précédent ont été séparées. Par exemple, une branche du dendrogramme précédent regroupait les genres Reggae, Dub, Dancehall/ragga, et Ska. Bien que le Ska soit proche de ces 3 autres genres, nous voyons que les auditeurs ayant l'habitude d'en écouter ont plutôt tendance à écouter également des styles proches du Rock et du Blues.

Une certaine prudence doit être observée quant à l'analyse de cet arbre, puisqu'il est issu de la normalisation de la matrice de co-occurrence. Ainsi, nous

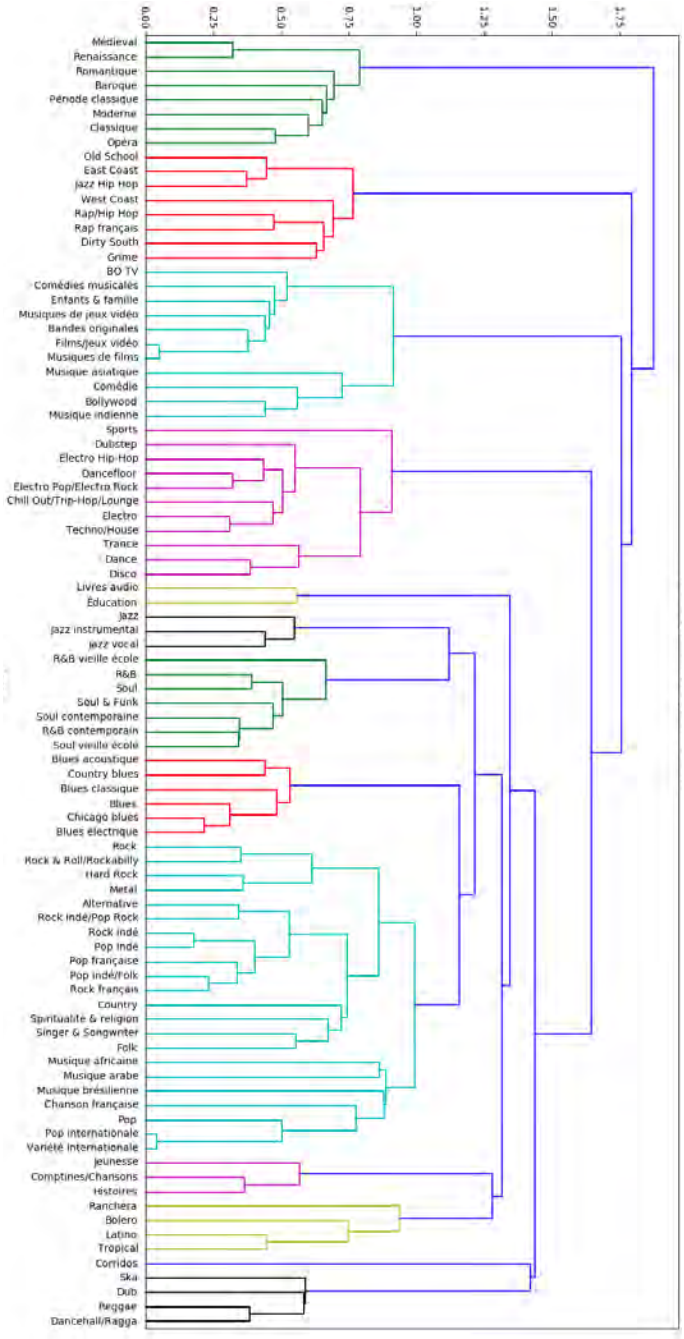

FIGURE 1.5: Dendrogramme construit à partir des annotations conjointes.

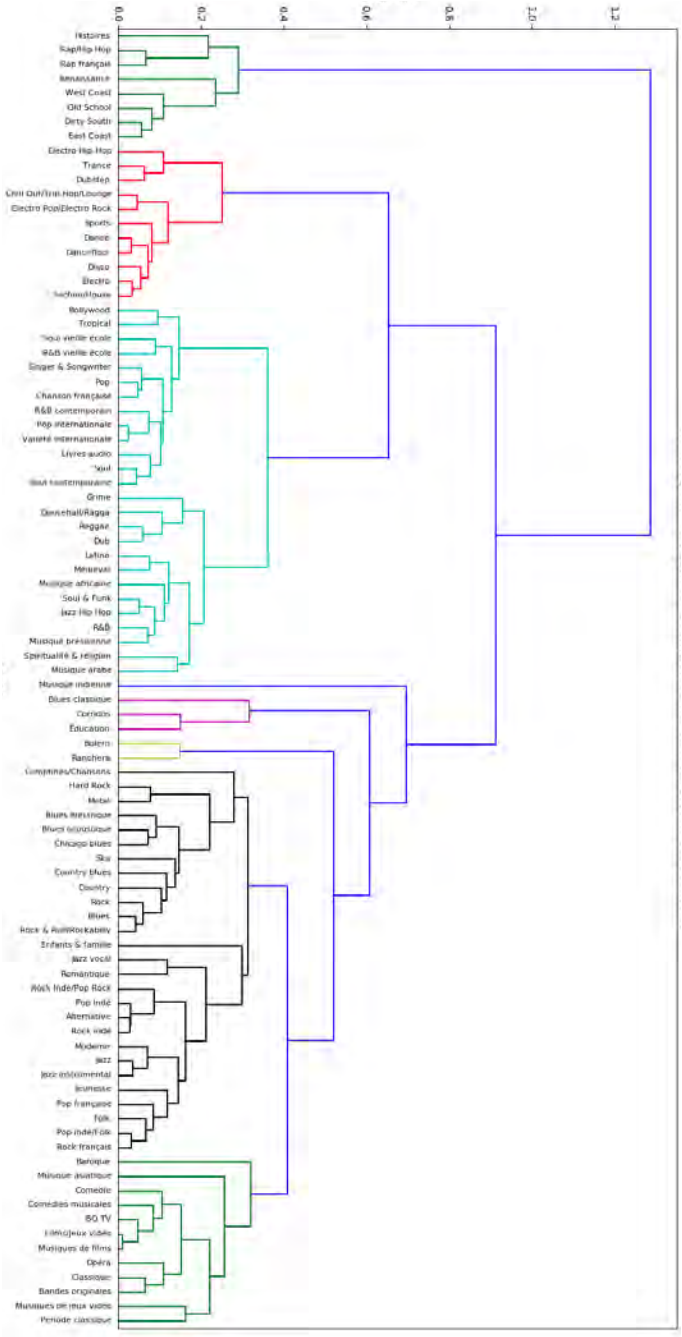

FIGURE 1.6: Dendrogramme construit à partir des écoutes des utilisateurs.

voyons que le genre renaissance est associé au hip-hop dit « west coast ». Cela est dû au fait que le peu d'auditeurs qui ont écouté des morceaux appartenant à la période de la Renaissance ont également écouté du hip hop west coast. Nous pouvons supposer qu'en nous concentrant sur un échantillon plus élevé d'auditeurs de musique de la Renaissance, nous pourrions obtenir d'autres associations.

Toutefois, nous pouvons observer que les genres populaires Rap/Hip hop et Rap français appartiennent à un groupe qui est à part de l'ensemble des autres genres. Une fois de plus, ces observations sont en adéquation avec la double dualité théorique abordée précédemment : musique savante/musique populaire et goûts univores/omnivores.

Cette approche dans le regroupement des genres peut s'approcher du filtrage collaboratif (abordé dans la partie 2.3 du chapitre suivant) : 2 morceaux appartenant à des genres distincts peuvent être considérés comme proches, car appartenant à des genres écoutés de manière conjointe par un certain nombre d'utilisateurs.

1.4.5 Influence du contexte temporel sur le comportement de l'utilisateur

Nous avons observé les heures et jours d'écoutes des utilisateurs afin d'étudier l'influence du contexte temporel sur l'écoute de musique.

L'histogramme de la figure 1.7 nous montre la concentration d'écoutes selon l'heure de la journée, moyennée sur la semaine d'observations. Nous constatons naturellement un creux durant la nuit, un pic d'activité aux alentours de 17h, ainsi qu'un deuxième pic aux alentours de 8h.

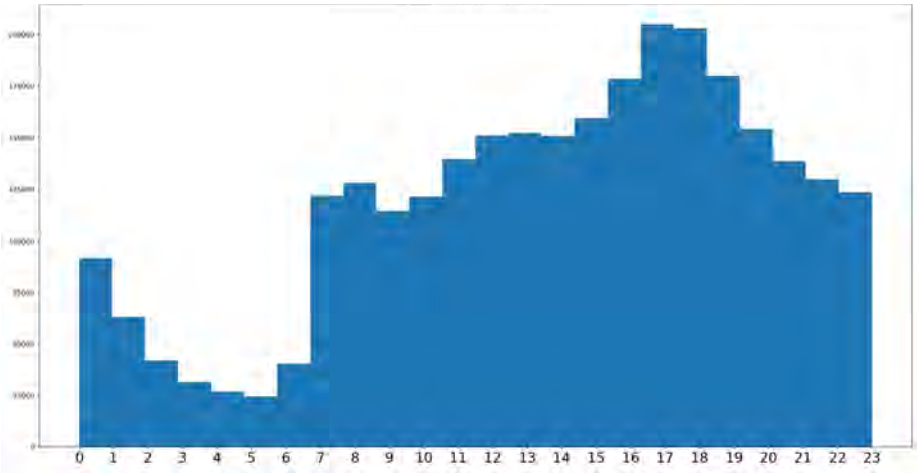

FIGURE 1.7: Nombre d'écoutes en fonction de l'heure de la journée.

Quand nous affichons la durée moyenne d'écoute en fonction de l'heure (voir

figure 1.8), nous obtenons le résultat inverse, i.e. un fort pic durant la nuit, et deux petits pics durant des « horaires de travail », aux alentours de 10h et 15h.

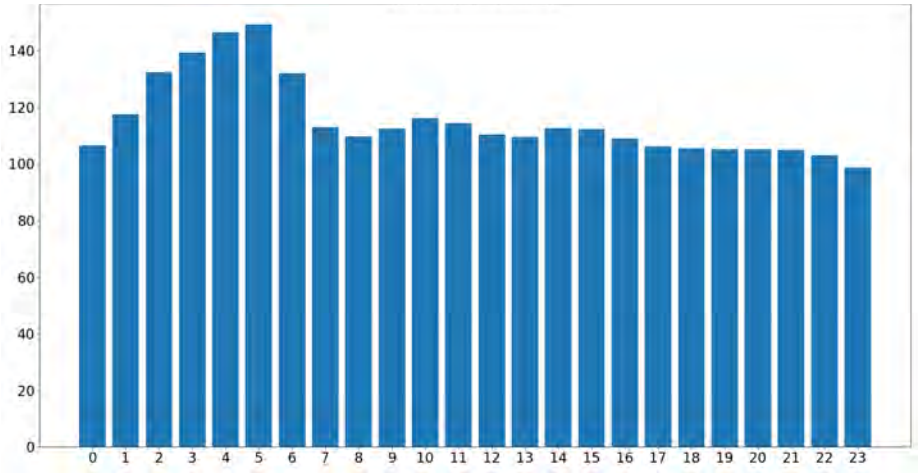

FIGURE 1.8: Durée moyenne en secondes d'écoute selon l'heure de la journée.

La figure 1.9 présente la concentration des écoutes tout au long de la semaine. Nous observons bien la répétition d'un motif régulier au fil des jours, avec un grand creux chaque nuit.

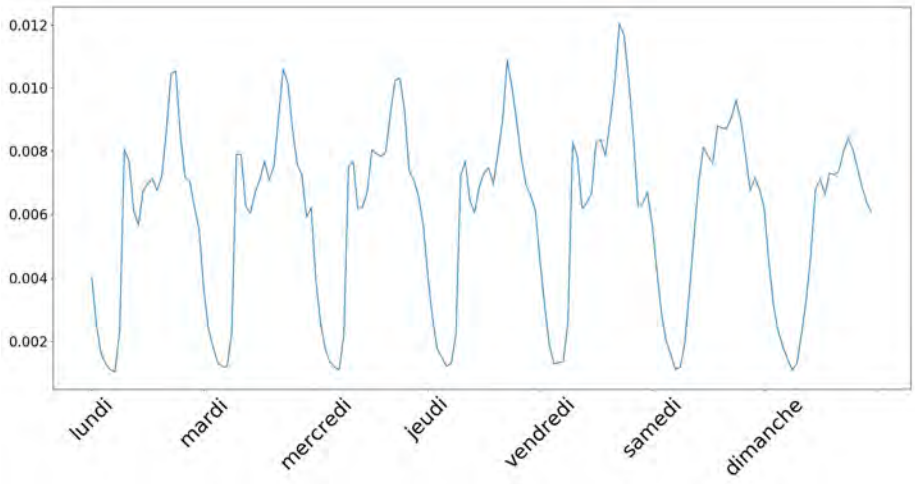

FIGURE 1.9: Densité des écoutes durant la semaine.

En allant plus dans le détail, nous voyons 2 pics par jours le matin et l'après-midi, particulièrement présents du lundi au vendredi. Ces pics correspondent probablement à une écoute dans les transports. Par ailleurs, nous observons un

pic particulièrement important le vendredi soir qui pourrait correspondre à un trajet plus long, comme un départ en week-end. Ce rythme observable du lundi au vendredi est moins observable le samedi et le dimanche, où la courbe est plus lissée. Par ailleurs, on observe des pics le vendredi et le samedi soir, peu avant minuit, qui pourraient correspondre à des écoutes dans un contexte festif.

Ces observations nous montrent donc que la quantité d'écoute varie selon l'heure et le jour de la semaine. Ces variations sont *a priori* dues à des activités (travail, transport, fête, etc.) pratiquées en même temps par une grande partie de l'échantillon observé.

1.4.6 Influence du contexte temporel sur les genres de musique

Nous avons observé différents pics d'écoutes qui d'après nos hypothèses proviennent de plusieurs contextes différents. D'après les travaux étudiés dans la littérature (voir partie 1.2.2) les genres écoutés peuvent dépendre de l'activité contextuelle à l'écoute. Nous avons voulu savoir si cette hypothèse était validée dans nos données. Nous avons donc étudié les genres écoutés par l'ensemble des utilisateurs observés sur différentes plages horaires. Pour chaque genre, nous avons donc tracé le nombre d'écoutes par heure, normalisé par le total d'écoutes cumulées par ce genre sur la semaine. Excepté pour le genre « Trance » qui semble avoir été écouté en grande majorité le lundi soir et le mardi soir, la plupart des concentrations d'écoutes sont fortement similaires entre les genres, et suivent le motif observé précédemment (voir figure 1.10).

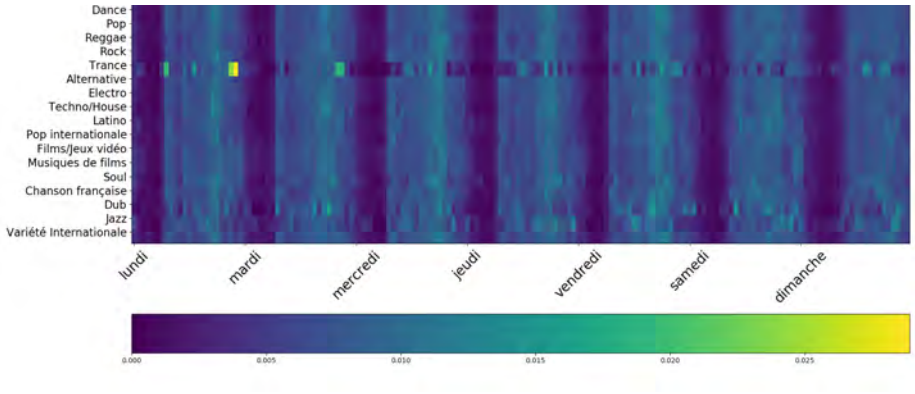

FIGURE 1.10: Concentration d'écoutes durant la semaine, selon le genre considéré.

Nous avons calculé la moyenne de ces concentrations sur tous les genres puis pour chaque genre nous avons observé sa déviation à la moyenne : voir figure 1.11.

Cette déviation à la « concentration moyenne d'écoute » peut également permettre de caractériser des genres et d'établir des similitudes : deux genres

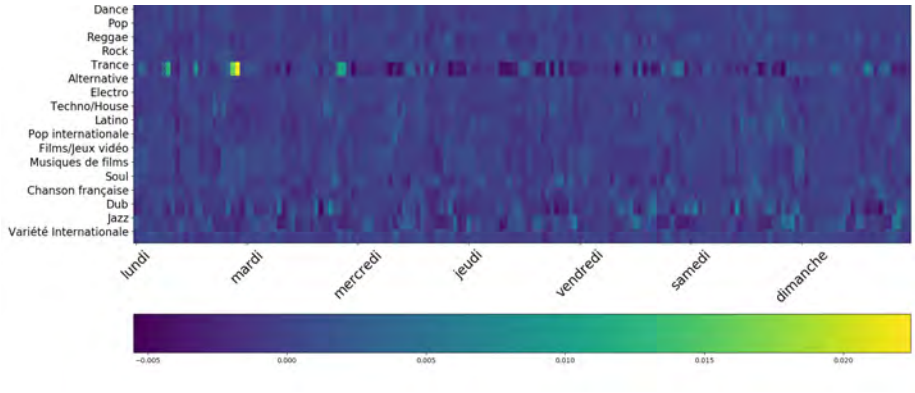

FIGURE 1.11: Écart de la concentration d'écoute par rapport à la moyenne, selon le genre considéré.

qui sont plus écoutés que la moyenne sur les mêmes plages horaires comportent probablement des similitudes musicales ou bien d'*usage* (musique pour travailler, se reposer, faire la fête...)

Sur ces figures rien ne semble indiquer qu'il existe une tendance globale sur des moments d'écoutes privilégiés selon les genres.

Nous avons constaté précédemment des pics d'écoutes à des moments distincts qui correspondent probablement à des activités pratiquées de manière synchrone par un grand nombre d'individus de notre échantillon. Si les transports ou la fête sont des moments à l'écoute de musique, rien n'indique que tout le monde écoute les mêmes genres de musique dans ces moments-là : les genres destinés au transport pour certains sont peut-être des musiques de fête pour d'autres. Pour aller plus loin, une étude des comportements individuels des utilisateurs peut révéler des motifs réguliers dans l'apparition de certains genres (voir partie 1.2.2).

Nous avons analysé les données de Deezer afin de de tester l'hypothèse suivante : « Individuellement, les genres écoutés diffèrent-ils selon le moment de la semaine ? ». En particulier, nous avons voulu vérifier si nous pouvions établir une dualité semaine/week-end, journée/soirée ou bien encore travail/repos.

Pour tester la première hypothèse, nous avons calculé pour chaque utilisateur 2 histogrammes : les genres écoutés durant la semaine (de lundi à vendredi) et les genres écoutés pendant le week-end (samedi et dimanche). Ces histogrammes donnent pour chaque genre le nombre d'écoutes de morceaux par l'utilisateur, sur la plage temporelle considérée. Ces histogrammes ont été normalisés en divisant chaque valeur par le nombre total de morceaux écoutés sur la période. Ainsi, nous ne retenons que la proportion d'écoutes entre les différents genres, en supprimant l'impact de la durée de la plage temporelle observée.

Pour chaque utilisateur, nous avons ensuite calculé l'intersection I entre ces deux histogrammes comme ce qu'il suit :

$$I(h_1, h_2) = \sum_{i=1}^{N} min[h_1(i), h_2(i)] \tag{1.4}$$

Du fait de la normalisation des histogrammes, leur intersection peut prendre des valeurs entre 0 et 1 : à 0, l'intersection entre les deux histogrammes est nulle et à 1 (ou 100%), les deux histogrammes sont identiques.

Sur l'ensemble des utilisateurs, nous avons observé une intersection moyenne de 76%, avec une variance de 0,04. La distinction entre semaine et week-end ne semble donc pas révéler de différence flagrante dans l'écoute des genres.

Sur la figure 1.12, nous voyons que l'utilisateur a écouté les mêmes genres durant la semaine et le week-end. Ce cas, avec une intersection forte de 87%, est un archétype des histogrammes observés pour une majorité des utilisateurs. Cependant, certains utilisateurs ont écouté des genres différents entre la semaine et le week-end.

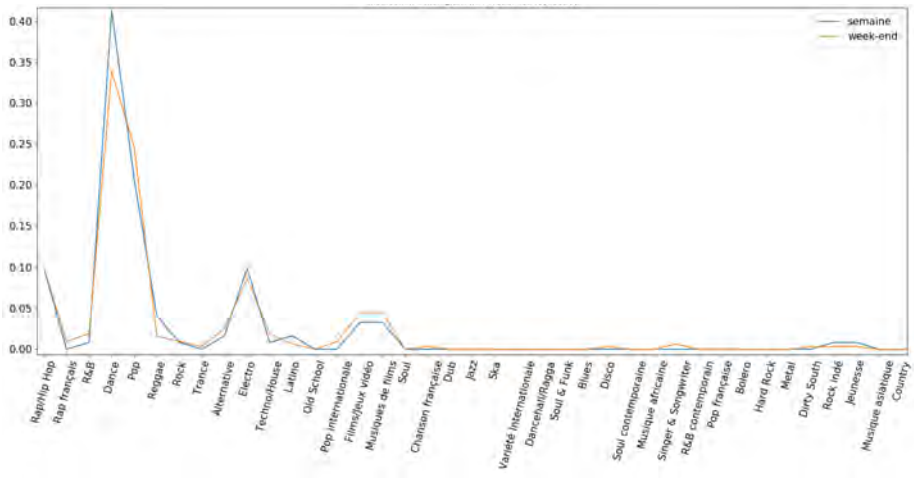

FIGURE 1.12: Exemple d'intersection forte des histogrammes de genre.

En figure 1.13 sont affichés les deux histogrammes d'un utilisateur n'ayant que 32% d'intersection entre la semaine et le week-end. Nous voyons que cet utilisateur a totalement délaissé le Rock, pourtant écouté très majoritairement la semaine, pour écouter du Rap/Hip Hop le week-end, de manière quasi exclusive.

Remarque : les histogrammes ont ici été représentés sous forme de courbes plutôt que de barres, afin de mieux distinguer les intersections.

Par ailleurs, nous avons également mesuré les intersections d'histogrammes de genre entre le jour (6h - 18h) et la nuit (18h - 6h). Nous avons obtenu une intersection moyenne de 75% avec une variance de 0,039.

De même, nous avons défini des horaires de travail, de 9h à 12h et de 14h à 18h du lundi au vendredi, afin de les comparer avec le reste de la semaine.

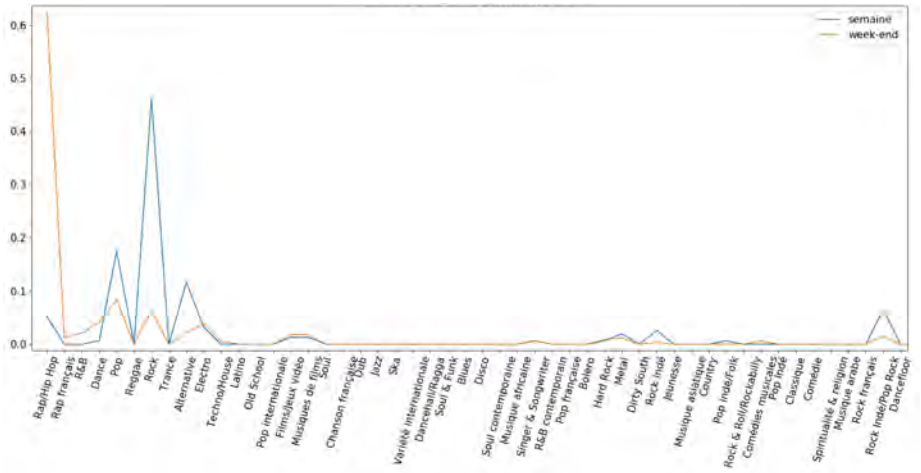

FIGURE 1.13: Exemple d'intersection faible des histogrammes de genre.

Comme précédemment, l'intersection des histogrammes est globalement forte, avec une moyenne de 75% et une variance de 0,033.

Dans l'ensemble, le moment de la semaine semble ne pas avoir d'influence sur les genres écoutés. Toutefois, il y a bien certains utilisateurs, comme le montre l'exemple de la figure 1.13, pour lesquels cette hypothèse peut être validée. Pour ces derniers, des algorithmes de recommandation tels que ceux présentés dans la partie 1.2.2 pourraient s'avérer particulièrement pertinents.

1.5 Uniformisation des goûts ?

Après avoir étudié les différents contextes ayant une influence sur l'écoute de certains genres, nous allons nous intéresser à la concentration des écoutes autour de certains morceaux, et au rôle des algorithmes de recommandation.

1.5.1 « The Long Tail », concentration des écoutes

Le phénomène dit de « Long Tail » (longue traîne en français) est apparu avec l'essor de la numérisation de la production et de la distribution de musique, qui en a drastiquement réduit les coûts en comparaison aux supports physiques : le stockage d'un album sur un serveur coûte bien moins cher que dans une boutique de disques. Ces réductions de coûts ont donc considérablement augmenté la quantité de musique disponible pour les auditeurs (voir figure 1.14).

Dans un article [3] paru en 2004 puis dans un livre [2] en 2006, Chris Anderson introduisait la théorie de la « Long Tail ». Selon cette théorie, l'augmentation de la taille des catalogues des plateformes allait être accompagnée d'un accès favorisé par internet à des niches musicales jusqu'alors très peu explorées. Ainsi, la

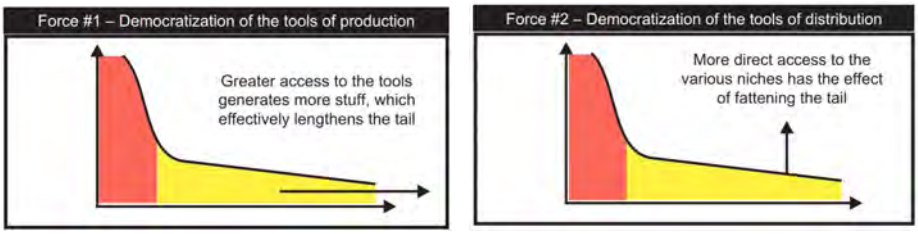

FIGURE 1.14: Les deux mécanismes de la « Long Tail ». Source : C. Anderson, The Long Tail, 2006 [2].

décennie passée aurait dû voir la distribution des écoutes par morceaux s'étaler davantage, en rendant le marché des musiques de niche davantage attractif.

Malheureusement, il a été montré par Mark Mulligan en 2014 [76], et dans d'autres études plus récentes [18, 30] que ce phénomène attendu n'a pas eu lieu dans l'industrie musicale : les parts de marché de la musique de niche ont plutôt eu tendance à diminuer. Les différentes pistes données pour expliquer ce phénomène sont les suivantes :

— La taille des catalogues des plateformes de musique en ligne est en constante augmentation avec des milliers de sorties quotidiennes. Plus la base de données à explorer est grande, plus l'algorithme de recommandation chargé de trouver des morceaux pertinents parmi ces données doit être performant. Les algorithmes de recommandation doivent ainsi s'améliorer aussi vite que la taille des catalogues augmente. De plus, les algorithmes étant souvent basés sur des algorithmes dits de *filtrage collaboratif*, ils sont particulièrement sensibles au problème du « cold start », i.e. lorsque des morceaux n'ont jamais été écoutés, ils ne peuvent pas être recommandés. Ce phénomène sera décrit plus en détail dans la partie 2.3. Ces problématiques liées au « Cold start » et à la « Long Tail » sont des thématiques de recherches récurrentes dans le domaine de la recommandation [124, 25, 64].

— La « Tyrany of choice » ou « l'excès de choix » : les faibles performances des algorithmes de recommandation incitent les auditeurs à se tourner vers des choses qu'ils connaissent déjà. En effet, il est montré dans [18] que choisir parmi un nombre trop élevé de possibilités représente un effort cognitif trop important pour les utilisateurs, qui préfèrent alors ne pas choisir, quitte à renoncer à la découverte.

— La « Dilution de la qualité » : il est moins risqué financièrement pour un artiste ou une maison de disque de diffuser sa musique sur les plateformes en ligne plutôt que via une sortie physique. Le moindre risque rend ainsi moins exigeante la distribution des morceaux vis-à-vis de leur qualité. Par conséquent, une grande quantité des nouveaux morceaux arrivant chaque jour ne sont donc pas écoutés, car ils n'ont pas la qualité attendue par les auditeurs.

Nous avons analysé le log mis à disposition par Deezer (voir partie 1.3.1) afin

d'observer la répartition des écoutes selon les morceaux. Sur cette période, plus de 200.000 morceaux différents ont été écoutés par 10.000 utilisateurs, pour un total de près de 3 millions d'écoutes au total, soit plus de 14 écoutes par titre en moyenne. Les 100 morceaux les plus écoutés ont accumulé plus de 375.000 écoutes, soit plus de 13% du total. Les 1000 morceaux (soit seulement 0,00179 % du catalogue de Deezer !) les plus écoutés ont accumulé plus d'un million d'écoutes, soit plus de 38% des écoutes. Le morceau le plus écouté a accumulé 16820 écoutes soit 0,58%. Enfin, les 10% et 20% de morceaux les plus écoutés ont accumulé respectivement 80% et 90% des écoutes. Nous observons donc une forte concentration des écoutes par une infime portion du catalogue disponible. Cette concentration est d'autant plus forte en comptant par auditeur : les morceaux du top 10 ont été écoutés en moyenne au moins une fois par 1/4 des utilisateurs.

Cette concentration de l'audience, appelée « Loi de Pareto » [83], a évidemment des conséquences économiques : il est d'ailleurs qualifié de « superstar music economy » par Mulligan. Comme nous le voyons sur la figure 1.15, une très grande part des revenus des ventes de musique n'est redistribuée qu'à une infime partie des artistes.

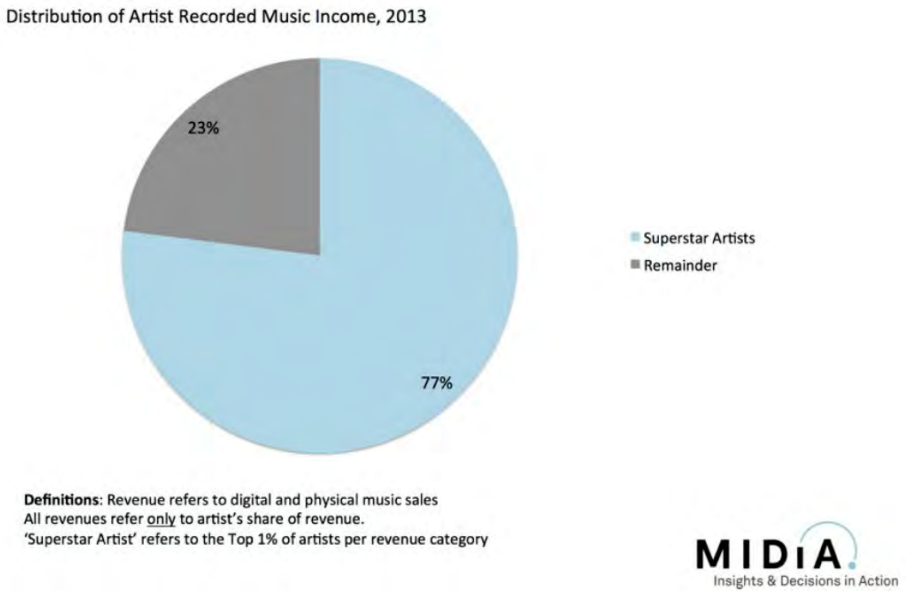

FIGURE 1.15: Répartition des revenus des ventes de musique en 2013 entre les artistes (Source : M. Mulligan The death of the long tail : The superstar music economy, 2014 [76]).

Par ailleurs, si une petite partie du catalogue des plateformes accumule la majorité des écoutes et donc des revenus, une partie non négligeable de la musique hébergée par les plateformes n'a jamais été écoutée. Ce phénomène a pris une telle ampleur que certaines plateformes ont pris la décision de retirer de

leurs plateformes certains titres afin de réduire leurs frais de stockage[5].

Une autre solution pour contrer ce phénomène repose bien entendu sur l'amélioration des systèmes de recommandation, qui permettrait de proposer à un public plus large ce contenu peu ou pas consommé.

1.5.2 Concentration des genres

Nous avons vu précédemment qu'un faible nombre de morceaux accumulent une très forte proportion d'écoutes. Nous allons à présent étudier la répartition des écoutes selon les différents genres. Les genres considérés ici ont été attribués par Deezer aux morceaux de leur bibliothèque. Ils ne sont pas directement visibles par l'utilisateur, mais sont accessibles via l'API. Plusieurs genres peuvent être affectés au même morceau.

Sur l'ensemble des morceaux écoutés par les 10.000 auditeurs durant une semaine, nous observons 90 genres différents avec une forte disparité. Si nous analysons les 1000 morceaux les plus écoutés, le nombre de genres différents est réduit à 30. Si nous observons les 100 morceaux les plus écoutés, ce nombre est réduit à 13.

Dans les tableaux 1.2 et 1.3 sont listés les genres avec lesquels ont été tagués les morceaux du top 100, c'est-à-dire les 100 morceaux les plus écoutés. Nous observons que la moitié de ces morceaux ont été annotés comme Rap/Hip hop, et que ces morceaux cumulent près de la moitié des écoutes du top 100.

TABLE 1.2: Nombre d'occurrences des différents genres dans les morceaux du top 100.

Genre	Occurences
Rap/Hip Hop	51
Pop	18
Dance	8
R&B	6
Rap français	5
Electro	2
Latino	2
Reggae	2
Techno/House	2
Chanson française	1
Singer & Songwriter	1
Pop internationale	1
Rock	1

Nous remarquons également une concentration des écoutes au niveau des genres. Malgré la centaine de genres disponibles sur la plateforme, le genre

5. https://www.traxmag.com/beatport-va-supprimer-tous-les-tracks-invendus-de-sa-plateforme/

TABLE 1.3: Nombre d'écoutes du top 100 selon le genre.

Genre	Ecoutes	%
Rap/Hip Hop	179709	48,34
Pop	74656	20,08
Dance	28235	7,59
R&B	27671	7,44
Rap français	22914	6,16
Pop internationale	7804	2,10
Rock	7804	2,10
Electro	7226	1,94
Singer & Songwriter	4548	1,22
Chanson française	3099	0,83
Techno/House	3018	0,81
Reggae	2748	0,74
Latino	2347	0,63

de musique rap/hip-hop accumule une très forte proportion des écoutes, et est omniprésent dans la liste des morceaux les plus écoutés. Par ailleurs, cette même étude avec les 1000 morceaux les plus écoutés et sur l'ensemble des morceaux écoutés nous permet de dresser le même constat (voir tables B.3 et B.4 en annexe).

Par ailleurs, les musicologues Antoine Vervier et Stéphane Escoubet ont mené une analyse sur ces données ainsi que sur le top 100 2018 du site Deezer, référençant les 100 morceaux les plus écoutés par l'ensemble des utilisateurs durant l'année 2018[6]. Le constat dressé est similaire pour les deux tops, avec une très forte prédominance des influences Hip hop et Pop. Au-delà des simples tags de genres fournis par l'API, l'analyse musicologique a donc confirmé l'homogénéité des morceaux les plus écoutés.

De plus, nous avons constaté un fort recouvrement entre les données dont nous disposions et le top 2018 : 59 morceaux figurent dans les deux listes. Ainsi, nous pouvons considérer que notre échantillon est représentatif des habitudes d'écoutes des utilisateurs.

1.5.3 Part de l'utilisation des outils de recommandation automatique

En 2019, J.S. Beuscart et al. ont publié les résultats de leur analyse sur des données provenant du site Deezer [14]. Sur les données observées datant de 2014, 75% des écoutes étaient d'origine autonome, 10% provenaient de recommandations personnalisées automatiquement, et 10% de recommandation dite traditionnelle (playlist éditoriale ou classement).

6. https://www.deezer-blog.com/fr/2018-le-bilan-musical-decouvrez-le-top-de-lannee/

Cette faible utilisation peut avoir diverses origines : un manque de confiance ou une insatisfaction des utilisateurs vis-à-vis des algorithmes, une volonté de choisir la musique écoutée ou bien un manque d'intérêt vis-à-vis de la découverte. Des algorithmes plus performants et plus transparents pourraient donc amener les utilisateurs des plateformes de streaming à utiliser davantage la recommandation automatique.

1.6 Conclusion

Dans cette partie, nous avons étudié l'influence des différents contextes sur les genres écoutés. Les contextes peuvent être soit à long terme comme le contexte socio-culturel de l'auditeur ou bien à court terme, comme une activité parallèle à l'écoute. Ces contextes sont montrés dans la littérature comme ayant une influence sur le goût des auditeurs, mais une étude des données fournies par Deezer nous a montré que les goûts allaient au-delà d'une simple distinction entre les genres. Le décompte des morceaux écoutés dans le log de Deezer nous montre que quelques morceaux de quelques genres de musique accaparent une grande majorité des écoutes. La diversité des morceaux écoutés selon les genres semble corroborer les théories sociologiques. Pour les genres associés aux musiques savantes, tels que le jazz et le classique, nous observons une plus grande dispersion des écoutes sur des morceaux et des artistes différents. Au contraire, sur les genres les plus populaires tels que pop, rap/hip-hop et rap français, nous remarquons une forte concentration des écoutes sur un nombre réduit de morceaux. Ces observations de diversité concernent aussi bien l'ensemble des auditeurs d'un genre donné, qu'une moyenne par auditeur. La part d'utilisation des algorithmes de recommandation de musique est encore marginale, l'amélioration de ces derniers pourrait mener à une plus grande diversité dans les horizons musicaux des auditeurs.

Dans le chapitre suivant, nous allons nous intéresser plus particulièrement aux algorithmes de recommandation musicale.

Chapitre 2

La recommandation de musique est-elle une prédiction de goûts ?

Nous venons de voir dans le chapitre précédent les différents facteurs pouvant influencer les goûts musicaux, et les liens tout relatifs que ces derniers pourraient avoir avec les genres. Nous avons constaté une grande concentration des écoutes des utilisateurs des plateformes de streaming autour d'un nombre très réduit de morceaux au regard de la taille de leur catalogue. Dans ce contexte, les algorithmes de recommandation musicale font donc partie des solutions plébiscitées pour amener davantage de diversité tout en respectant les goûts des auditeurs.

Dans ce chapitre, nous allons nous intéresser aux différents algorithmes de recommandation musicale, afin de déterminer à quel point cette tâche est comparable à une prédiction de goût.

Les méthodes employées relevant de l'apprentissage automatique, nous introduirons tout d'abord cette notion et les algorithmes les plus usités. Nous nous intéresserons ensuite aux données utilisées en recommandation de musique : à quel point reflètent-elles le goût des utilisateurs ? Viendra ensuite un état de l'art sur les différentes méthodes de recommandation de musique, en particulier sur les méthodes dites de filtrage collaboratif et les méthodes fondées sur le contenu. Ces dernières étant basées sur des paramètres acoustiques, nous étudierons le lien entre la perception des auditeurs et ces paramètres, afin de questionner la pertinence et les limites de leur utilisation. Enfin, nous verrons sur quelles bases les différents algorithmes de recommandation sont testés et évalués, afin d'en mesurer le lien avec le goût des utilisateurs.

2.1 L'apprentissage automatique

Dans cette section, nous allons présenter les concepts clés de l'apprentissage automatique ainsi que les méthodes les plus utilisées dans ce domaine. Un état de l'art détaillé ainsi qu'une comparaison des différents algorithmes ont été réalisés notamment dans [122] et [104].

2.1.1 Qu'est-ce que l'apprentissage automatique ?

L'apprentissage automatique (ou Machine Learning) désigne un ensemble de méthodes dont le but est d'estimer une fonction à partir de données. Si cette fonction prédit une variable quantitative, alors la fonction estimée est une fonction de régression. Si la fonction prédit une variable qualitative alors il s'agit d'une fonction de classification. Lorsqu'une méthode s'appuie sur une annotation *a priori* des données il s'agit d'une méthode dite *supervisée*. Dans le cas contraire, il s'agit de classification *non-supervisée*.

Dans la plupart des techniques d'apprentissage automatique, des paramètres sont calculés au préalable sur les signaux à traiter. Par exemple, un signal audio comportant des centaines de milliers de points se verra représenté par un nombre plus réduit de paramètres sous forme d'un vecteur. Pour obtenir de bons résultats, le choix doit se porter sur les paramètres les plus discriminants possibles vis-à-vis de la tâche à résoudre. L'extraction de ces paramètres constitue un domaine à part entière qui sera abordé plus en détail ultérieurement (section 2.5.2).

2.1.2 Apprentissage non-supervisé

Les méthodes d'apprentissage non-supervisées ont pour but d'identifier des liens et des structures dans un jeu de données, soit à des fins de regroupement (ou *clustering*) soit à des fins de réduction de dimension. Parmi le grand nombre de méthodes d'apprentissage non-supervisées et de variantes existantes, voici une liste non exhaustive de celles utilisées au cours de cette thèse et couramment utilisées dans l'état de l'art en recommandation de musique :

1. k-means : il s'agit d'une méthode plutôt ancienne, mais toujours très utilisée en clustering [74]. Le but de cet algorithme est d'obtenir k partitions des données observées. À chaque itération, chaque point est affecté au cluster dont le barycentre est le plus proche selon une métrique donnée, les barycentres de chaque cluster sont ensuite recalculés tant que des points changent de cluster. À l'initialisation, c'est l'utilisateur qui choisit le nombre de barycentres ainsi que leur position, de manière aléatoire ou non.
2. GMM (Gaussien Mixture Model) : cette méthode peut être considérée comme une extension de k-means, où l'écart-type de chaque cluster est considéré en plus de sa moyenne. À chaque itération, un point est affecté au cluster pour lequel la vraisemblance (voir [43]) d'y appartenir est maximale, compte tenu des covariances et moyennes du cluster [90].

3. Classification ascendante hiérarchique : cette méthode a déjà été présentée de manière détaillée dans la partie 1.4.3.
4. Réduction de dimension : parmi les méthodes de réduction de dimension les plus populaires, nous comptons l'ACP [121], et la t-SNE [73], plus récente. Ces méthodes n'aboutissent pas à une classification à proprement parler, mais à une projection des données dans un nouvel espace de dimension inférieure, dont les axes ne correspondent plus aux variables initiales. Le but de ces méthodes est de réduire la taille des données, afin de les rendre d'une part plus lisibles afin d'en obtenir une représentation acceptable par l'humain, et d'autre part de réduire le temps de calcul lors de traitements ultérieurs. La création des nouvelles variables qui correspondent à des combinaisons linéaires des variables initiales se fait de manière à réduire au maximum le nombre de dimensions tout en conservant un maximum d'informations. Ces méthodes sont souvent utilisées en amont de méthodes de classification supervisées ou non, afin d'en faciliter le calcul. Quand des méthodes permettent de restituer l'information initiale avec un minimum de pertes, elles peuvent être utilisées à des fins de compression.

2.1.3 Apprentissage supervisé

De même que précédemment, voici une liste non exhaustive de méthodes d'apprentissage supervisé couramment rencontrées en recommandation de musique :

1. k plus proches voisins : la méthode des k-PPV, ou kNN en anglais a été proposée pour la première fois en 1967 [34]. Chaque individu à classer est affecté à la classe majoritaire des k points les plus proches dans l'espace où sont projetées les données. La variable k est déterminée par l'utilisateur de l'algorithme.
2. SVM : les machines à vecteurs de support (ou Support Vector Machine en anglais) sont souvent utilisées dans des problèmes de classification dans des cas où les classes ne sont pas linéairement séparables [31]. La projection des données dans un espace de dimension supérieure à l'aide d'une fonction noyau *non-linéaire* permet alors de les séparer linéairement.
3. Arbre de décision [20] : dans un arbre de décision, chaque nœud compare la valeur d'un des paramètres de l'individu d'entrée à un seuil et l'affecte à une des deux branches selon si le seuil est dépassé ou non. Après une succession de nœuds et de branches, l'individu est affecté à une des classes de sortie. Différentes méthodes existent afin d'entraîner l'arbre à partir de données d'apprentissage, l'une des plus populaires étant les Random forest [19].
4. Réseaux de neurones artificiels, du perceptron [91] au deep learning [68] : les réseaux de neurones profonds (ou Deep Neural Network) ont vu leur utilisation exploser au cours des dernières années. La partie suivante

donnera une description d'un type particulier : les réseaux de neurones convolutionnels. Ceux-ci seront utilisés durant le prochain chapitre pour la prédiction de goûts musicaux.

2.1.4 Les réseaux de neurones convolutionnels

Les réseaux de neurones convolutionnels (ou CNN) font partie des méthodes de traitement d'images dites « *end-to-end* ». Ces méthodes ont pour particularité de se passer de l'extraction de paramètres : les modèles peuvent recevoir en entrée un signal sous sa forme brute. Dans le cas des CNN, le signal d'entrée est généralement stocké sous la forme d'un tableau à n dimensions. Une succession de filtres de convolution permet d'extraire directement dans l'image des paramètres pertinents pour la classification ou la régression. À mesure que les couches se succèdent, les paramètres deviennent de plus en plus de « haut niveau », jusqu'à atteindre la couche de sortie qui renvoie l'information de plus haut niveau.

Initialement, les CNN étaient utilisés pour le traitement d'images et plus récemment ils le sont pour traiter l'audio qui est dans ce cas représenté à travers un spectrogramme (voir partie 2.5.2). L'application des CNN à l'audio est donc un cas particulier d'utilisation, où l'entrée du modèle n'est pas à proprement parler une image. Sur une image, les deux axes peuvent être confondus : l'image peut par exemple subir une rotation sans pour autant en affecter son contenu, la forme représentée sera toujours reconnaissable : il s'agit de la propriété d'invariance par rotation [67]. En revanche sur un spectrogramme les deux dimensions n'ont pas la même signification : l'une correspond aux fréquences et l'autre au temps, on ne peut donc pas appliquer de rotation à un spectrogramme sans affecter son contenu. Cependant, malgrès cette différence entre une image et un spectrogramme, les CNN ont su montrer leur efficacité sur de nombreuses applications au traitement de la parole [62] et de la musique [32].

Un réseau de neurones convolutionnel est composé de différents éléments dont nous allons donner une brève explication : filtre de convolution, couches denses, pooling...

— Filtre de convolution
  En traitement du signal, l'opération de convolution peut s'appliquer à une matrice ou à un vecteur. La matrice à laquelle est appliquée l'opération est parcourue par une matrice noyau (kernel). En sortie, le résultat correspond à la somme des éléments de la matrice d'entrée pondérée par les coefficients de la matrice noyau (voir figures 2.1 et 2.2).
  La nature des coefficients du noyau définit le type de filtrage : passe-haut, passe-bas, etc. Dans le cas d'un réseau de neurones convolutionnel, les coefficients du noyau sont directement appris par le modèle durant la phase d'apprentissage.
— Pooling
  Une opération de pooling, qui consiste à effectuer un sous-échantillonnage, est effectuée entre les différentes étapes de convolution. Ce sous échantillonnage peut se faire selon différentes règles : dans le cas du *max pooling*,

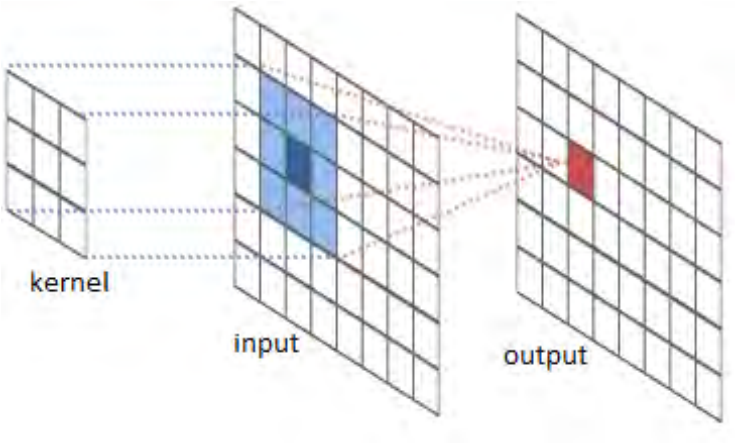

FIGURE 2.1: Noyau, entrée et sortie d'un filtre de convolution.

c'est la valeur la plus forte dans le voisinage de pixels qui est retenue, dans le cas d'un *average pooling* nous conservons la moyenne des pixels adjacents. Les opérations de pooling ont pour intérêt de réduire la quantité de données et donc de calculs, mais aussi de rendre le modèle invariant par translation [101].

— Couche dense

À la suite de la succession d'opérations de convolution et de pooling, un CNN comporte des couches denses. La sortie $\hat{y}$ d'un neurone d'une couche dense correspond à la somme pondérée par les poids W_i de tous les éléments X_i de la couche précédente, à laquelle est appliquée une non-linéarité (voir équation 2.1 et figure 2.3).

$$\hat{\mathrm{y}} = f(\sum_{i=1}^{n}(\mathrm{X}_{i+1}\mathrm{W}_i) + \mathrm{W}_0) \tag{2.1}$$

Dans un CNN, les couches denses ont pour but d'effectuer la tâche de classification ou de régression à partir des paramètres extraits par les couches de convolution.

Différents types de fonctions non linéaires peuvent être utilisés, dont notamment : Sigmoid, Tanh, ReLU, soft ReLU (ou leaky ReLU) (voir figure 2.4). Le but de ces fonctions est de modéliser les relations non linéaires dans les données, où les relations entre les variables ne sont pas proportionnelles.

Par exemple, si nous voulons modéliser l'état liquide ou solide de l'eau à pression ambiante en fonction de la température, il suffit d'un neurone. Nous pouvons considérer que l'état de la matière est une fonction non linéaire de la température, et que pour une pression donnée, elle ne dépend que de la température. Le modèle n'a qu'un seul paramètre d'entrée, la température, auquel nous pouvons appliquer une sigmoïde, qui vaudra

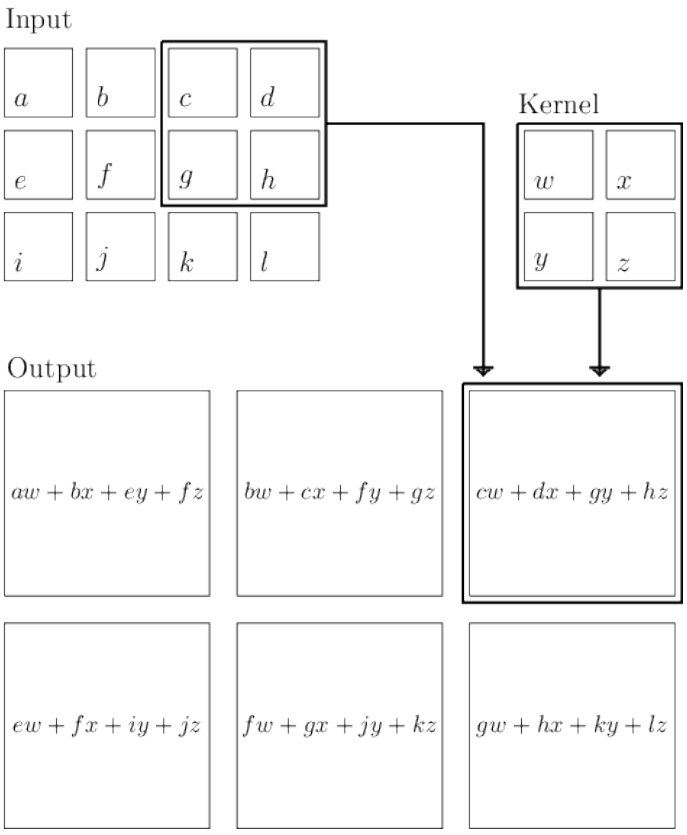

FIGURE 2.2: Calcul d'une convolution.

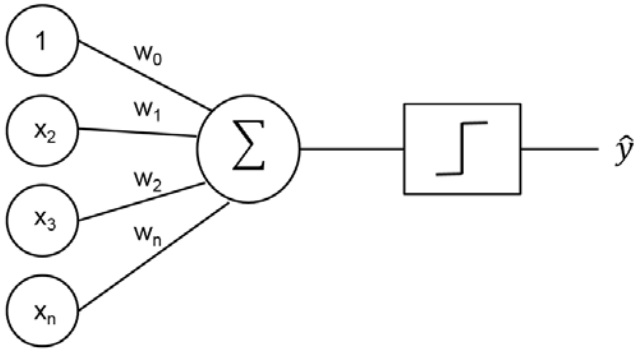

FIGURE 2.3: Structure d'une couche dense.

1 pour les températures positives, et 0 pour les températures négatives. Ainsi, si la sortie vaut 1 (resp. 0) notre modèle prédit un état liquide

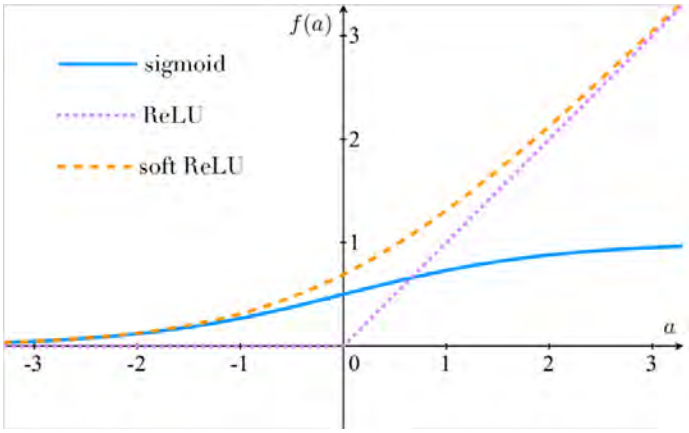

FIGURE 2.4: Non-linéarités utilisées dans les réseaux de neurones artificiels.

(resp. un état solide).

— Softmax
Le rôle de la fonction Softmax est de transformer un vecteur Z de taille K contenant n'importe quelles valeurs en un vecteur $\sigma(Z[j])$ dont la somme des composantes est égale à 1. Cette fonction est utilisée dans les tâches de classification où les classes sont mutuellement exclusives.

$$\sigma(Z[j]) = \frac{e^{Z[j]}}{\sum_{i=1}^{K} e^{Z[i]}} \forall j \in \{1, \ldots, K\} \tag{2.2}$$

La figure 2.5 décrit l'architecture générale des réseaux de neurones convolutionnels. Le nombre de couches, leur dimension et leur organisation peuvent différer d'une tâche à une autre.

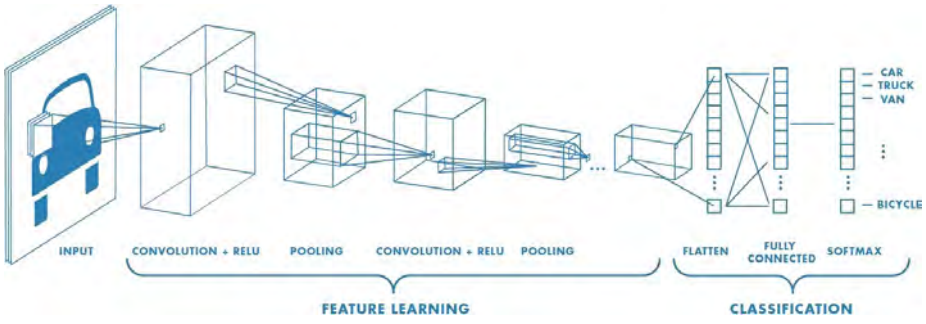

FIGURE 2.5: Architecture d'un réseau de neurones convolutionnel (Source : Matworks).

Fonction de coût, optimisation

Lors de l'entraînement du modèle, les paramètres des réseaux de neurones, tels que les poids des couches denses et des filtres de convolution, sont ajustés automatiquement. Cet ajustement se fait par la résolution d'un problème d'optimisation : la majorité du temps, la fonction de coût est fonction de la différence entre les prédictions et les classes connues des données d'apprentissage. Le but de l'optimisation est de réduire au maximum l'erreur commise (fonction de coût) par le modèle en ajustant automatiquement les poids. Différents algorithmes d'optimisation existent pour le deep learning et ce domaine fait toujours l'objet de recherches [27].

2.1.5 Données d'entraînement, de validation, de test et sur-apprentissage

En apprentissage supervisé, les algorithmes sont entraînés sur des données dites d'entraînement, et sont testés sur des données de test. De plus, des données de validation peuvent être utilisées afin de limiter le sur-apprentissage d'un modèle. Le sur-apprentissage désigne le fait qu'un modèle soit plus performant sur des données d'entraînement que sur des données de test. Dans ce cas, nous considérons qu'il est allé au-delà de la généralisation du problème et qu'il a « appris les exemples par cœur ». Cela peut se produire lorsque le nombre de paramètres du modèle est grand devant le nombre de paramètres des données d'entrées [77], [56]. Par exemple, lorsque des données sont linéairement séparables, 2 paramètres a et b suffisent à décrire la droite y de séparation des données : $y = ax + b$.

Sur la figure 2.6, nous voyons que la courbe noire, qui est modélisée à partir de moins de paramètres, a davantage généralisé les données. Si la courbe verte a une meilleure précision sur les données d'entraînement, elle risque en revanche d'être erronée sur de nouvelles données.

Sur des méthodes d'apprentissage profond, le sur-apprentissage est un problème récurrent qui peut être limité en utilisant des méthodes telles que le dropout [108].

Par ailleurs, nous pouvons mesurer le sur-apprentissage en testant régulièrement le modèle sur des données de validation. Tant que la fonction de coût sur les données de validation diminue de manière corrélée avec celle des données d'entraînement, le modèle apprend normalement. Une fois que les deux fonctions divergent, nous pouvons considérer que le modèle est en train de sur-apprendre sur les données d'entraînement. Il est alors temps d'interrompre l'entraînement du modèle, puisqu'il ne généralisera pas davantage. Par exemple, sur l'exemple en figure 2.7, on voit que la fonction de coût de validation commence à croître à partir de la 8ème itération, et la fonction de coût d'entraînement semble atteindre un plateau : il faut alors arrêter l'entraînement.

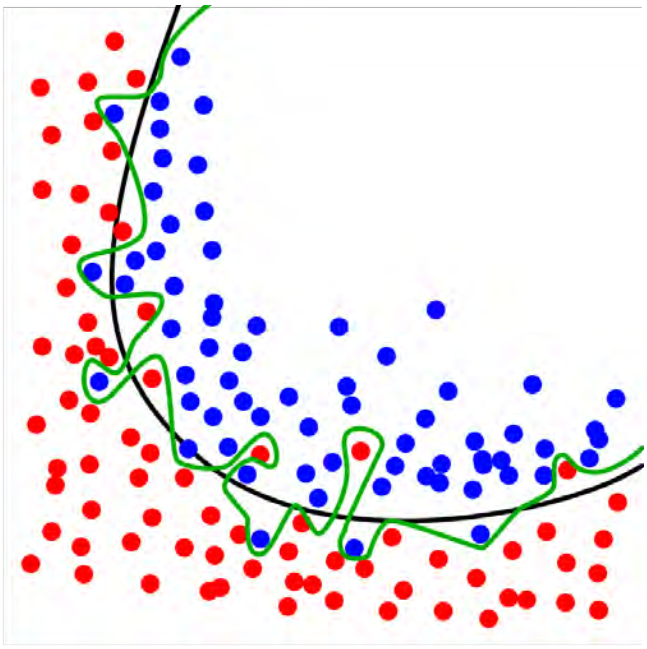

FIGURE 2.6: Apprentissage correct (courbe noire) et sur-apprentissage (courbe verte) (source Wikipedia).

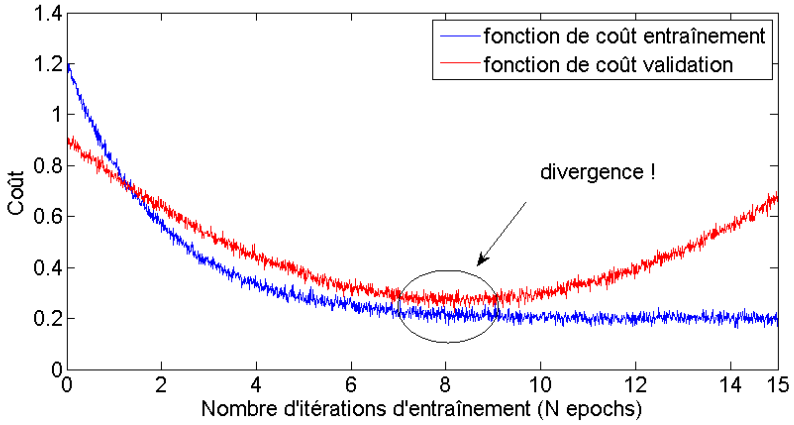

FIGURE 2.7: Fonctions de coût de test et de validation au cours de l'apprentissage.

2.1.6 Apprentissage automatique pour la recommandation de musique

La recommandation de musique repose sur des méthodes d'apprentissage automatique avec des spécificités qui sont propres à ce domaine. Nous allons dans les parties suivantes nous intéresser aux données, aux techniques de modélisation et aux modes d'évaluation mis en oeuvre en recommandation de musique (voir figure 2.8), afin de savoir à quel point il s'agit d'une prédiction de goût.

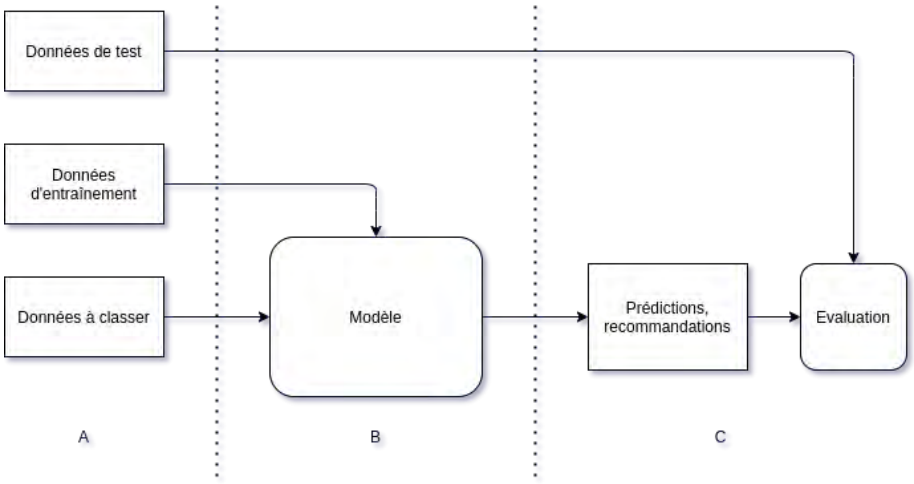

FIGURE 2.8: Les 3 grandes parties de la recommandation de musique : (A) données, (B) techniques de modélisation, (C) modes d'évaluation.

2.2 Qu'est-ce que les données comportementales permettent d'identifier ?

Nous venons de décrire un ensemble de méthodes d'apprentissage automatique qui peuvent être appliquées à la recommandation de musique et à la prédiction de goûts. La grande majorité de ces algorithmes sont supervisés, c'est-à-dire qu'ils ont besoin d'une quantité suffisante de données dites « d'apprentissage » pour être entraînés (voir partie 2.1.3). Dans cette section, nous allons voir à partir de quelles données les industriels appliquent leurs algorithmes de recommandation. Nous allons également nous intéresser à la pertinence des données utilisées dans ce contexte.

2.2.1 Feedbacks Explicites/Implicites

Dans le cas des applications industrielles, les données utilisées pour l'entraînement des algorithmes de recommandation sont collectées lorsqu'un utilisateur

utilise la plateforme. Ses réactions face un morceau sont alors relevées, nous les appelons des *feedbacks*. Nous pouvons distinguer 2 types de feedbacks :

— Feedbacks explicites : lorsque l'utilisateur déclare explicitement aimer ou non un morceau. Selon les plateformes, un feedback explicite peut être un *like*, un *dislike* ou une note. Un like définit l'action de presser un bouton « j'aime » sur l'interface utilisateur. Cette option est omniprésente sur les plateformes de musique en ligne sous différents noms tels que *« j'aime », « like », « love », « coup de coeur », « favoris »*, etc. Pour l'utilisateur, la fonction du like est généralement d'enregistrer un morceau afin d'y avoir rapidement accès plus tard. Pour la plateforme, cette information peut être utilisée pour alimenter le système de recommandation. Certaines plateformes donnent la possibilité à leurs utilisateurs de donner une note pour un morceau ou un album. L'information n'est alors plus binaire, mais quantifiée, elle peut être positive ou négative. Enfin, le dislike est l'option inverse du like. En permettant à l'utilisateur de déclarer son aversion pour un morceau, la plateforme peut alors affiner ses recommandations. C'est par exemple le cas sur Deezer, où cette fonction prend le nom de *« ne plus me recommander ce titre »*. Cette possibilité reste plutôt marginale : Spotify, le leader du marché du streaming ne propose pas de moyen d'exprimer une aversion pour un titre.
— Les feedbacks implicites sont généralement des données de navigation, comme la durée d'écoute d'un morceau, le nombre d'écoutes, le fait d'être ou non passé au morceau suivant (Skip)...

2.2.2 Skip

L'action « Skip » désigne le fait de passer au morceau suivant avant la fin du morceau courant. Cette action est un feedback implicite auquel la plateforme Spotify accorde beaucoup d'attention, au point même d'y avoir consacré un challenge, le Skip Prediction Challenge [22], où les participants avaient pour but de prédire avec la meilleure précision les Skips des utilisateurs.

Cette tâche constitue toujours un défi en intelligence artificielle de par la nature même du phénomène à prédire. En effet un Skip n'a pas toujours la même signification. Si un utilisateur peut effectivement être lassé par un morceau, rien n'indique que son Skip n'a pas une fonction simple de navigation : il aime le morceau en cours, mais il passe au suivant, car il en cherche un autre. Dans ce cas-là, « Skip » ne doit pas être interprétée comme une marque d'aversion pour le morceau en cours. Ainsi, même si un algorithme permettait de prévoir avec précision le Skip (ou l'absence de Skip) d'un utilisateur sur un morceau, il ne serait pour autant pas certain de la corrélation entre un Skip et l'aversion de l'utilisateur pour ce morceau.

Les données à disposition des participants au Skip Prediction challenge étaient de 2 types :

— descripteurs sur les morceaux : paramètres acoustiques, métadonnées, etc.

— logs de session : durée des écoutes, présence de Skip, contexte d'écoute, etc.

À l'aide de toutes ces informations sur une partie des sessions d'écoutes d'utilisateurs de Spotify, les participants au challenge avaient pour but de prédire avec précision les Skips de la deuxième partie de session.

Avec l'aide du stagiaire Jérémie Huteau nous avons ré-implémenté les algorithmes [60] ayant donnés les meilleurs résultats. Ainsi, nous avons pu obtenir les *cartes de saillances* de ces algorithmes (voir figure 2.9).

Ces cartes nous montrent sur quels paramètres les algorithmes reposent pour faire leurs prédictions [107]. L'observation de ces cartes de saillance nous a montré que :

— les paramètres acoustiques n'ont eu que très peu d'influence sur la prédiction de Skip. Autrement dit, le modèle entraîné automatiquement sur les données ne tient pas compte de la nature acoustique des morceaux.
— l'un des paramètres les plus importants pour le modèle pour ses prédictions était les skips sur les morceaux précédents ainsi que des informations sur le contexte d'écoute et l'activité de l'utilisateur.

En conclusion, pour prédire le plus efficacement possible les Skips, le plus simple reste encore d'observer les Skips d'un utilisateur donné sur les morceaux précédents. Si l'utilisateur a sauté les morceaux précédents, alors il y a de fortes chances qu'il saute le morceau courant, et ce quelle que soit la nature du morceau.

Dans ce contexte, se baser sur les Skips afin d'en déduire les goûts des utilisateurs semble donc peu justifié.

2.2.3 Nature et abondance des différents feedbacks

Dans les logs mis à notre disposition par Deezer (voir section 1.3.1), les skips sont présents en abondance. Sur près de 3 millions d'écoutes effectuées par les 10.000 utilisateurs suivis, plus de 1 million de Skips ont étés observés. C'est donc plus de 37% des écoutes qui ont été interrompues par un Skip de l'utilisateur. À titre de comparaison, le « Ban », bouton servant à signifier à la plateforme une insatisfaction vis-à-vis d'une recommandation, n'a été utilisé que 5.000 fois. La différence entre ces deux chiffres illustre parfaitement les quantités relatives de données entre les feedbacks explicites et implicites : il y a 200 fois plus de données sur les Skips que les Bans. De plus, on ne compte qu'un peu plus de 20000 likes, soit 50 fois moins que de skips. L'information implicite des skips est donc bien plus abondante que celle des ban et love explicites. Cette différence d'abondance peut expliquer que le skip soit privilégié pour l'entraînement des algorithmes de recommandation de musique, certaines méthodes de deep learning nécessitant une grande quantité de données. Par ailleurs, une plus grande quantité de données peut amener à une plus forte robustesse statistique pour l'entraînement des modèles de manière générale.

En compilant toutes les durées des écoutes qui ont été interrompues par un Skip, nous avons tracé l'histogramme des durées d'écoutes des morceaux

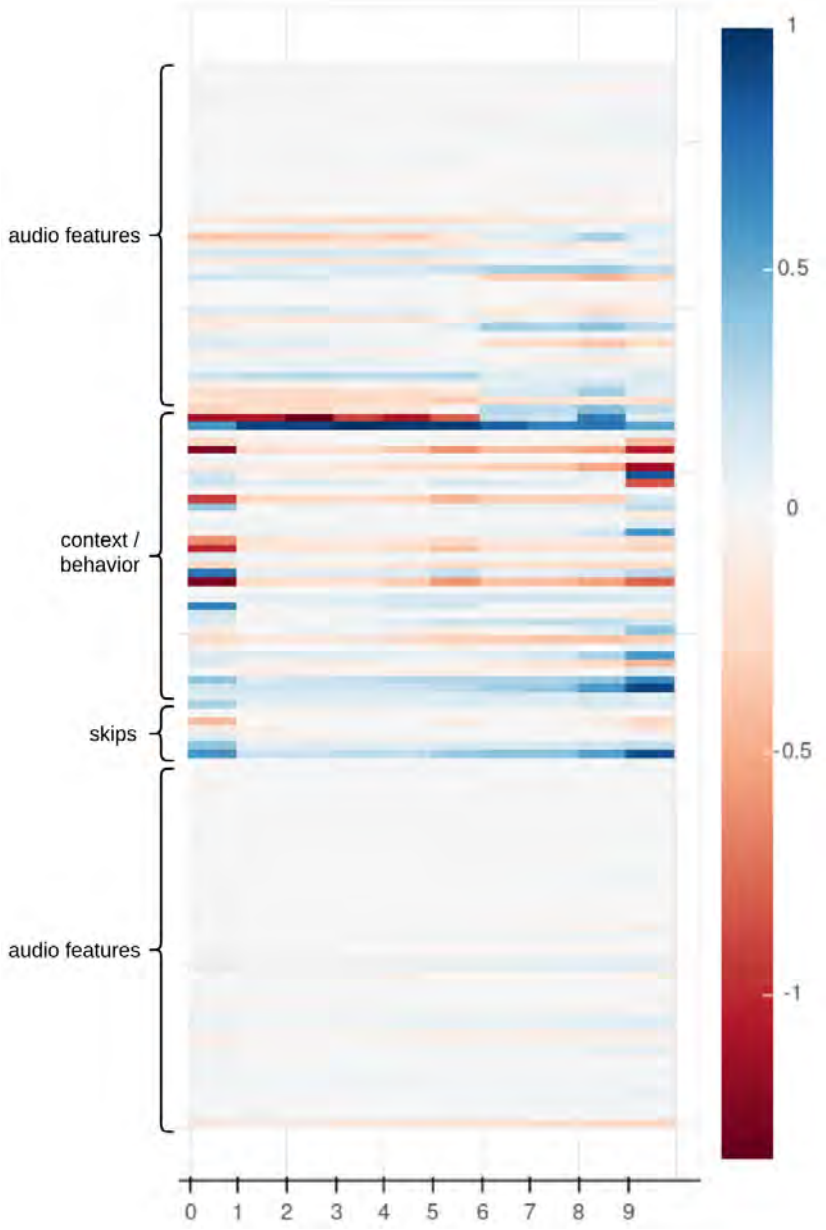

FIGURE 2.9: Carte de saillance. Chaque ligne indique un paramètre, chaque colonne correspond à un morceau, et l'échelle de couleur correspond à l'intensité du gradient.

« skippés ». Sur les figures 2.10 et 2.11 nous voyons que dans la très grande majorité des cas, les morceaux sont abandonnés au bout de quelques secondes.

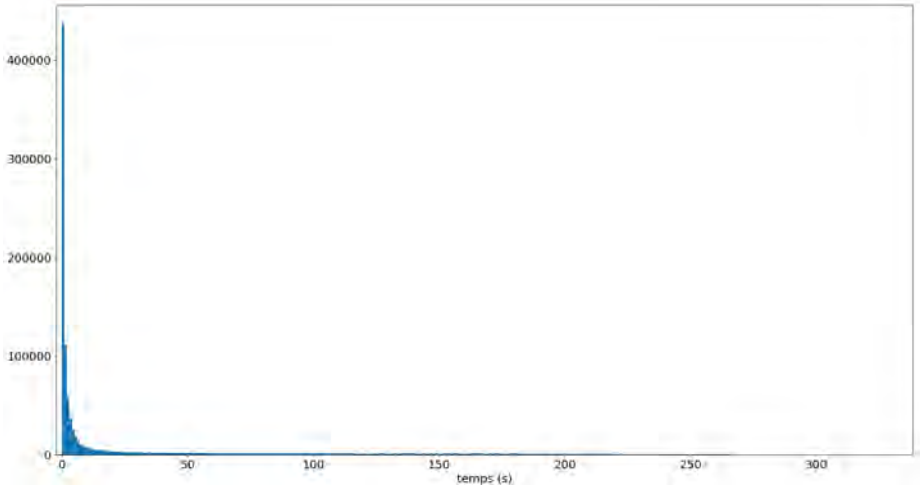

FIGURE 2.10: Histogramme de la durée d'écoute précédant un skip.

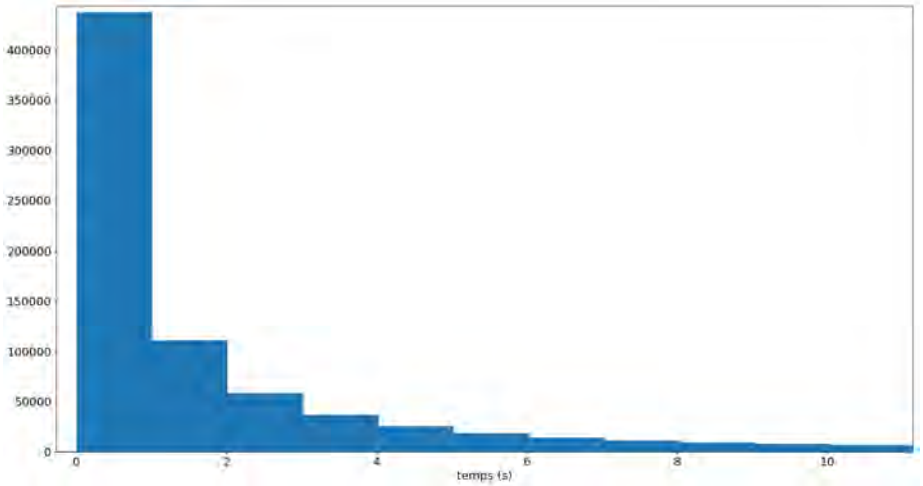

FIGURE 2.11: Zoom sur l'histogramme de la durée d'écoute précédant un skip.

Sur plus d'un million de Skips enregistrés dans les données, 66% ont eu lieu au bout de 2 secondes d'écoutes ou moins ! Par ailleurs, une analyse du top 100 du site Deezer par Antoine Vervier, un musicologue impliqué dans notre projet PERMUSES, nous a montré que 90 titres ont une intro d'une durée de 12 secondes en moyenne, les deux tiers des intros durent au moins 10 secondes. Pour les morceaux conformes à ces formats dominants, les écoutes de moins de 12 secondes ont donc de grandes chances de ne comprendre que l'intro.

Au vu de la très courte durée de ces écoutes, nous pouvons donc nous interroger une fois de plus sur la pertinence du Skip comme un indicateur d'aversion pour un morceau. Un utilisateur peut par exemple utiliser cette commande pour naviguer parmi une playlist ou un album dans le but de trouver un morceau précis et cela ne veut pas dire pour autant qu'il n'aime pas les morceaux « skippés » : il cherche peut être simplement autre chose.

2.2.4 Impact de l'heure d'écoute

À l'aide des données fournies par Deezer, nous avons affiché le nombre d'écoutes par heure, ainsi que le nombre de skips, voir figure 2.12. Sans surprise, ces deux grandeurs sont fortement corrélées et une forte baisse d'activité est constatée pendant la nuit.

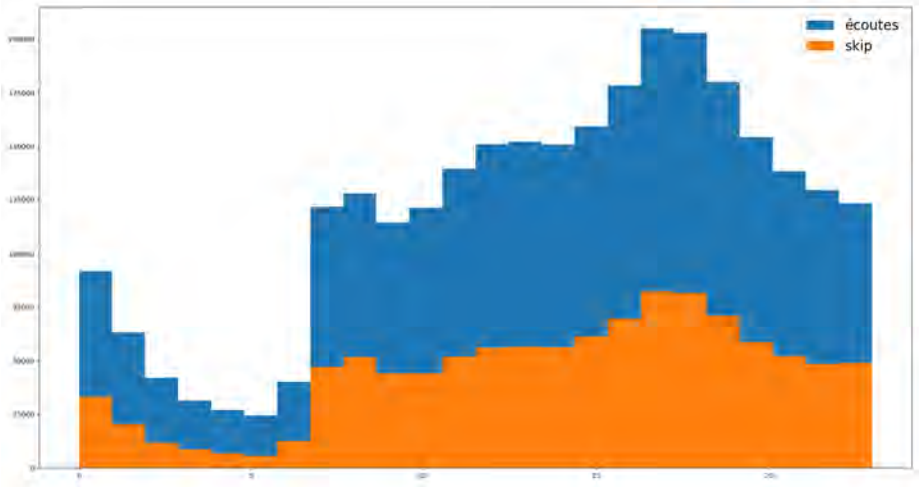

FIGURE 2.12: Nombre de skips par heure.

Nous avons ensuite normalisé le nombre de skips par heure, par le nombre d'écoutes : voir figure 2.13.

Nous voyons sur cette figure que le nombre de skips par écoute diminue fortement durant la nuit. De plus, nous observons également des creux de 10h à 15h ainsi qu'aux alentours de 21h. Cela nous montre que l'attention des utilisateurs est plus faible sur certains créneaux horaires. Cette baisse d'attention peut être liée à une une écoute davantage en « musique de fond » par les utilisateurs. Ainsi, se baser sur le skip peut à la fois biaiser la connaissance de l'algorithme de l'utilisateur, mais également biaiser l'évaluation de cet algorithme. Sur ces horaires, les « non-skips » peuvent être perçus comme un signal positif provenant de l'utilisateur alors qu'en réalité son attention est tout simplement diminuée.

De plus, les données affichées ici ne montrent qu'une tendance sur 10.000 utilisateurs. En réalité, chacun possède ses propres habitudes et les moments

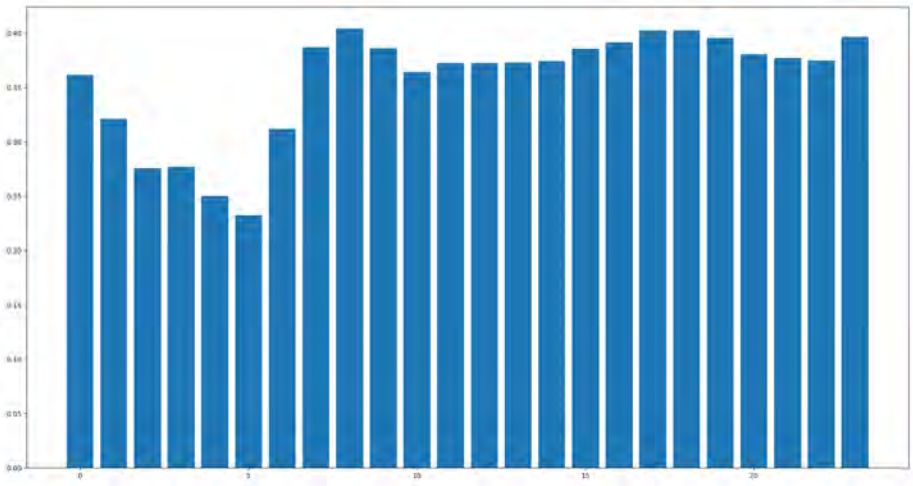

FIGURE 2.13: Nombre de skips par heure, normalisé par le nombre d'écoutes.

d'écoutes actives et passives peuvent différer d'un utilisateur à un autre. Ainsi, le skip pourrait être davantage fiable en intégrant pour chaque utilisateur les habitudes en termes d'écoute active et passive.

Dans [52], les auteurs, en se basant sur des données d'utilisateurs de Spotify ainsi que des questionnaires, ont pu isoler plusieurs types de profils différents. Selon les attentes qu'ils ont de la plateforme et leur activité en parallèle à l'écoute de musique, ils n'auront pas la même tendance à utiliser le bouton skip. Ainsi, il apparaît que selon le cas d'utilisation, il serait pertinent d'utiliser des profils différents pour le même utilisateur. Les métriques d'évaluation de la recommandation de musique pourraient alors s'adapter aux différents usages.

2.2.5 Ecouter = aimer ?

L'écoute d'un morceau par un utilisateur constitue le feedback le plus basique qui puisse être collecté. Certaines méthodes de recommandation sont évaluées sur des corpus où le seul feedback disponible est l'écoute des morceaux par les utilisateurs, tels que dans [125]. Dans ce cas, des méthodes proposées dans [80] sont utilisées. Deux cas extrêmes sont évoqués : tous les morceaux non-écoutés par un utilisateur sont considérés comme « dislikés » par cet utilisateur, ou bien les feedbacks sont considérés comme inconnus pour les morceaux non écoutés. Dans [125], les auteurs ont décidé de considérer les morceaux écoutés par un utilisateur comme des exemples positifs et de constituer un ensemble d'exemples négatifs à partir de morceaux choisis aléatoirement parmi ceux qui n'ont pas étés écoutés par l'utilisateur. Dans [80], les auteurs ont montré que cette solution était un bon compromis entre « tous les exemples manquants sont négatifs » et « tous les exemples manquants sont inconnus » dans le cadre du

filtrage collaboratif.

Néanmoins, nous pouvons nous interroger sur la pertinence d'une telle démarche : ici, que prédisons-nous et qu'évaluons-nous alors ? Nous avons vu précédemment (partie 2.2.3) qu'une grande partie (37 %) des morceaux écoutés sont skippés. Dans certaines recherches, un morceau skippé est considéré comme un exemple négatif : il y aurait donc 37% de morceaux « mal annotés » dans un corpus s'appuyant uniquement sur les écoutes. À l'inverse, nous avons pu voir que certains skips ne traduisaient pas particulièrement l'aversion d'un utilisateur pour le morceau concerné, mais seulement un geste de navigation sur l'interface, auquel cas le morceau pourrait être considéré comme non-écouté.

Si nous pouvons avoir des doutes sur une partie des morceaux annotés comme « liké » car écoutés, qu'en est-il des morceaux considérés comme « dislikés » car non écoutés ? Si un utilisateur n'a jamais écouté un morceau, rien ne nous indique qu'il ne l'aime pas ! Le problème que pose ce genre d'hypothèse est qu'elle tend à enfermer les utilisateurs dans une bulle de recommandation : si les morceaux inconnus de l'utilisateur sont considérés comme « dislikés » par ce dernier, alors ils ne lui seront pas recommandés, et de même pour les morceaux considérés comme similaires.

Un système de recommandation étant supposé fournir à un utilisateur des morceaux qu'il n'a potentiellement jamais écoutés et qu'il pourrait aimer, l'hypothèse initiale sur laquelle s'appuie un tel système est donc erronée... Les bons résultats affichés par les méthodes évoquées ici ne prouvent pas le contraire : ces méthodes ont peut-être réussi à prédire quels morceaux sont écoutés par un utilisateur, mais il est difficile de dire si les goûts ont été prédits.

2.2.6 Conclusion sur les feedbacks

Les feedbacks implicites représentent une quantité de données bien plus importante que les feedbacks explicites. Le feedback implicite le plus utilisé est le skip. Le sens à donner à son usage peut changer selon la personne, l'heure de la journée, l'usage fait de l'application, etc. Ainsi, utiliser les skips à des fins de recommandation requiert une interprétation, et donc une prudence quant au sens à donner aux résultats. Bien qu'il ait été montré que des algorithmes puissent prédire des évènements d'écoutes ou des skips, seuls les comportements des utilisateurs sont modélisés, et non pas leurs goûts. Ainsi, l'usage de feedbacks *explicites* apparaît comme crucial dès lors que nous voulons prédire les goûts à proprement parler.

2.3 Filtrage Collaboratif

Après avoir introduit les algorithmes d'apprentissage automatique dans leur ensemble ainsi que les variables prédites en recommandation de musique, nous allons à présent expliquer les méthodes couramment utilisées pour la prédiction, ainsi que leurs intérêts et leurs limites. Ces algorithmes étant mis en application par des industriels, le but ici n'est pas d'en donner le fonctionnement détaillé

puisque cette information relève du secret industriel. Nous allons dans un premier temps nous intéresser au filtrage collaboratif.

Dans le domaine de la recommandation au sens large, nous considérons des utilisateurs et des objets entre lesquels sont observées des interactions. Les objets en question peuvent être des morceaux de musique, des films, ou n'importe quel produit sur un site de vente en ligne. Les interactions considérées peuvent être un achat, une écoute, un visionnage, ou bien encore une simple consultation. Ainsi, nous pouvons construire une matrice d'interaction r entre tous les utilisateurs et tous les objets :

$$r = \begin{array}{c} \begin{array}{cccc} \text{Utilisateur 1} & & \text{Utilisateur i} & \\ \downarrow & & \downarrow & \end{array} \\ \begin{pmatrix} r_{1,1} & \dots & r_{1,p} \\ \vdots & & \vdots \\ r_{n,1} & \dots & r_{n,p} \end{pmatrix} \end{array} \begin{array}{l} \\ \leftarrow \text{Objet 1} \\ \\ \leftarrow \text{Objet u} \end{array} \tag{2.3}$$

Si nous considérons N objets et P utilisateur, la matrice r est de dimensions $\{N, P\}$. Dans la majorité des domaines de recommandation, cette matrice est parcimonieuse, car chaque utilisateur ne peut consulter qu'une petite partie du catalogue disponible. Le but du filtrage collaboratif est d'estimer une matrice $\hat{r}$ non parcimonieuse, permettant ainsi de prédire une interaction inconnue entre un utilisateur i et un objet u.

De manière générale, la matrice $\hat{r}$ est estimée à partir de la factorisation de deux matrices : A, la matrice des utilisateurs de dimension $\{N, L\}$ et M la matrice des objets, de dimensions $\{L, P\}$:

$$\hat{r} = A \times M \tag{2.4}$$

À chaque utilisateur i est associé le vecteur q_i et à chaque objet le vecteur p_u, tous les deux de longueur L. Ainsi, l'interaction estimée entre un utilisateur i et un objet u est donnée par :

$$\hat{r}_{u,i} = q_i^T p_u \tag{2.5}$$

Le calcul des vecteurs q_i et p_u se fait généralement par la minimisation de l'erreur quadratique entre les interactions connues $r_{u,i}$, et l'estimation $\hat{r}_{u,i}$ correspondante :

$$\min_{q,p} \sum_{u,i \in \kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(||q_i||^2 + ||p_u||^2) \tag{2.6}$$

κ désigne l'ensemble des interactions $\{u, i\}$ connues, et λ est un coefficient constant. Le terme de droite est un terme de régularisation ayant pour but de limiter le sur-apprentissage (voir section 2.1.5), en pénalisant la norme des vecteurs q_i et p_u.

Les vecteurs q_i et p_u contiennent des variables latentes, ce qui veut dire que ces variables ne sont pas directement observables dans les données. Ce terme vient du domaine de la réduction de données, et le filtrage collaboratif peut

être considéré en tant que tel. En effet, dans la matrice d'interaction chaque morceau est caractérisé par les utilisateurs qui l'ont écouté, ou non. Ainsi, le vecteur représentant chacun des morceaux contient N variables binaires, pour les N utilisateurs. Après la factorisation de matrice, chaque morceau est représenté sous la forme d'un vecteur de variables latentes de dimension L. La taille de ce vecteur est fixée par l'utilisateur de l'algorithme, et en général nous avons $L << N$ (par exemple $L = 50$, pour des millions d'utilisateurs). Plus ce vecteur est grand, plus il contiendra de variables latentes et pourra décrire avec précision les morceaux correspondants.

De nombreuses méthodes de factorisation de matrices existent [63], et certaines sont basées sur de l'apprentissage profond [123].

La méthode du filtrage collaboratif est très largement utilisée, au-delà de la recommandation de musique, et elle permet de recommander des objets sans aucune connaissance sur leurs caractéristiques, seules les interactions comptent. C'est pourquoi le choix du type d'interaction est déterminant : certains sites recommandent selon qui a *consulté* un produit, alors que d'autres s'intéresseront aux *achats* uniquement (voir figures 2.14 et 2.15).

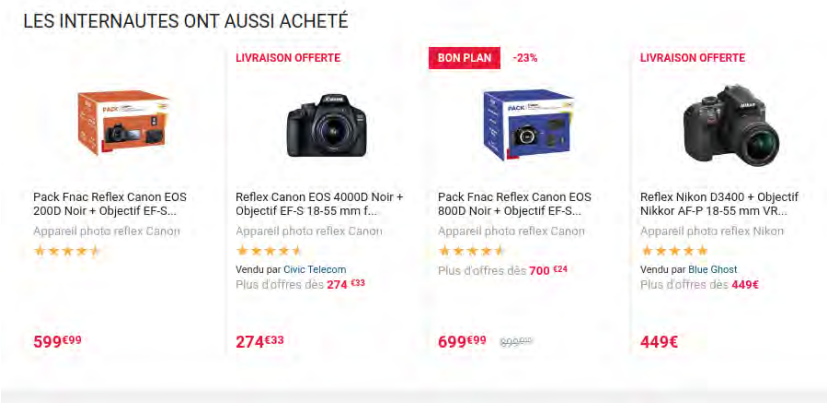

FIGURE 2.14: Exemple de recommandation « achat ».

De même que pour la musique, voulons-nous prédire si un utilisateur va **écouter** un morceau ou s'il va **l'aimer** ? Dans le premier cas, les données seront abondantes et la matrice d'interaction moins parcimonieuse, mais laisseront beaucoup de place à l'ambiguïté : comme nous venons de le voir, ce n'est pas parce qu'un utilisateur a écouté un morceau qu'il l'a aimé.

Avantage

L'avantage le plus évident du filtrage collaboratif est qu'il est simple et très rapide à mettre en œuvre. En inférant *a priori* un sens à une action, nous évitons la diversité et les incohérences pourtant quelquefois inhérentes à une action.

FIGURE 2.15: Exemple de recommandation « consultation ».

La relation action / sens est choisie *a priori* par les développeurs ou gestionnaires de sites, calquant généralement leurs choix sur leurs propres comportements. Cette méthode est simple, rapide, et répond dans certaines conditions à l'exigence de résultat.

Défauts

Son avantage majeur (la simplicité) est également son plus gros défaut puisque les inférences sont trop rigides pour expliquer l'ensemble des raisons pour lesquelles nous pouvons faire une seule action. Dans le cas de la musique par exemple, nous écoutons (ou nous « likons ») parce que nous connaissons déjà, parce que nous ne connaissons pas, parce que quelqu'un nous en a parlé, parce que j'ai lu des articles sur ce titre, parce que je suis curieux, parce que c'est le même interprète que j'aime d'habitude, etc.

Il existe également un certain nombre de problèmes plus «techniques» que nous évoquerons à la suite de cette partie.

— Le problème du *cold-start* (ou « démarrage à froid » en français) concerne les morceaux ou utilisateurs n'ayant eu peu ou aucune interaction. Pour un utilisateur ayant écouté ou noté peu de morceaux, les recommandations risquent d'être aléatoires, puisque le système n'a pas eu suffisamment d'informations pour apprendre de ses actions et ne peut donc pas trouver d'utilisateur aux actions similaires. À l'inverse, un morceau qui n'a jamais été écouté ou noté ne pourra pas être recommandé [85]. Le problème du cold-start est davantage pénalisant pour les morceaux que pour les utilisateurs : un utilisateur inconnu par la plateforme finira par ne plus l'être de par ses différentes actions. À l'inverse, un morceau jamais écouté risque de le rester s'il n'est pas mis en avant par l'algorithme de recommandation. Le problème du cold-start fait toujours partie des challenges majeurs en recommandation de musique [113, 29, 79, 100].

— L'uniformisation des recommandations. Dans les faits, les morceaux recommandés sont bien souvent déjà connus du grand public. Par exemple, dans la capture d'écran en figure 2.16, le morceau cible était 'Lump Sum'

de l'artiste Linval Thompson, qui comptabilise environ 50.000 écoutes, que nous pourrions donc considérer comme « de niche ». Nous voyons que parmi les 8 recommandations associées à ce morceau, 4 ont plus d'un million d'écoutes et 3 ont plus d'une centaine de milliers d'écoutes. Un seul morceau a un nombre limité d'écoutes et pourrait constituer une découverte, mais il s'agit en fait d'un morceau du même chanteur. Une étude, menée par A. Ferrar [44] a montré que le problème des recommandations uniformes était inhérent à l'usage du filtrage collaboratif.

FIGURE 2.16: Exemple de recommandation de musique.

En conclusion, si le filtrage collaboratif est facile à mettre en place et ne demande aucune connaissance *a priori* sur les objets à recommander, il semble pénaliser les morceaux les moins connus et favoriser les plus connus. Parmi les solutions pour répondre au problème du cold-start, la plus répandue est l'approche « basée contenu » que nous allons décrire dans la section suivante.

2.4 Approche basée contenu

Dans cette section, nous allons définir les approches basées contenu pour la recommandation de musique ainsi ses défauts et ses avantages.

L'approche basée contenu repose sur des descripteurs pour recommander des objets [1]. Ces descripteurs peuvent être des métadonnées, comme l'année de sortie d'un film ou une annotation en genre d'un morceau, ou bien extraits

automatiquement à partir du signal. Nous pouvons ainsi recommander à un utilisateur des objets similaires à ceux avec lesquels il a interagi.

Quand ces méthodes s'appuient sur des paramètres acoustiques, leur mode d'extraction et leur choix sont déterminants. Les différents descripteurs couramment utilisés pour la recommandation de musique sont décrits dans la partie 2.5.

2.4.1 Calcul de proximité/similarité entre 2 morceaux

À partir de métadonnées ou de paramètres acoustiques, un morceau peut être représenté sous forme d'un vecteur à N dimensions. Ainsi, en projetant l'ensemble des morceaux d'une base de données dans un espace, nous pouvons connaître, pour un morceau *requête*, le ou les morceaux de la base de données les plus proches de ce morceau (voir figure 2.17). Ce type de méthode peut être employé à partir d'un seul feedback positif d'un utilisateur : *« si vous aimez ce morceau, alors vous aimerez peut être ceux-ci »*.

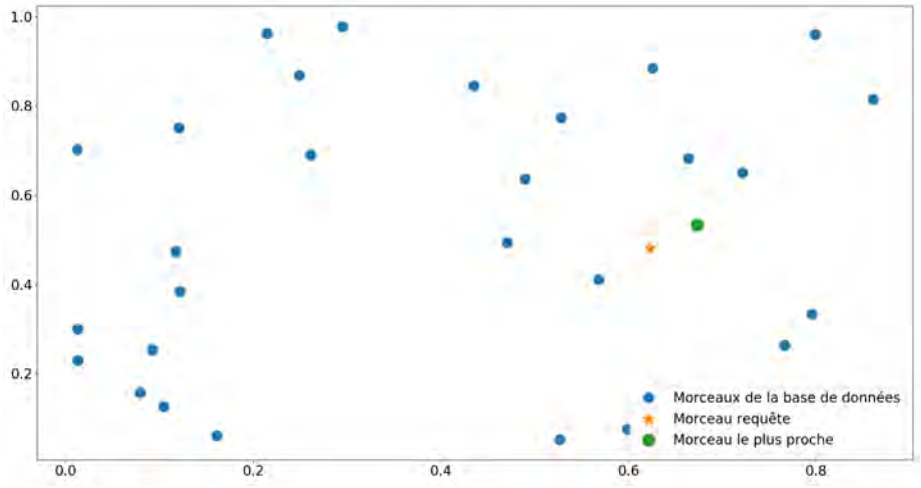

FIGURE 2.17: Recommandation par similarité.

Les méthodes basées sur la mesure de similarité sont largement utilisées en recommandation basée sur le contenu : [105], [128], [5]. Il existe une multitude de moyens de mesurer la similarité entre deux morceaux, et à d'autres fins que la recommandation de musique, comme l'identification de morceau [106], [54].

2.4.2 Modélisation des goûts par un ensemble de morceaux

Si la méthode précédente peut s'avérer efficace pour renvoyer un morceau similaire à un morceau requête, elle reste néanmoins éloignée d'une réelle prédiction de goûts. Si un seul morceau suffisait à définir nos goûts musicaux, la

recommandation de musique ne constituerait pas un tel challenge à l'heure actuelle ! Afin d'être décrits de manière suffisamment exhaustive, les goûts d'une personne doivent être définis par un ensemble de morceaux représentatifs. Cependant, si une personne a des goûts variés, leur projection selon des paramètres acoustiques peut mener à la constitution de différents clusters. Dans une telle situation, il apparaît donc périlleux de représenter l'ensemble des goûts (pluriels !) d'une personne à travers un seul vecteur « moyen » de paramètres acoustiques (voir figure 2.18).

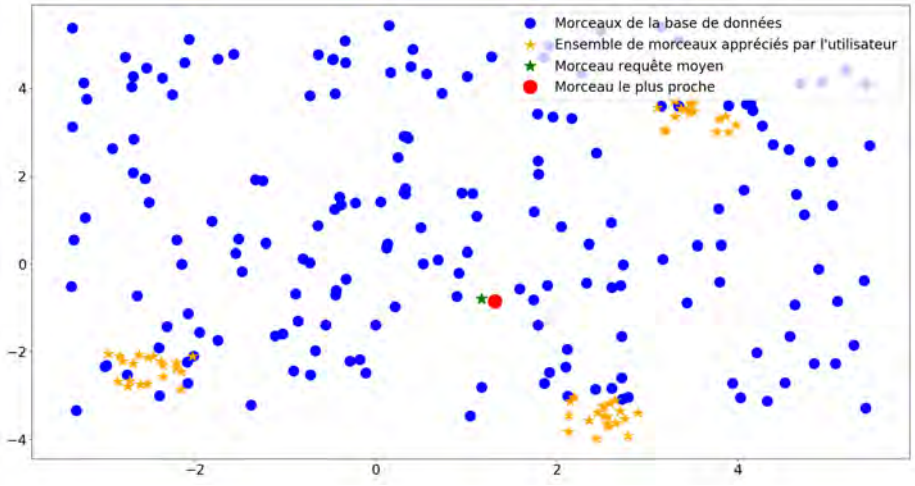

FIGURE 2.18: Recommandation par similarité, morceau requête moyen.

Par exemple, si une personne est amatrice de Rock et de Classique, il n'est pas dit que les morceaux se situant « entre les deux » dans une représentation acoustique lui conviennent.

Une solution pour répondre à cette problématique est de représenter un utilisateur sous la forme de plusieurs sous-profils. Par exemple, dans [15] et [125], les auteurs représentent les goûts de chaque utilisateur par un ensemble de morceaux, rassemblés en différents clusters réalisés à partir des méthodes k-means et GMM.

2.4.3 Méthodes hybrides

Certaines méthodes s'appuient à la fois sur le filtrage collaboratif et sur l'approche basée sur le contenu.

Estimation de facteurs latents

Nous avons vu dans la partie précédente que les méthodes de filtrage collaboratif s'appuient sur des méthodes de factorisation de matrices afin de représenter

chaque utilisateur et chaque morceau sous la forme de vecteurs de variables latentes. Dans [114], les auteurs utilisent des réseaux de neurones profonds afin d'estimer ces variables latentes à partir de paramètres acoustiques extraits sur le signal audio. Ainsi, ils sont parvenus à partir d'informations acoustiques à estimer des variables provenant des données d'interaction avec une précision de 77%. De plus, l'analyse qualitative menée par les auteurs montre qu'une recommandation basée sur les vecteurs estimés à partir de données acoustiques présentait plus de variabilité, ce qui représente un avantage compte tenu des défauts du filtrage collaboratif.

Les méthodes hybrides n'étant pas au cœur des travaux de cette thèse, elles ne seront pas détaillées davantage ici. Des états de l'art récents des méthodes de recommandation de musique sont donnés dans [96] et [24].

2.5 Différents types de contenus

Dans cette section nous allons définir les différents types de contenus utilisés pour la recommandation de musique.

2.5.1 Contenus textuels

Des morceaux de musique peuvent être annotés de diverses manières. Les morceaux disponibles sur les plateformes de streaming sont associés à des métadonnées qui donnent des informations sur la date de sortie, l'artiste, l'album ou bien encore le genre du morceau. En général, les informations contenues dans les métadonnées sont fournies à la plateforme par les maisons de disques. Ces métadonnées peuvent être récupérées via les API des plateformes, comme celle de Deezer notamment[1]. Certaines plateformes comme lastFM[2] permettent aux auditeurs de taguer eux-mêmes les morceaux de musique. Ce type d'annotation présente plusieurs défauts majeurs, tels que la présence de fautes de typo engendrant plusieurs tags pour la même classe (« métal », « metal », « Mtal »), ou des tags erronés (« disco » à la place de « funk »). En revanche, ces annotations sont souvent abondantes, ce qui a pu permettre leur utilisation dans la création de corpora tels que LFM-1b [95] et The Million Song Dataset [13].

La plateforme Pandora[3] a quant à elle fait appel à des musicologues pour annoter sa collection de morceaux. Ces annotations sont utilisées par l'algorithme de recommandation.

Les contenus textuels sont utilisés dans d'autres domaines de recommandation. Les méthodes mises en oeuvre pour les traiter ne sont pas spécifiques à la recommandation de musique [16].

1. https://developers.deezer.com/api
2. https://www.last.fm/fr/
3. http://www.pandora.com/about/mgp

2.5.2 Les paramètres acoustiques

Les paramètres acoustiques sont calculés automatiquement sur le signal. Il en existe de plus ou moins « haut niveau ». Les paramètres de bas niveau sont *calculés* alors que les paramètres de plus haut niveau sont *estimés*. Par exemple, la puissance du signal possède une définition stricte, son calcul ne souffrira donc d'aucune ambiguïté. Des paramètres comme le tempo sont également basés sur une définition « simple » mais il s'agit d'une mesure qui peut avoir une imprécision, donc d'une *estimation*.

Durant cette thèse, nous avons utilisé la toolbox MIR (Music Information Retrieval) [65] afin d'extraire ces paramètres. Dans cette partie, les différents paramètres que nous avons utilisés sont définis, ainsi que les méthodes permettant de les calculer. Ces dernières sont détaillées de manière plus approfondie dans le manuel de cette toolbox [65], et un schéma décrivant le lien entre toutes ces fonctions est fourni en Annexe D.

Les paramètres acoustiques peuvent être distingués en quatre grandes catégories : paramètres de dynamique, de rythme, de timbre et de hauteur.

Paramètres de dynamique

En traitement du signal, et de manière générale, la dynamique décrit la plage de variation des différentes valeurs prises par un signal. En musique, la dynamique décrit le rapport entre des sons d'amplitudes fortes et faibles. La dynamique dépend du jeu des musiciens, mais est également dépendante des traitements effectués sur la musique lors de la production ainsi que du support via laquelle elle est diffusée. Certains supports disposent de plages dynamiques plus élevées que d'autres : par exemple, le CD peut atteindre jusqu'à 100 dB de plage dynamique pour seulement 60 dB pour un disque vinyle.

Niveau RMS La valeur efficace d'un signal aléatoire ergodique x sur un intervalle temporel τ est la racine carrée de la moyenne de ce signal au carré, ou autrement dit, la racine carrée de sa puissance moyenne. Dans la pratique, pour un signal à temps discret, le niveau RMS est calculé sur n échantillons :

$$Xrms = \sqrt{\frac{1}{n}\sum_{i}^{n} x_i^2} \tag{2.7}$$

Il est très commun pour les producteurs de musique d'appliquer un compresseur de dynamique aux signaux (instruments seuls, groupés ou à tout un morceau). Le but de ce procédé est d'augmenter le niveau RMS - le niveau sonore ressenti par l'auditeur - sans pour autant faire saturer le signal. Cette pratique est plutôt récente et dépend des styles de musiques. Le niveau RMS du signal peut donc constituer un indice pour distinguer des morceaux entre eux. Par exemple, en musique classique le niveau peut se situer autour des -15 dB RMS alors qu'il peut atteindre -3 dB RMS pour de la techno. Par ailleurs, l'évolution temporelle du niveau RMS est équivalente à l'enveloppe du signal.

Taux de faible énergie (ou Low Energy Rate) Le taux de faible énergie correspond au nombre de points dont la valeur est inférieure à la valeur RMS du signal. Pour un signal comportant des pics au niveau RMS élevé, ce taux sera élevé alors que pour un signal au niveau RMS plutôt constant, ce taux sera faible (voir Figure 2.19). Ainsi, le Low Energy Rate nous informe sur le contraste en termes d'intensité présent dans un morceau de musique.

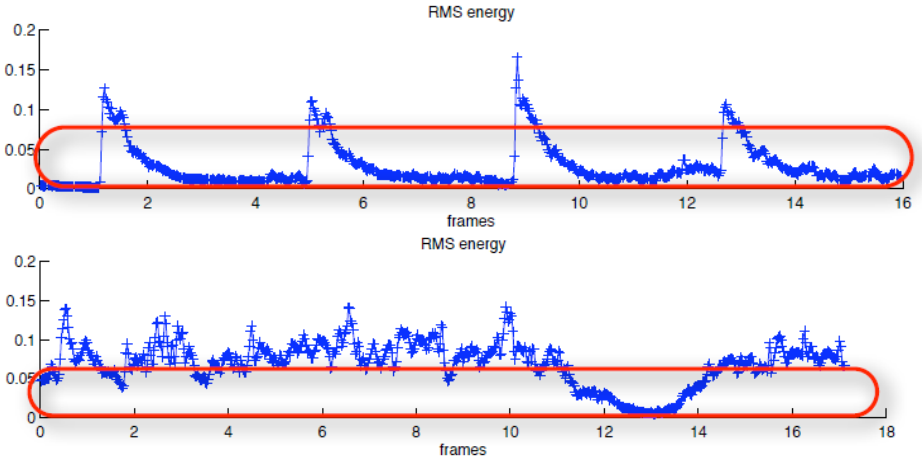

FIGURE 2.19: Taux de faible énergie : en haut, de fortes variations et en bas, de faibles variations. (Source : MIR manual)

Paramètre rythmiques

Le rythme décrit la localisation temporelle des événements sonores ainsi que leur durée. Dans la musique occidentale conventionnelle, une pulsation régulière détermine les temps, une mesure étant composée de plusieurs temps. Dans une partition, le rythme est décrit par les différentes figures de notes (croche, noire, blanche...) et de silences (pause, soupir...) ainsi que par le chiffrage.

Détection des événements Tous les paramètres rythmiques s'appuient dans un premier temps sur la localisation temporelle de chaque événement. Pour cela, nous utilisons un algorithme de détection de pics sur l'enveloppe du signal (voir Figure 2.20).

Densité d'événements Une fois que les pics sont détectés, nous pouvons ensuite calculer le nombre d'événements par seconde. Pour N événements détectés sur une durée τ nous avons :

$$N_{event} = \frac{N}{\tau} \tag{2.8}$$

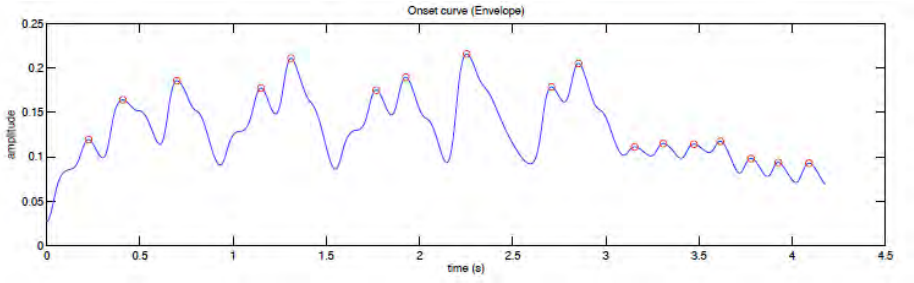

FIGURE 2.20: Exemple de détection de pics sur un extrait de 4 secondes. (Source : MIR manual)

Tempo Le calcul du tempo s'appuie sur une détection de la périodicité des événements, et sélectionne le pic le plus élevé. La détection de périodicité s'effectue à l'aide de la fonction d'autocorrélation [65].

Clarté de la pulsation La clarté de la pulsation (pulse clarity) est calculée selon la méthode détaillée dans [66]. Ce paramètre décrit à quel point la pulsation est dominante dans le rythme, ou autrement dit, à quel point l'accent est mis sur les temps. Par exemple, la clarté de la pulsation est forte pour des rythmes disco, et est souvent faible pour des rythmes complexes, comme ceux du jazz.

Paramètres de timbre

Le timbre décrit la composition spectrale d'une note, c'est-à-dire l'amplitude des harmoniques et la variation dans le temps de ces harmoniques. C'est ce qui distingue par exemple deux notes jouées à la même hauteur par un piano et une guitare.

Attaque La toolbox MIR dispose de fonctions permettant de détecter la durée, l'amplitude ainsi que la pente des attaques des notes. Ces fonctions s'appuient sur la détection d'événements : quand une note est détectée, nous repérons dans les instants précédents le début de l'attaque, puis nous calculons la différence d'amplitude ou la durée entre les deux points (la pente est obtenue à l'aide de ces deux informations). Un exemple est donné sur la Figure 2.21.

Taux de passage par zéro (ou ZCR : zero crossing rate) Le taux de passage par zéro est calculé sur le signal original en faisant la multiplication de toutes les paires d'échantillons successifs, et en itérant une variable lorsque le produit est négatif (changement de signal). Nous divisons ensuite par la durée pour obtenir le taux. Pour un signal d'une durée N, échantillonné à la fréquence

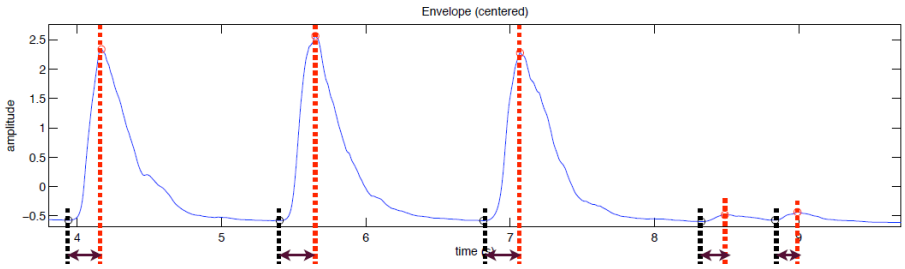

FIGURE 2.21: Illustration de la durée des attaques. (Source : MIR manual)

F_e, le ZCR est décrit par l'équation 2.9, où $f_z(x)$ est une fonction égale à 1 pour $x < 0$ et 0 sinon.

$$ZCR = \frac{1}{N.F_e} \sum_{n=1}^{N-1} f_z(x[n].x[n-1]) \tag{2.9}$$

Spectre, Spectrogramme Le spectre d'un signal sonore est sa représentation fréquentielle. Pour passer d'un signal temporel échantilloné à un spectre, on utilise la transformée de Fourier des signaux discrets (TFD). Pour un signal discret x de dimension N, sa transformée de fourier discrète X se calcule ainsi, avec $k = 0, ..., N-1$:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{2i\pi \frac{nk}{N}} \tag{2.10}$$

X est de même dimension que x. Si x est un signal échantilloné à une fréquence f_e, alors les N points de $X(k)$ correspondent à un échantillonage de la transformée de Fourier du signal analogique original sur N points de fréquence. On a alors $f_k = kf_e/N$, ce qui pour $k = 0, ..., N-1$ nous donne $0 < f < f_e$. En informatique, le spectre d'un signal est calculé à partir d'algorithmes de transformée de Fourier rapide, ou FFT (Fast Fourier Transform) [116]. Le spectre d'un signal contient ses différentes composantes fréquentielles, et est essentiel à l'étude du timbre sonore. Afin d'obtenir l'évolution temporelle des différentes composantes fréquentielles, on peut calculer une succession de FFT à intervalles réguliers. Les vecteurs obtenus pour chaque interval temporel sont concaténés dans une matrice, appelée spectrogramme. Une dimension de cette matrice correspond à l'axe temporel, l'autre à l'axe fréquentiel. Les spectrogrammes sont utilisés pour l'analyse temps-fréquence des signaux. Dans la toolbox MIR, ils servent de base à l'extraction de nombreux paramètres de haut niveau voir annexe D. Dans le chapitre 3 de cette thèse, nous avons utilisé les spectrogrammes conjointement aux réseaux de neuronnes convolutionnels pour une tâche de prédiction de goûts.

Fréquence de roulement (ou Rolloff frequency) La fréquence de roulement nous informe de la quantité d'énergie présente dans les basses fréquences. Sur un spectre, nous calculons la fréquence en dessous de laquelle 85% de l'énergie est contenue [111]. Plus cette fréquence est basse, plus l'énergie est concentrée dans les basses fréquences (cf. figure 2.22).

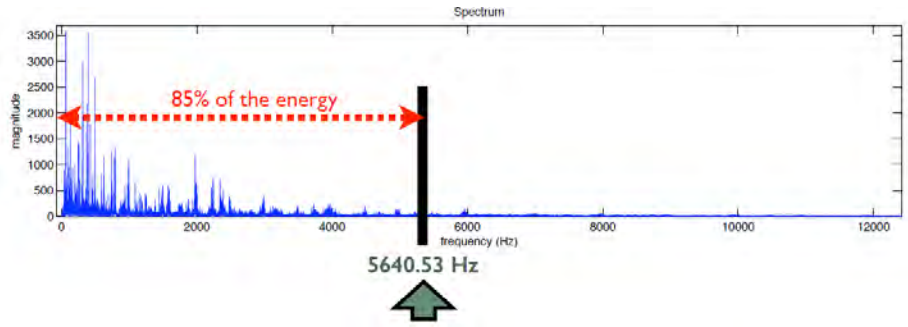

FIGURE 2.22: Illustration du Rolloff. (Source : MIR manual)

Pour un signal discret avec un spectre d'amplitude M_t et de fréquence de roulement discrète R_t cela nous donne :

$$\sum_{n=1}^{R_t} M_t = 0.85 \sum_{n=1}^{N} M_t \tag{2.11}$$

Brillance (ou Brightness) La brillance nous informe sur la quantité d'énergie présente dans les hautes fréquences. Sur un spectre, nous calculons la quantité d'énergie présente au delà d'une fréquence fixée, en général (et par défaut sur MIR) nous avons $fb = 1500$ Hz (cf. figure 2.23).

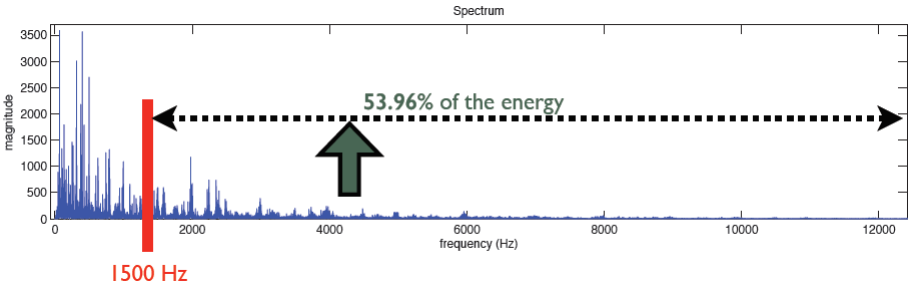

FIGURE 2.23: Illustration de la brillance. (Source : MIR manual)

Pour un spectre de module $|H(f)|$, cela nous donne :

$$B = 2 \int_0^{fb} |H(f)| \, \mathrm{df} \tag{2.12}$$

Paramètres statistiques de la distribution spectrale Il est possible de calculer des statistiques ainsi que des moments de différents ordres sur le spectre : centroïde (barycentre), spread (étalement), skewness (assymétrie), kurtosis (courbure), flatness (platitude...), ainsi que l'entropie.

MFCC (Mel Frequency Cepstral Coefficients) Les MFCC sont des coefficients cepstraux calculés par une transformée en cosinus discrète appliquée au spectre de puissance d'un signal. Les différentes bandes de fréquences sont déterminées selon l'échelle perceptive Mel. L'échelle de Mel est une approximation du système d'audition humain. En général, la quasi-totalité de l'information est contenue dans les 13 premiers coefficients, ce qui est dû à l'utilisation de la transformée en cosinus discrète. Le premier coefficient est assimilable à l'énergie moyenne du signal. Ces coefficients sont très utilisés en traitement audio : parole, musique et sons environnementaux (voir figure 2.24).

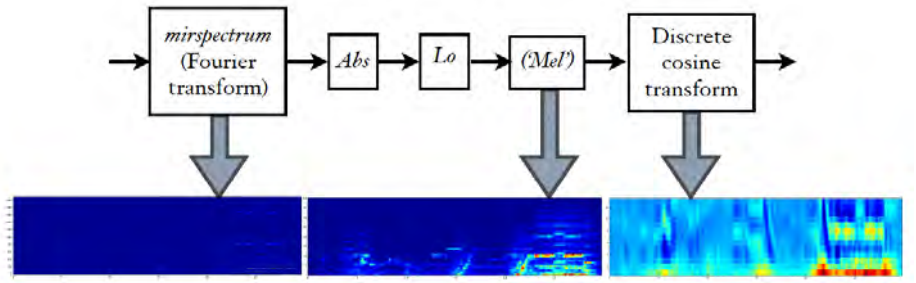

FIGURE 2.24: Illustration du calcul des MFCC. (Source : MIR manual)

Dissonance sensorielle (ou Roughness) La dissonance sensorielle décrit le phénomène de « battement » audible en présence de deux fréquences proches. Deux notes espacées d'un demi-ton (ou moins) généreront une forte dissonance sensorielle, qui diminue à mesure que l'espacement augmente. La dissonance sensorielle est quasi nulle à partir de 5 demi-tons (voir figure 2.25).

Un morceau avec des harmonies complexes donnera un score plus élevé qu'un morceau ne comportant seulement que des quintes et des octaves.

Irrégularité du spectre L'irrégularité d'un spectre est le degré de variation de deux pics (harmoniques ou non) successifs du spectre. Si nous considérons

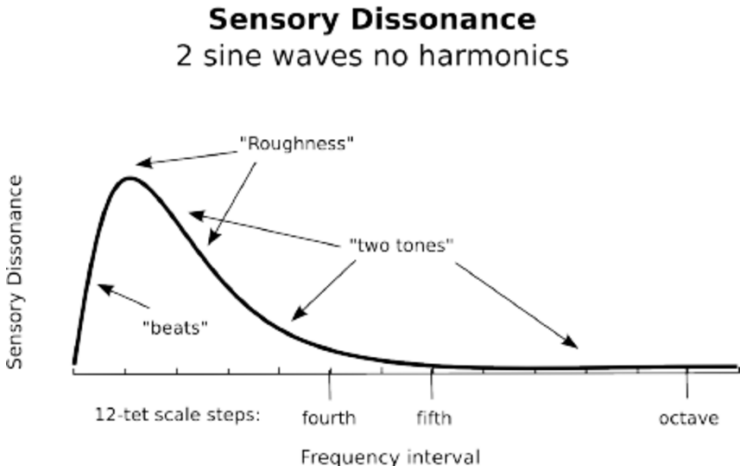

FIGURE 2.25: Illustration de la dissonance sensorielle. (Source : Plompt et Levelt (Sethares, 1999))

a_k les différents pics du spectre détectés, la formule de l'irrégularité du spectre est la suivante [59] :

$$I = \frac{(\sum_{k=1}^{N}(a_k - a_{k+1})^2)}{(\sum_{k=1}^{N} a_k^2)} \tag{2.13}$$

Paramètres de hauteur et d'harmonies

La hauteur décrit la fréquence fondamentale d'une note jouée par un instrument, qui définit la note jouée. Par exemple, le $La3$ a pour fréquence fondamentale (notée F_0) 440 Hertz alors que pour le $Do3$, F_0 vaut $261,6$ Hz.

Détection de notes (hauteur) La méthode utilisée par défaut pour détecter les notes (ou « pitch ») est de décomposer le signal en plusieurs bandes de fréquences, de calculer ensuite l'autocorrélation (voir figure 2.26) et enfin de détecter les pics afin d'obtenir une estimation des notes (voir figure 2.27).

Détection d'harmonies À partir de la détection de notes, il est ensuite possible de détecter des harmonies, c'est-à-dire des combinaisons de notes (accords). Il est également possible de calculer la tonalité d'un extrait, ainsi que l'évolution temporelle de tous ces paramètres. En pratique, MIR toolbox détecte le flux du centroïde tonal, qui correspond à une projection des accords selon le cycle des quintes, des tierces mineures et des tierces majeures [55].

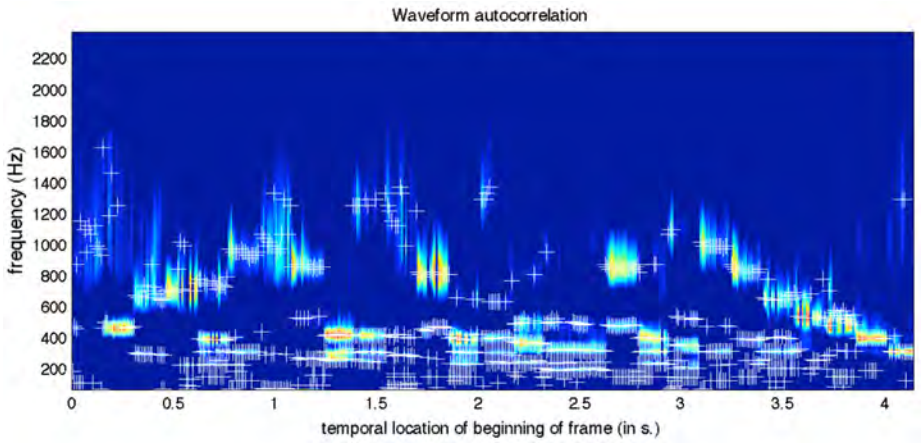

FIGURE 2.26: Exemple de calcul d'autocorrélation du spectrogramme. (Source : MIR manual)

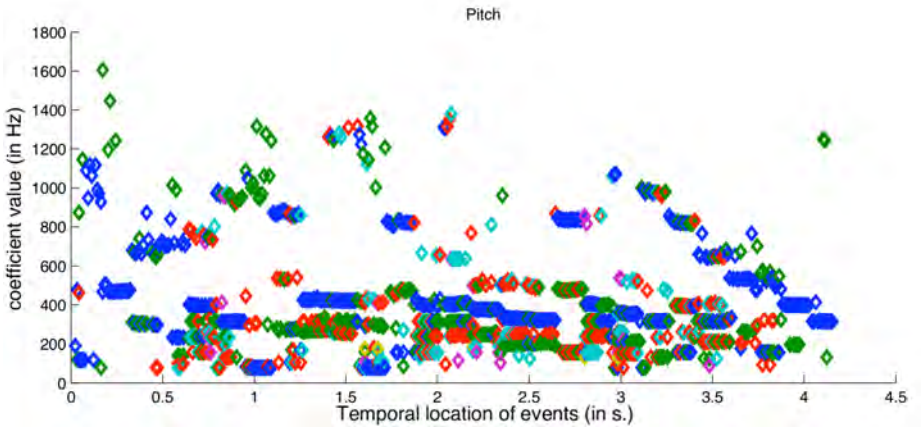

FIGURE 2.27: Estimation de différentes notes sur 4,5 secondes. (Source : MIR manual)

Extraction

1. L'extraction de nombreux paramètres passe dans un premier temps par une transformée de Fourier à court terme (FFT, Fast Fourier Transform).

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{2\pi j}{N} kn} \tag{2.14}$$

Cette FFT s'effectue par défaut sur des fenêtres de 50 ms avec un recou-

vrement de 50%. La fenêtre utilisée est Hamming, qui est la plus adaptée pour cette utilisation [65].

2. Tous les paramètres décrits jusqu'à présent peuvent être extraits sur différentes trames (ou "Frames"). L'utilisateur peut déterminer la durée de ce découpage ou bien utiliser la valeur par défaut de chaque paramètre. En effet, selon le paramètre à extraire, certaines durées seront plus ou moins pertinentes. Par exemple, les paramètres rythmiques comme la clarté de la pulsation sont extraits par défaut sur des fenêtres de 5 secondes, alors que l'extraction des notes et des paramètres de timbre se fait sur des fenêtres de 50 ms.

   De plus, certaines fonctions utilisent une segmentation automatique afin d'isoler chaque note (voir figure 2.28). Ainsi, nous pouvons considérer que chaque "Frame" respecte l'hypothèse de stationnarité et d'ergodicité.

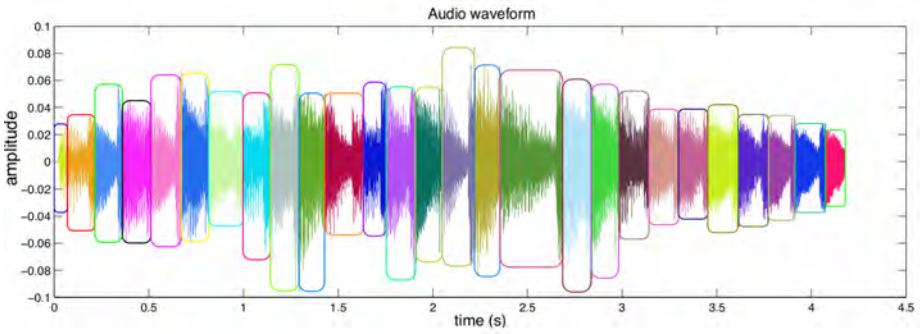

FIGURE 2.28: Segmentation des notes. (Source : MIR manual)

3. Avant d'effectuer un traitement avec les fonctions de la toolbox MIR, il faut extraire les échantillons sonores des fichiers avec la fonction « miraudio ». Lors de cette opération, le fichier est par défaut converti en une piste mono en effectuant la somme des deux canaux dans le cas d'un fichier stéréo, pour plusieurs raisons. D'une part, la plupart des paramètres présentent peu de variation selon le canal. Et, d'autre part, dans le cas de la stéréo certaines fonctions renvoient une valeur par canal alors que d'autres n'en renvoient qu'une au total : ceci pose quelques problèmes pour uniformiser les données. Cela permet également de gagner du temps de calcul lors de l'extraction des paramètres ainsi que lors de leur traitement.

2.5.3 Les descripteurs de haut niveau

Des descripteurs de haut niveau peuvent être extraits à partir du signal audio ou bien de paramètres acoustiques en vue d'une recommandation de musique. Ces descripteurs peuvent concerner l'artiste ou l'instrument comme dans [28],

le genre [103], les émotions [82]. Par ailleurs, Spotify propose via son API[4] des paramètres de haut niveau, tels que : danceability, instrumentalness, liveness, speechiness, etc.

Les descripteurs de haut niveau sont généralement obtenus via des méthodes de machine learning, en particulier basées sur l'apprentissage profond. Ces méthodes s'appuient soit sur des paramètres acoustiques de bas niveau, soit directement sur le signal audio (ou bien sur une analyse temps-fréquence de ce dernier). Les sources des annotations utilisées pour entraîner ces algorithmes peuvent donc être diverses : annotations d'experts, annotation d'auditeurs, annotation de maisons de disques, etc. (voir partie 2.5.1). À terme, le but de ces algorithmes est d'éviter le travail d'annotation manuelle pour obtenir directement des descripteurs de haut niveau qui pourront ensuite être utilisés en recommandation. Les annotations sur lesquelles sont basés ces algorithmes sont produites par des humains. Ces sources d'annotations, qu'elles soient expertes ou novices, sont donc sujettes à la subjectivité et au jugement des personnes qui les auront produites. Ainsi, si un algorithme est supposé estimer le genre de musique d'un morceau avec une forte précision, il faut garder à l'esprit qu'il estimera le genre du morceau d'après la personne qui a annoté le corpus d'entraînement. Par ailleurs, un système de prédiction de genres donnera des résultats différents selon qu'il a été entraîné sur un corpus généraliste comme GTZAN [111] ou sur un corpus spécialisé sur la musique sud-américaine comme [117]. Ces méthodes demeurent performantes du moment où leur contexte d'application reste cohérent vis-à-vis de leur apprentissage.

2.6 Catégorisation libre d'extraits musicaux et analyse automatique.

Nous avons vu que les méthodes de recommandation basées sur le contenu reposent notamment sur des paramètres acoustiques. Ces paramètres sont sélectionnés pour une tâche donnée et recouvrent de manière plus ou moins complète différentes caractéristiques de la musique : *rythme*, *dynamique*, *timbre* et *harmonie*. À l'aide de ces paramètres et de méthodes plus ou moins sophistiquées d'apprentissage automatique, des similarités sont mesurées entre différents morceaux. Cependant, dans le cadre de la recommandation de musique, il faut que les morceaux proposés à un utilisateur soient similaires selon des critères humains.

Par exemple, si nous soumettions à un algorithme la question illustrée en figure 2.29, est-ce que leur réponse serait la même que celle donnée par des humains ? Ces derniers répondraient probablement que l'image C est la plus proche de la A, puisqu'elles comportent toutes les deux un chat. Un algorithme « naïf », basé sur les composantes RGB de l'image, aurait probablement tendance à associer les images A et B ensemble en raison de la dominante verte, tandis

4. https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

qu'un algorithme « sophistiqué » de reconnaissance d'image telle que ceux entraînés sur ImageNet [92] identifirait probablement avec succès un chien dans l'image B et des chats dans les images A et C, les associant alors.

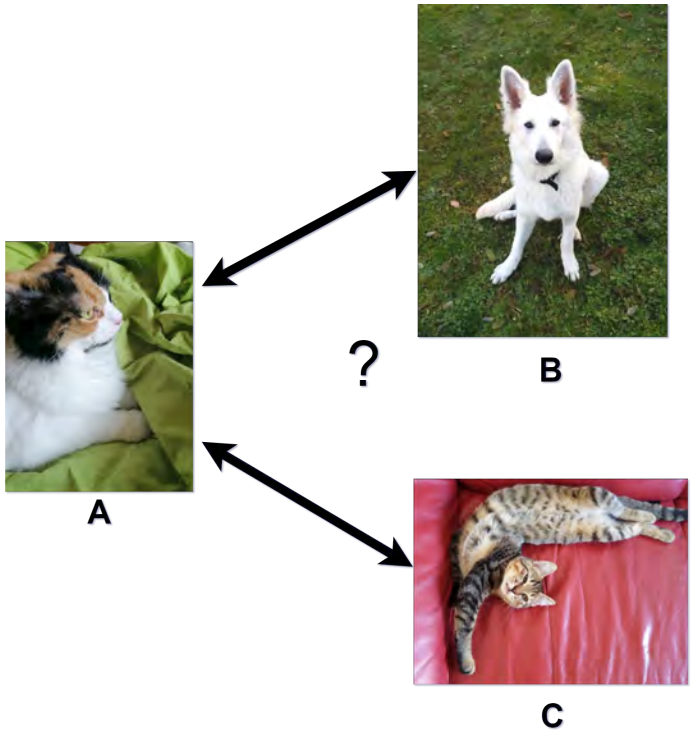

FIGURE 2.29: Quelle image (B ou C) est la plus proche de l'image A ?

Il demeure cependant délicat de déterminer lequel de ces deux algorithmes a raison. Un humain pourrait tout à fait considérer que les images A et B sont plus similaires, car les deux représentent des animaux domestiques sur un fond vert. Il ne s'agit que de sa perception subjective des images, et de la question posée. Dans ce contexte, si nous voulons déterminer quel algorithme est juste, nous pourrions d'une part rendre la question posée plus précise (*« quelle image est la plus proche selon le critère suivant ? »*) Ou bien estimer quel algorithme est en adéquation avec la perception d'un humain ou d'un groupe d'humains. Ainsi, une consigne explicitant le but de la catégorisation peut aiguiller le choix du paramètre discriminant.

Si des paramètres de bas niveau permettent une analyse acoustique précise de la musique, ils restent néanmoins éloignés de la perception subjective des humains (voir figure 2.30).

À l'inverse, outre une incapacité à les modéliserles numériquement, les paramètres s'approchant le plus de la cognition humaine peuvent souffrir d'un

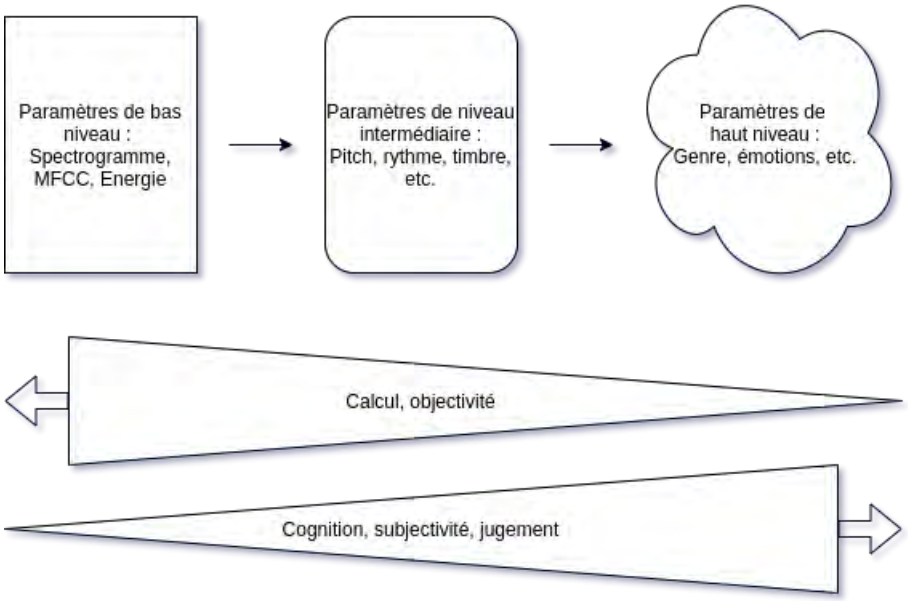

FIGURE 2.30: Des paramètres de différents niveaux d'objectivité.

manque d'objectivité puisqu'ils ont été entraînés à partir de données annotées selon un jugement humain. Parmi tous les paramètres que nous pouvons extraire d'un signal audio, lesquels permettent de décrire et d'expliquer les affinités des auditeurs ? Dans le domaine de la musicologie, les experts possèdent leurs propres critères pour décrire les morceaux. Comment pouvons-nous lier ces paramètres acoustiques à des paramètres musicologiques ?

Quelques expériences ont déjà été menées dans ce domaine, mais M. Schedl, l'un des chercheurs les plus actifs en MIR et en recommandation de musique, déplore que cette discipline ne soit pas assez centrée sur l'humain [97, 98]. Dans [70], les auteurs ont proposé plusieurs méthodes afin de calculer des métriques de distance personnalisées entre des morceaux, basées sur les préférences de l'utilisateur. En 2016, Flexer et. al montrait que l'accord entre différents humains sur la similarité entre les morceaux est limité [46]. Dans [47], il a étudié les différents facteurs menant à davantage d'accords entre les utilisateurs.

Un article [84] de Parizet et. al dresse un bilan de différentes expériences de catégorisation libre d'extraits audio, où le tri permet d'évaluer la dissimilarité entre plusieurs sons.

Nous avons mené une expérience similaire dans le but d'identifier à la fois les paramètres acoustiques et les critères musicologiques ou « non-experts » selon lesquels les sujets classent les morceaux. Cette expérience, qui a donné lieu à deux publications dans des conférences nationales et internationales ([36,

38]), s'est inscrite dans le cadre du projet ECREME : Expertise Cognitive et musicologique pour une REcommandation Musicale personnaliséE[5].

2.6.1 Constitution du corpus selon des critères musicologiques

L'une des premières étapes du projet ECREME a été de constituer un corpus qui devait répondre à plusieurs exigences :

- présenter un large panel de genres musicaux,
- disposer d'extraits de bonne qualité : CD Audio (stéréo, 16 bits, 44,1 kHz),
- posséder des extraits suffisamment longs ($\simeq$ 20 s) et suffisamment nombreux (45 extraits),
- utiliser, de préférence, une base de données en accès libre de droits.

Le corpus a été constitué par Ludovic Florin et Paul Albenge, deux musicologues de l'Université de Toulouse - Jean Jaurès (Laboratoire Lettres, Langages et Arts). Dans un premier temps, ils ont listé un ensemble de 15 critères définissant de manière la plus exhaustive possible la musique de leur point de vue musicologique : *Qualité d'enregistrement, Prédominance d'un instrument, Voix, Espace, Travail Mémoire, Dynamiques, Aspect Narratif, Lisse/Strié, Sensorimoteur, Representations, Règles, Énergie, Niveau de technicité, Inscription dans une sphère culturelle, Distance chronologique.* Ces critères seront définis de manière détaillée dans ce qui suit.

D'ordinaire, la majorité de ces critères sont qualitatifs. Ils ont cependant été quantifiés par les musicologues afin de pouvoir mesurer la diversité du corpus. L'annotation de ce corpus peut être comparée à l'annotation des morceaux de la plateforme Pandora[6]. Cette tâche de quantification de critères musicologiques n'est absolument pas courante et a donc soulevé un certain nombre de remarques et de questionnements de la part des musicologues.

Qualité d'enregistrement Elle quantifie de 1 à 3 la perception du support et medium d'enregistrement (bruits, spectre sonore, intensité...). 1 : Forte altération du son, influe fortement sur l'écoute, 2 : Perception du support d'enregistrement, 3 : Pas d'information permettant d'identifier le support.

Prédominance d'un instrument Il s'agit de la saillance d'un timbre particulier. Répertorie le ou les instruments se trouvant au premier plan.

Voix Elle précise la présence ou l'absence de voix : 1 étant présence, 0 étant absence. Le type de voix (chantée, parlée, criée...) n'est pas pris en compte.

5. Cette étude a été financée par la Maison des Sciences de l'Homme et de la Société de Toulouse

6. http://www.pandora.com/about/mgp

Espace Correspond à la sensation et à la représentation d'un espace de diffusion de la musique. Il s'agit d'une valeur entre 1 et 5 qui représente la grandeur de l'espace perçu, 1 étant proche et fermé, 5 étant lointain et ouvert. 3 représente une espace proche de celui perçu en situation de spectateur d'un concert avec une perception d'une distance moyenne et d'un son provenant, en majorité, de devant.

Travail mémoire Il s'agit de la présence d'un ou de plusieurs éléments mémorisables, qui peut se manifester par la répétition d'un élément (stricte ou bien similaire) ou bien la perception d'une forme ou d'une logique. Cette notion englobe également la notion de prédictibilité de la musique. Elle est quantifiée de 1 à 5, 1 représentant la quasi-absence de travail autour de la mémoire.

Dynamiques Cela correspond aux changements de quantité ou de densité d'événements, et définit le contraste dans le développement musical. Valeurs de 1 à 5, 1 représentant une dynamique faible.

Aspect Narratif Il s'agit de l'évolution d'éléments musicaux, de la présence de différentes parties relativement distinctes. Quantifié de 1 à 5.

Lisse/Strié Ceci reflète l'absence ou la présence de pulsation, la diversité des éléments. Variation dans la quantité d'éléments (dans un temps court). Quantifié de 1 à 3 : 1 représente un temps lisse et 3 un temps strié. Le 2 signifie l'usage des deux conceptions au sein d'un même morceau ou un rapport complexe entre elles.

Sensori-motrice Il s'agit de musiques d'instrumentistes, animées en grande partie par un désir du geste et par une recherche de l'effet sensoriel du son. Elle est représentée par un 0 ou 1, si la musique comporte ou ne comporte pas un aspect sensori-moteur.

Représentations Ceci désigne des morceaux qui représentent de façon visible ou cachée le réel sur le plan plastique, par des trajectoires, des vitesses, des chocs ou même des bruits réalistes (grégorien, romantisme occidental, beaucoup de musiques contemporaines). Ce critère est binaire, la musique comporte ou ne comporte pas un aspect représentatif.

Règle Toutes musiques d'écriture, qu'elles soient réellement écrites, comme le contrepoint classique, ou transmises de mémoire comme les polyphonies pygmées ou les jeux de trompes M'baka. Elle est représentée par un 0 ou 1, si la musique comporte ou ne comporte pas une notion de règle.

Énergie Il s'agit de l'intensité, l'implication corporelle ou l'implication du musicien. Elle est quantifiée de 1 à 5, de manière croissante.

Niveau technicité Il s'agit de technicité instrumentale et de composition : la perception de l'assurance des intentions du musicien. Elle reflète aussi la complexité d'une musique – densité polyphonique, traitements des dissonances, densité harmonique, etc. Valeurs de 1 à 5, 5 étant le niveau le plus technique.

Inscription dans une sphère culturelle Elle est représentée sur une échelle de 1 à 5. Il s'agit de l'inscription plus ou moins précise et explicite dans une sphère culturelle de l'imaginaire et du savoir commun. 5 représente une musique clairement inscrite dans une pratique définie et identifiée. 1 quant à lui illustre une musique difficilement assimilable à une pratique culturelle précise.

Distance Chronologique Il s'agit de la distance chronologique du morceau musical par rapport à l'époque actuelle. La graduation est exponentielle afin de représenter la perception des différentes époques musicales et leurs différentiations par un auditeur de nos jours. Ainsi, 1 représente les musiques ne rappelant pas une époque précise et paraissant intemporelles – plus particulièrement les musiques traditionnelles. Elle est représentée sur une échelle de 1 à 10 : 10 (2010's) ; 9 (2000's) ; 8 (1990's) ; 7 (1980's) ; 6 (1970's) ; 5 (1960's) ; 4 (1ere moitié du XXe siècle) ; 3 (Classique, Romantique) ; 2 (Ancienne, Baroque) ; 1 (Intemporelle).

Ces explications sont issus d'études musicologiques. Nous avons laissé les définitions telles que données par les musicologues pour ne pas en altérer le sens. Une fois ces critères définis, les musicologues ont constitué un corpus de 100 morceaux recouvrant au maximum l'ensemble de ces critères.

Parmi ces 100 morceaux, 45 d'entre eux ont été sélectionnés afin de recouvrir au maximum les 15 critères musicologiques définis précédemment. Les extraits choisis contiennent différentes caractéristiques musicales qui permettent de proposer un ensemble éclectique. Au final, un corpus de 45 extraits de 20 secondes a été défini pour cette expérience. Le corpus entier de 100 morceaux (voir table 2.1) a par ailleurs été utilisé dans une expérience présentée dans le chapitre suivant de cette thèse. La playlist des 100 morceaux est disponible à l'écoute sur le site Deezer[7].

2.6.2 Conditions expérimentales

Afin de limiter l'impact de l'âge des participants sur les résultats, nous avons fait appel à des volontaires de 20 à 25 ans (au nombre de 30).

Pour l'expérience, nous avons utilisé l'outil TCL-labX[8] [51]. L'interface était présentée de manière identique à tous les volontaires à qui nous avons demandé de trier librement des extraits et de former ainsi autant de catégories qu'ils le souhaitaient (voir 2.31).

7. https://www.deezer.com/fr/playlist/5189043344
8. https://mycore.core-cloud.net/index.php/s/EhMhbvgc4NSa9w5

TABLE 2.1: Corpus de 100 morceaux constitué par les musicologues.

Les Doubles Six – Au Bout du Fil (Meet Benny Bailey)	Mark Guiliana – A Quote Machine
Bill Bruford – One Of A Kind (Part 1)	Guillaume de Machaut – Kyrie, La Messe de Notre Dame
Negro Prison Blues And Songs - No More, My Lawd	Meredith Monk – Dusk
Rautavaara, Cantus Arcticus 2e mvt	Paul Bley – Speak Easy
Hector Berlioz - Symphonie Fantastique, Op. 14, Songe d'une nuit de sabbat	Golden Gate Jubilee Quartet – Preacher And The Bear
Jean-Christophe Maillard - Branle de village sur le 8ème ton	Carmen Miranda – Alô, Alô
Namibie : bushmen et himba - Chant himba	Octurn/ The Tibetan Monks of Gyuto – Lhasa
Han Bennink and Willem Breuker - Mr. M.A. de R. in A.	The Thewaprasit Ensemble – Pleng Sen Lao Na
Death Grips – Hot Head	Maalim Abdi – Natamani uwa
Big Satan – Geez	Ravi Shankar – Spring
Jazzoo – Pics et l'hirondelle	Skip James – Devil Got My Woman
Lords Of The Underground – Here Come The Lords	Lord Invader – Out the Fire
Pierre Laurent Aimard/ Aka Pygmies – Yangisa	Ikuta Ryu – Kojo No Tsuki
Tuvalu – Aliamu Ma Ana Mea Tufuga	Goldie – Vanilla
James Brown – Mother Popcorn	Bob Log III – Guitar Party Power
Jakob Dietrich/Ernst Frick - Zäuerli with Schelleschötte	Abe Schwartz – Sher Pt.2
Horace Silver, Capverdian Blues	Boban Markovic – Izvorski biseri
Awa Poulo – Dimo Yaou Tata	Orchestre National de Barbès – Alaoui
Tony Williams – There Comes a Time	Sélébéyone – Are You in Peace
David Fiuczynski – Moonring Bacchanal	James Chance and The Contortions – Contort Yourself
André Minvielle – Madame Mimi	Glenn Branca – Symphony #3, 2nd Movement
Edgard Varèse – Un Grand Sommeil Noir	Harry Partch – The Street
The Residents – This Is Man's World	Dan Deacon – Sheathed Wings
Theo Bleckmann and Ben Monder – Late Green	Le Mystère des Voix Bulgares – Zableyalo Mi Agantze
Aphex Twin – Circlont14 [Shrymoming Mix]	Totó La Momposina - Chí Chí Maní
Tim Hecker – Aerial Silver	Supersilent – 6.5
Bugge Wesseltoft - Dreaming	Don Cherry – Mahakali
Ali Farka Touré – Sabu Yerkoy	Hawk House – My Mind Is The Weapon
Lilli und Lars – A Ram Sam Sam	Austin Peralta – Capricornus
Sleepytime Gorilla Musuem – The Putrid Refrain	Carla Bley – Musique Mecanique I
John Zorn – Forbidden Fruit	Björk – Crystalline (Matthew Herbert Remix)
Liadov – Baba Yaga	Le Poème Harmonique – Una musiqua
Don Ellis – Niner Two	Panzerballett – Birdland
Ligeti - Quatuor à cordes No. 2 - come un meccanismo di precisione	Car Bomb – Black Blood
Arvo Pärt – Ludus du Tabula Rasa	Steps Ahead – Both Sides Of The Coin
Deru – Echoes of Me	King Oliver – Chattanooga Stomp
Jaco Pastorius – Come On, Come Over	Théo Ceccaldi – Amanda Dakota
Bach/ Glenn Gould - The Art of the Fugue, BWV 1080- Contrapunctus III	Slint – Breadcrumb Trail
Laurent DeWilde – Jungle Hard Bop	Lightning Bolt – Dracula Mountain
Dayton – Krackity Krack	Bill Frisell – One Of These Days
Aka Moon – For Drummers Only	The Adolescents – Kids of the Blackhole
Botch - To Our Friends in the Great White North	César Franck – Quatuor en Ré majeur, 1er mouvement
Primus – Tommy The Cat	Daniel Johnston – I Had Lost My Mind
Bruckner - Symphony No. 9 in D Minor, WAB 109 : II. Scherzo	Joy Division – Atmosphere
Naked City – Osaka Bondage	Goto80 – Boys Say Go
Luciano Berio – Visage	Manu Le Malin – On The Way Home
Karlheinz Stockhausen – Gesang Der Juenglinge	Georgette Plana – La Valse Brune
Magic Malik – XP Contrepoint mécanique à six voix	Manolo Sanlucar – El Poeta Pide A Su Amor Que Le Escriba
Steve Lehman – Digital Ambush	Franz Liszt – La Lugubre Gondola
Jakob Bro – Full Moon Europa	Giacinto Scelsi – Quattro Pezzi su una nota sola, 4eme mouvement

Pour cela, les utilisateurs pouvaient écouter autant de fois que nécessaire les extraits et pouvaient les déplacer et les regrouper librement sur l'interface.

2.6.3 Données

Une fois le tri terminé (voir figure 2.32), le programme génère pour chaque volontaire un fichier dans lequel est indiquée la répartition des extraits dans les différentes classes.

Le logiciel génère également un fichier « mouchard » contenant l'historique des opérations effectuées par l'utilisateur : déplacement des icônes et écoute des extraits. Nous pouvons ainsi rejouer toutes les actions effectuées par le volontaire. De plus, le logiciel permet d'effectuer directement une analyse de ces fichiers et ainsi d'extraire plusieurs statistiques sur les participants. La durée moyenne de l'expérience a été de 37 minutes, la durée maximale a dépassé une heure (1h 2min) et la durée minimale fut de 15 minutes. L'écart-type sur la durée de l'expérience est de 10 minutes. En moyenne, les participants ont formé

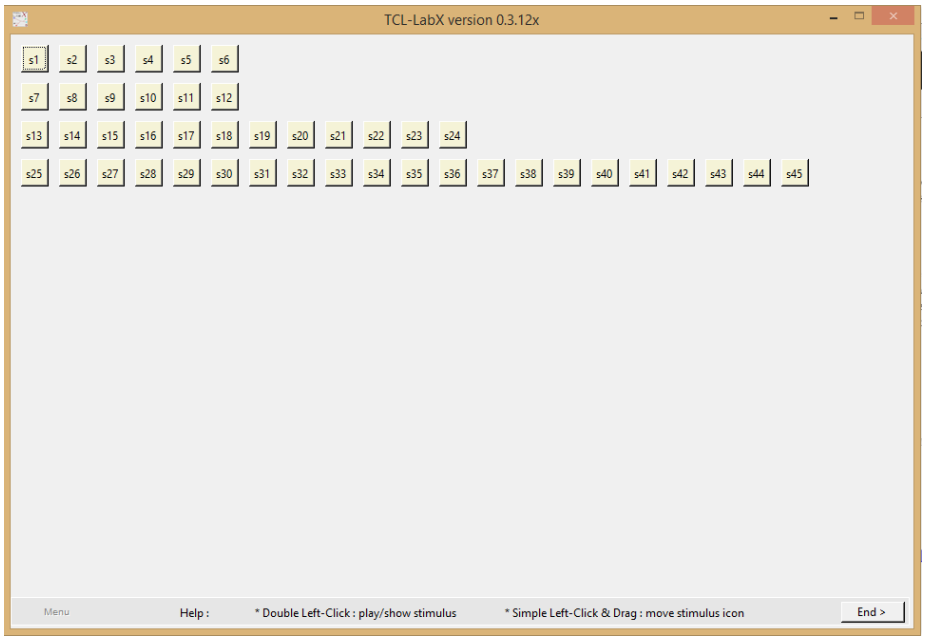

FIGURE 2.31: Interface de l'outil TCL-labX présentée au départ de la passation.

15,5 classes, le minimum est de 8 classes et le maximum 20. L'écart-type sur le nombre de classes est de 3,2.

2.6.4 Analyse des résultats

Pour réaliser cette tâche, nous nous sommes basés sur les travaux de [8] et [84] : il s'agit d'une classification hiérarchie ascendante, expliquée dans la partie 1.4.

Nous avons obtenu le dendrogramme de la figure 2.33.

Il nous renseigne sur les liens établis entre les morceaux par les utilisateurs. L'axe vertical représente la distance entre les morceaux ou groupes de morceaux. Ainsi, deux morceaux ayant été très souvent placés ensemble par les volontaires ont une liaison basse, comme par exemple le 11 et le 17 ou le 15 et le 37.

2.6.5 Interprétation musicologique

Le dendrogramme obtenu précédemment a été interprété d'un point de vue musicologique afin de comprendre comment les volontaires avaient effectué cette classification. Le dendrogramme a été annoté avec les critères communs aux morceaux appartenant à la même branche, voir figure 2.33.

Le nom indiqué sous chaque nœud correspond à un critère musicologique commun à tous les morceaux qui sont situés en dessous. Les quatre grandes

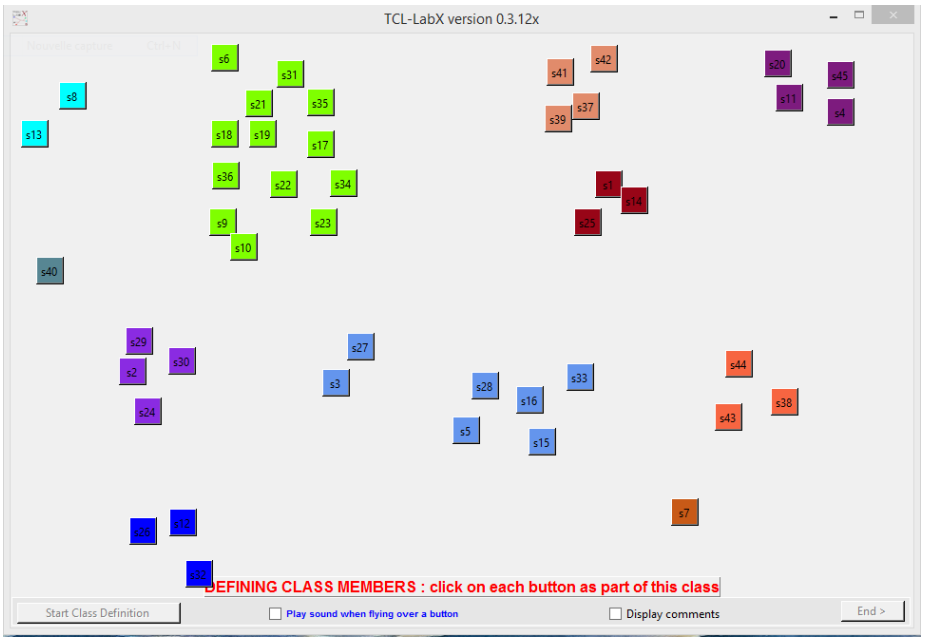

FIGURE 2.32: Interface après le tri.

catégories que l'on retrouve tout en haut de la figure et représentant les quatre grandes classes effectuées par les participants sont décrites ci-dessous.

1. **Audio-tactile** : les musiques présentes dans cette catégorie sont toutes très rythmées et appartiennent au genre jazz ou funk et plus globalement à tendance afro-américaine. L'audiotactilité désigne un rapport particulier avec le corps. À l'intérieur de cette catégorie, les morceaux ont étés distingués par l'instrument prédominant.
2. **Musique savante** : cette catégorie est un hybride regroupant d'une part la Musique Savante Occidentale, d'autre part la musique évoquant un aspect sacré ou spirituel et enfin les morceaux où la voix est prédominante.
3. **Non Conventionnel** : cette catégorie regroupe la musique construite en dehors des règles conventionnelles régissant la musique occidentale notamment basée sur la mélodie et l'harmonie. Nous y retrouvons, par conséquent, les musiques sans hauteur de notes précises et les musiques extra-occidentales. Toutes musiques ne suivant pas les hiérarchies présentes dans les codes occidentaux sont forcément perçues comme un groupe à part.
4. **Électronique et Énergique** : la plupart des extraits de cette catégorie ont étés produits à partir d'instruments et de traitements électroniques. Cette catégorie regroupe également des morceaux présentant des contrastes en termes d'énergie. Cependant, la distance reste particulièrement conséquente entre ces deux sous-catégories.

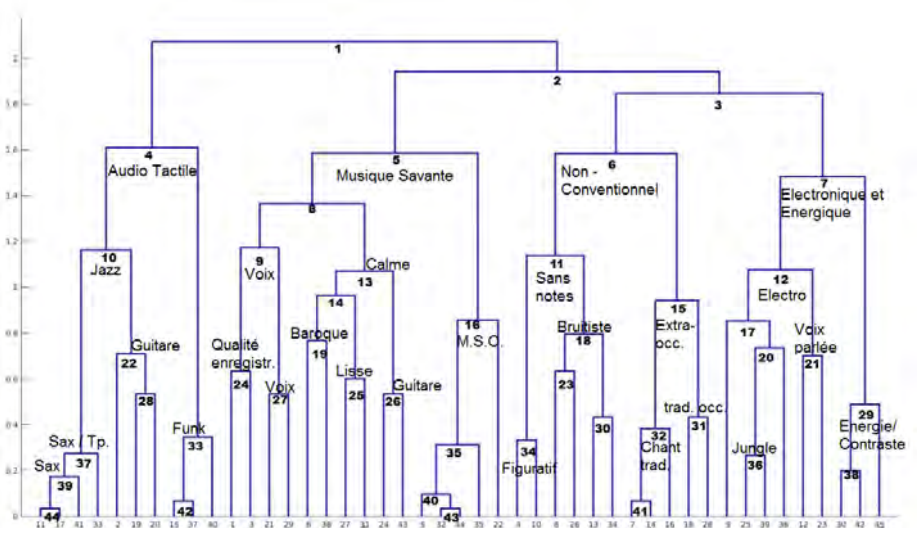

FIGURE 2.33: Dendogramme annoté par les musicologues. L'axe y correspond à la distance entre deux extraits ou groupe d'extraits, l'axe x indique les numéros d'extraits.

L'objectif n'était pas de retrouver dans les résultats de la passation les critères musicologiques définis en début d'expérience, mais de voir quelles dimensions sont les plus prégnantes à l'écoute pour les volontaires. Ainsi, nous pouvons remarquer que les critères musicologiques qui ont servi à établir le corpus ne se retrouvent pas à travers la classification des participants non-experts. Cela illustre différents types d'analyses et d'appréciations musicales. Les critères musicologiques nous ont permis d'obtenir un corpus très varié et les catégories identifiées par les participants révèlent d'autres critères plus accessibles pour des non-experts. C'est sur ces derniers que nous allons nous appuyer pour effectuer une classification automatique.

2.6.6 Sélection des paramètres acoustiques

Dans un premier temps, nous avons calculé 31 paramètres audio (voir figure 2.34) sur chaque morceau du corpus à l'aide de MIR Toolbox [65]. L'extraction de ces paramètres est décrite en détail dans la partie 2.5.2.

Nous avons moyenné chaque paramètre afin d'obtenir une matrice de la forme $N \times P$ avec $N = 45$ (morceaux) et $P = 31$ (paramètres). Notons qu'en ne conservant qu'une moyenne, nous perdons l'évolution temporelle, mais cela nous permet de n'avoir qu'un seul scalaire par morceau et par paramètre. Pour chaque paramètre, nous avons pu calculer la distance pour chaque paire de morceaux et ainsi former une matrice de dissimilarité P^i pour chaque paramètre i. Ensuite, nous avons voulu établir un modèle de la matrice de dissimilarité de la passation

	Paramètre
1	MFCC
2	MFCC
3	MFCC
4	MFCC
5	MFCC
6	MFCC
7	MFCC
8	MFCC
9	MFCC
10	MFCC
11	MFCC
12	MFCC
13	MFCC
14	Taux de passage par zero
15	Rolloff
16	Etalement du spectre
17	Platitude du spectre
18	Entropie du spectre
19	Dissonance Sensorielle
20	Brillance
21	Mode
22	Clarté de la clé
23	Détection de changement d'harmonies
24	Taux de faible énergie (dynamique)
25	Clarté de la pulsation
26	Nombre d'évènements par seconde
27	Tempo
28	Attaque
29	Irregularité
30	Inharmonicité
31	Kurtosis du spectre

FIGURE 2.34: Paramètres extraits et numéros associés.

à partir d'une combinaison linéaire des matrices des paramètres :

$$M_{modele} = \sum_{i=1}^{31} a_i P^i \tag{2.15}$$

Les valeurs contenues dans chaque matrice de dissimilarité ont été normalisées entre 0 et 1 afin de rester cohérentes avec les valeurs de la matrice de la passation. Plutôt que d'utiliser toutes les matrices P^i, nous avons sélectionné les matrices les plus pertinentes en calculant le coefficient de corrélation de chaque matrice de paramètres avec la matrice des volontaires et nous avons sélectionné les N plus corrélées. En effet, les matrices les plus corrélées avec la matrice de dissimilarité établie lors de la passation sont par définition les plus « ressemblantes ».

2.6.7 Régression

Matrice complète

Avec ces N premières matrices, nous avons utilisé un algorithme de descente de gradient afin de trouver la meilleure combinaison linéaire, le critère à optimiser étant l'erreur quadratique entre cette combinaison linéaire et la matrice de

dissimilarité de la passation. Cet algorithme renvoie donc les coefficients a_i par lesquels sont multipliées les matrices de dissimilarité afin d'obtenir la matrice la plus ressemblante à la matrice de dissimilarité formée par l'ensemble des résultats des volontaires. Ces coefficients nous informent de l'importance de chaque paramètre : si un coefficient est faible alors il est peu influent pour les volontaires pour trier les morceaux, et inversement. Afin de simplifier les calculs, les matrices de dissimilarité ont été transformées en vecteurs V_p et V_d de longueur $L = 45 \times 45 = 2025$.

Pour N matrices de dissimilarité de paramètres utilisées, l'équation de l'erreur quadratique est définie par :

$$J = \sum_{j=1}^{L} \left[\left(\sum_{i=1}^{N} a_i V_p^{i,j} \right) - V_d^j \right]^2 \tag{2.16}$$

Le gradient de cette erreur est :

$$\overrightarrow{\text{grad}}\ J = \begin{bmatrix} \frac{\partial J}{\partial a_1} \\ \dots \\ \frac{\partial J}{\partial a_k} \\ \dots \\ \frac{\partial J}{\partial a_N} \end{bmatrix} \tag{2.17}$$

avec :

$$\begin{array}{ll} \frac{\partial J}{\partial a_k} & = \sum_{j=1}^{L} \frac{\partial \left[\left(\sum_{i=1}^{N} a_i V_p^{i,j} \right) - V_d^j \right]^2}{\partial a_k} \\ & = 2 \sum_{j=1}^{L} \left[\left[\left(\sum_{i=1}^{N} a_i V_p^{i,j} \right) - V_d^j \right] V_p^{k,j} \right] \end{array} \tag{2.18}$$

Nous avons testé l'algorithme avec les N « meilleurs » paramètres, au sens de la corrélation. En augmentant successivement le nombre de paramètres N, l'erreur quadratique totale diminue jusqu'à 220. À partir de 7 paramètres, l'erreur augmente de nouveau. Les 6 premiers paramètres sont : *Irrégularité, Brillance, Rolloff, Détection de changement d'harmonie, Entropie du spectre, Attaque.*

À partir de la matrice estimée avec 6 paramètres, nous avons généré un nouveau dendrogramme (cf. figure 2.35) afin de pouvoir comparer visuellement le résultat de cette estimation avec le dendrogramme obtenu à l'issue de la passation (cf. figure 2.33) : nous pouvons constater une certaine dissemblance entre les deux dendogrammes.

Ceci peut s'expliquer par le fait que les participants n'ont pas utilisé la même « règle » pour classer tous les extraits. Par exemple, certains extraits ont été groupés par rapport à des similarités rythmiques, et d'autres vis-à-vis de leur mélodie. Ainsi il est difficile d'établir une règle générale sur les paramètres pour estimer avec une bonne précision la classification globale. Cela est sûrement dû au fait que les extraits présentaient une grande variabilité. De plus, les participants ont également une perception et une expérience différente de la musique, qui ammène à des catégorisations différentes. Cependant, si nous considérons des sous-parties du dendrogramme, les extraits appartenant à chacune de ces sous-parties sont plus similaires entre eux et nous pouvons donc supposer qu'il

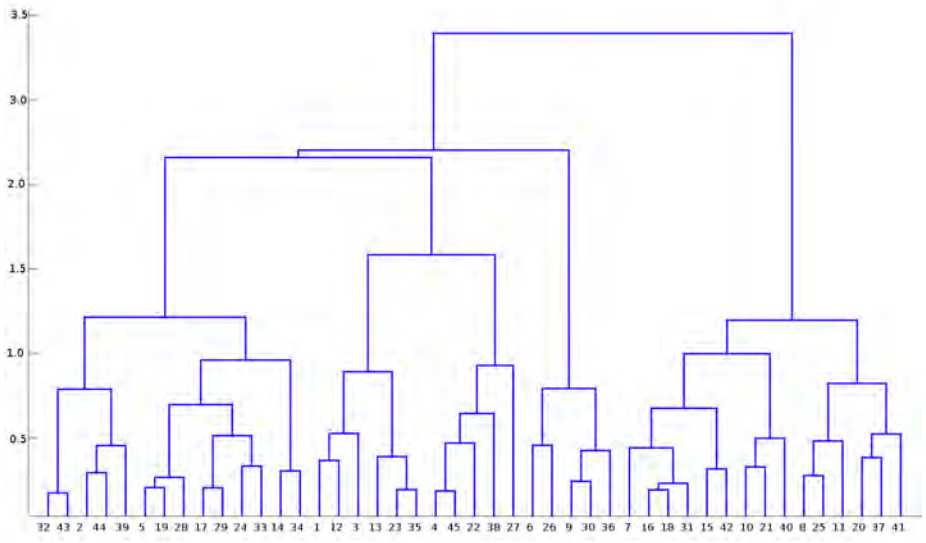

FIGURE 2.35: Dendrogramme estimé à partir de 6 paramètres.

sera plus aisé d'isoler des paramètres discriminants pour chacune de ces sous-parties.

Sous matrices/parties

Pour chacune des parties nommées par les musicologues, Audio-Tactile, Musique Savante, Non conventionnel et Électronique/Énergique, nous avons recalculé les critères les plus corrélés, et les avons utilisés à nouveau pour former des combinaisons linéaires. Nous avons utilisé la même méthode que précédemment afin de retracer les dendrogrammes correspondants à chacune de ces parties.

1. **Audio-Tactile** Les paramètres les plus corrélés sont : *Brillance, Irrégularité, MFCC(10), Attaque, MFCC(4), Entropie du spectre, Clarté de la pulsation, Taux de passage par zéro, Low energy, Kurtosis du spectre, MFCC(3), Rolloff, Tempo, MFCC(8).*

   À l'issue de la descente de gradient, l'erreur quadratique totale est de 5,5. Nous remarquons que plusieurs MFCC interviennent dans la classification. Nous pouvons expliquer cela par le fait que pour cette catégorie, les volontaires ont beaucoup distingué les extraits selon les instruments prédominants. Le dendrogramme a été bien reconstitué, mis à part pour les extraits 5 et 6 qui ont été échangés.

2. **Musique Savante** Pour cette catégorie, les résultats étaient plutôt mitigés, même en utilisant tous les paramètres. En effet, à l'issue de la descente de gradient, l'erreur quadratique totale est de 9,5. Ces résultats

plus faibles peuvent s'expliquer par le fait que cette catégorie contient des extraits disparates, il est donc plus difficile de généraliser une règle de catégorisation issue uniquement des données acoustiques des extraits.

3. **Non conventionnel** Ici, la méthode a fourni de bons résultats avec 18 paramètres. À l'issue de la descente de gradient, l'erreur quadratique totale est de 3,8. Les paramètres les plus corrélés sont : *Dissonance sensorielle, Attaque, Nombre d'événements par seconde, Taux de passage par zéro, Kurtosis du spectre, Clarté de la clé, Entropie, MFCC(8).*
4. **Électronique/Énergique** C'est pour ce groupe que la méthode a été la plus performante cf. figure 2.36).

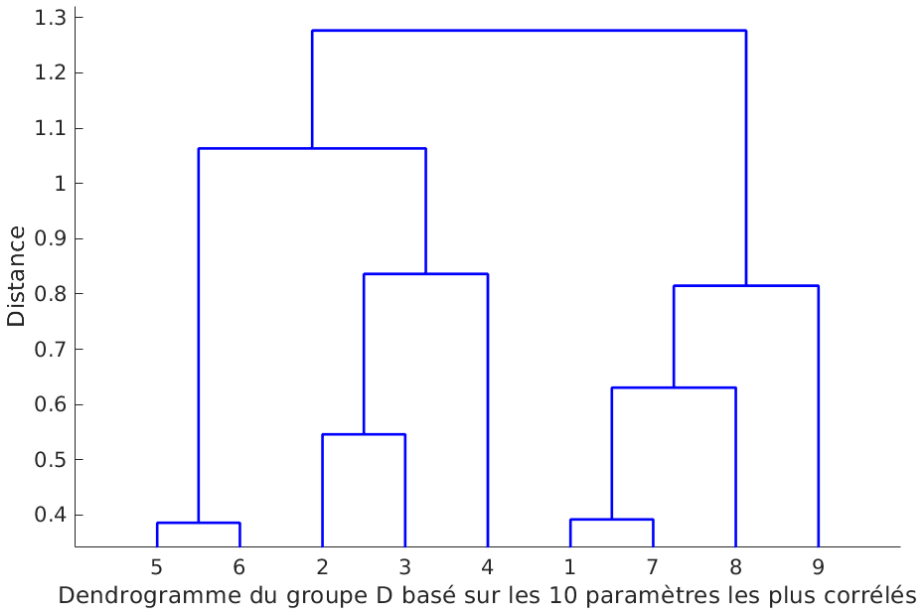

FIGURE 2.36: Dendrogramme obtenu pour la catégorie « Electronique/Energique » avec 10 paramètres. Excepté pour le premier extrait, le dendrogramme de ce groupe a été bien reconstitué. Pour chaque sous-groupe i de taille N_i, les indices initiaux des morceaux ont été remplacés par des indices allant de 1 à N_i. Ainsi, si le dendrogramme d'un sous-groupe a bien été reconstitué, les individus sont placés dans l'ordre croissant.

Nous avons obtenu une erreur quadratique totale de 2,8. Nous avons utilisé les 10 paramètres suivants (du plus corrélé au moins corrélé) : *Attaque, MFCC(3), MFCC(8), MFCC(11), Rolloff, Brillance, MFCC(4), Taux de passage par zéro, MFCC(0) (i.e. énergie).*

Globalement, les résultats sont plutôt satisfaisants, car nous avons ainsi pu reconstituer les dendrogrammes de chacune des catégories avec un nombre d'er-

reurs limité.

2.6.8 Classification de nouveaux extraits

L'objectif de cette partie était de trouver une méthode permettant d'attribuer à un « nouvel » extrait la bonne catégorie. Dans toutes les méthodes qui suivent, nous avons considéré successivement chaque extrait comme un nouvel individu, en prenant soin de le retirer de la base d'apprentissage. Le score de chaque méthode est donc compris entre 0 et 45 (cas où tous les extraits ont été attribués aux bonnes catégories). De plus, les paramètres ont été centrés réduits afin de supprimer l'influence de l'unité de mesure utilisée pour chacun d'eux.

La première méthode a consisté à calculer le barycentre de chaque catégorie selon les 31 paramètres, et à attribuer ensuite le nouvel extrait à la classe ayant son barycentre le plus proche : nous avons ainsi obtenu 30 attributions correctes (soit environ 67%).

Dans la seconde méthode, nous avons retenu un nombre restreint de paramètres : nous avons observé quels paramètres étaient pertinents pour la classification dans les sous-catégories (section 2.6.7) et nous avons conservé seulement ceux qui étaient les plus corrélés dans les 4 classifications. Il s'agit de *Détection de changement d'harmonie, Attaque, Entropie du spectre, Rolloff MFCC(0) et Irrégularité*. Nous avons ainsi obtenu 22 attributions correctes (33%) : les paramètres sélectionnés n'étaient donc pas particulièrement pertinents.

Dans la troisième méthode, nous avons utilisé les N paramètres les plus corrélés en établissant le classement de la même manière que dans la section 2.6.7. Sur la figure 2.37, nous voyons que le score augmente globalement avec le nombre de paramètres, mais qu'il diminue parfois lorsque nous en utilisons un nouveau. Le score maximal (32 attributions correctes, soit 71%) est atteint pour 25 paramètres. Cette méthode s'est donc avérée être la meilleure.

2.6.9 Conclusion

L'analyse des résultats de la passation nous a permis d'établir une classification moyenne des morceaux par les volontaires qui a été représentée sous la forme d'un dendrogramme dans lequel apparaissent quatre groupes ainsi que des sous-groupes.

Afin de reconstruire de manière automatique cette classification humaine, nous avons établi une hiérarchie dans la pertinence des paramètres selon leur corrélation avec la classification des volontaires. Nous avons vu que cette reconstruction automatique est plus efficace pour distinguer les sous-groupes à l'intérieur d'un groupe, plutôt que les groupes entre eux. Au final, les paramètres identifiés pourront être privilégiés pour une application en recommandation musicale.

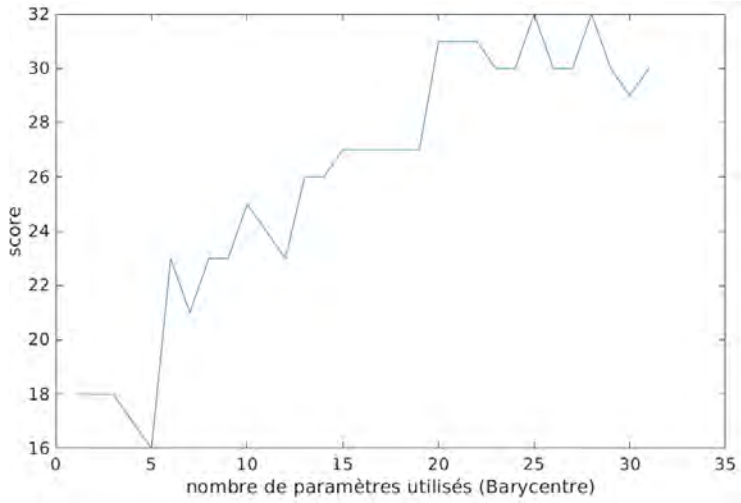

FIGURE 2.37: Score d'attribution en fonction du nombre de paramètres utilisés.

2.7 Avantages et défauts de l'approche basée sur le contenu

2.7.1 Avantages

L'approche basée sur le contenu permet d'éviter le problème du démarrage à froid pour les morceaux, puisqu'ils n'ont pas besoin d'avoir été écoutés pour être recommandés. De plus, le nombre d'écoutes des morceaux n'entre pas en compte dans cette méthode, qui n'a pas tendance à favoriser des morceaux populaires, contrairement au filtrage collaboratif. Cette caractéristique peut néanmoins constituer un défaut pour certains auditeurs en quête uniquement de musique populaire. Si la recommandation est utilisée dans le but d'obtenir un morceau **similaire** selon des paramètres identifiables, alors la méthode basée sur le contenu peut être privilégiée.

Enfin, un des avantages majeurs des méthodes basées sur le contenu demeure dans son potentiel d'explicabilité et de contrôle de l'utilisateur sur la recommandation. Des paramètres identifiés par le système comme particulièrement pertinents pour la recommandation pourront être utilisés pour justifier celle-ci, et ainsi aiguiller l'utilisateur dans ses choix.

Par exemple, si un utilisateur a écouté de nombreux morceaux de rock des années 80, l'algorithme pourra lui proposer un morceau appartenant à cette période, un morceau rock sans regard sur la période, ou bien un morceau rock des années 80. En indiquant les paramètres (ici période et/ou genre) qui ont motivé le choix de l'algorithme, l'utilisateur pourra alors profiter de cette explication pour motiver son choix final. Le choix du morceau à écouter sera ainsi issu

d'une collaboration entre l'algorithme, qui aura sélectionné 3 morceaux parmi une base de données en contenant plusieurs millions, et l'utilisateur, qui a le dernier mot sur la sélection finale. Son choix permettra par exemple de s'adapter à ses goûts, au contexte, ou bien à un désir de variété vis-à-vis des morceaux précédemment écoutés, qui sont des paramètres difficilement identifiables par un système automatique. De même, si un utilisateur possède des goûts variés regroupés en différents clusters (voir partie 2.4.2), l'algorithme pourra proposer plusieurs recommandations correspondant aux différents centres d'intêrét musicaux.

2.7.2 Défauts

Le problème du démarrage à froid subsiste pour les nouveaux utilisateurs puisqu'ils n'ont pas encore pu expliciter leurs goûts. Cependant, dans le cadre de la recommandation personnalisée ce problème est par définition irrésolvable en l'absence de données *a priori* sur l'utilisateur. Dans certains cas, des informations sur l'utilisateur tels que l'âge, des critères démographiques [49], ou bien encore des liens sur les réseaux sociaux [93] permettent d'établir des similarités entre de nouveaux utilisateurs et des anciens aux goûts mieux connus par la plateforme. Ces similarités peuvent être utilisées à des fins de recommandation dans un premier temps, en recommandant aux nouveaux utilisateurs, des morceaux aimés par d'anciens utilisateurs au profil similaire.

De plus, il peut être attendu d'un système de recommandation qu'il suggère des éléments complémentaires, et non pas similaires. Or, une approche basée uniquement sur la distance minimale selon le contenu aura tendance à recommander les morceaux les plus proches possibles du morceau cible. Cela peut donc poser un problème vis-à-vis de la diversité de la recommandation. Par exemple, il a été montré dans [48] que des systèmes basés sur la similarité acoustique peuvent avoir tendance à recommander des morceaux du même artiste, voire du même album. Néanmoins, ce cas extrême pourrait facilement être évité à l'aide d'un filtre.

2.8 Comment évalue-t-on les performances de la recommandation de musique ?

Nous avons vu que les systèmes de recommandation de musique sont des algorithmes d'apprentissage automatique basés sur des données de feedback des auditeurs parfois explicites, souvent implicites. Dans le cas de la recommandation basée sur le contenu, les morceaux sont représentés à partir de descripteurs provenant d'annotations manuelles ou automatiques.

Nous allons à présent nous intéresser aux différents modes d'évaluation de la recommandation de musique.

2.8.1 Evaluation OFFLINE

Une évaluation est dite « hors-ligne » lorsqu'un algorithme est testé sur un jeu de données statiques. Certains jeux de données comme The Million Song Dataset [13], LFM-1b [95], ou bien encore The Music Streaming Sessions Dataset [22] ont été mis à disposition du public afin de favoriser la recherche dans le domaine de la recommandation. Ces jeux de données contiennent des feedbacks implicites ou explicites associés à des utilisateurs, sur un grand nombre de morceaux. Des descripteurs acoustiques ou des métadonnées sont parfois associés aux morceaux.

Les modèles sont entraînés sur une partie de ces données, le restant est utilisé pour le test. Connaissant un nombre donné d'interactions entre les utilisateurs et les morceaux de musique, le but est de prédire les interactions contenues dans le jeu de test. Ici, ce n'est donc pas la capacité des modèles à effectuer de bonnes recommandations qui est évaluée, mais plutôt une estimation d'une partie « manquante » des données. Le terme de prédiction peut également être employé, en gardant cependant à l'esprit que les informations ainsi prédites sont seulement occultées et non pas futures.

Des métriques pertinentes sont nécessaires afin d'évaluer les performances des différents systèmes de recommandation et de les comparer entre eux, dans le but de désigner le plus performant. Si nous considérons un ensemble de N prédictions $\hat{R} = \{\hat{r}_1, ..., \hat{r}_i, ..., \hat{r}_N\}$ et une *vérité terrain* R, les métriques les plus communément utilisées pour évaluer les performances d'un système de recommandation sont les suivantes.

Erreur Absolue Moyenne (ou Mean Absolute Error, MAE)

L'erreur absolue moyenne est la moyenne des valeurs absolues de l'erreur commise sur chaque estimation.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |(\hat{r}_i - r_i)| \tag{2.19}$$

Cette métrique est l'une des plus utilisées pour évaluer les systèmes de recommandation et correspond à la moyenne de la valeur absolue des écarts entre les variables prédites et la vérité terrain [57]. La MAE est une métrique simple qui indique à quel point les prédictions sont proches de la réalité.

Erreur Quadratique Moyenne, EQM (ou Mean Squared Error, MSE)

L'erreur quadratique moyenne est la mesure de précision d'un estimateur en statistique. L'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ peut s'exprimer en fonction de son biais et de sa variance :

$$EQM(\hat{R}) = Var(\hat{R}) + Biais(\hat{R})^2 \tag{2.20}$$

En statistiques, un estimateur idéal doit avoir un biais nul et une variance minimale. L'erreur quadratique moyenne permet donc de mesurer la qualité d'une prédiction, au sens statistique du terme.

$$EQM = \frac{1}{N}\sum_{i=1}^{N}(\hat{r}_i - r_i)^2 \tag{2.21}$$

Root-Mean-Square Error, RMSE

Il s'agit de la racine carrée de l'EQM :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{r}_i - r_i)^2} \tag{2.22}$$

Le fait de calculer la racine carrée de la moyenne a pour effet de pénaliser davantage les plus fortes erreurs qu'avec l'erreur absolue moyenne. Ainsi, l'usage de la métrique RMSE est pertinent dans un cas où une somme de « petites erreurs » est moins grave qu'une « grosse erreur ».

Par exemple, pour un système de recommandation chargé de prédire la note entre 0 et 5 donnée par un utilisateur à plusieurs morceaux, nous pouvons considérer que prédire un 4 au lieu d'un 1 sur un seul morceau est pire qu'un petit écart récurrent sur tous les morceaux, voir table 2.2.

TABLE 2.2: Différence entre l'erreur absolue moyenne et la RMSE

	morceau 1	morceau 2	morceau 3	RMSE	MAE
vérité terrain	0	5	4		
prédictions A	1	4	3	$\sqrt{3}$	3
prédictions B	2	4	4	$\sqrt{5}$	3
prédictions C	0	5	1	3	3

Précision au rang K (ou P@K)

Les métriques décrites précédemment sont calculées sur l'ensemble des N estimations données par un modèle. Compte tenu de la taille des catalogues, les recommandations n'ont généralement pas besoin d'être exhaustives : seuls les morceaux qui ont les plus fortes probabilités de plaire seront présentés à un auditeur. Il peut donc s'avérer plus pertinent de calculer la précision au rang K, c'est-à-dire de calculer la précision sur les K prédictions ayant les notes estimées r les plus élevées pour un utilisateur donné. Il faut donc ordonner les notes prédites par ordre décroissant, et calculer la proportion de vrais positifs (ou True Positives, TP) parmi le nombre de prédictions.

$$Precision@K = \frac{TP}{K} \tag{2.23}$$

Normalized Discounted Cumulative Gain (ou NDCG)

Les métriques précédentes ne tiennent pas compte du rang de chaque proposition. Or, lorsque nous proposons K morceaux, une erreur sur un morceau proposé en premier peut être considérée comme plus pénalisante qu'une erreur en bas de classement. Ici, la notation rel_i (pour "relevance" en anglais) nous indique la pertinence du morceau recommandé. Dans un cas binaire de type « j'aime/je n'aime pas », si le morceau proposé i est pertinent, $rel_i = 1, 0$ sinon. La pertinence de chaque morceau est pondérée par l'inverse de son rang dans l'ensemble des recommandations, sur une échelle logarithmique binaire. L'échelle logarithmique a l'intérêt de croître plus lentement qu'une échelle linéaire. Par exemple, la pertinence d'un morceau au rang 31 sera pondérée d'un facteur 5.

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \tag{2.24}$$

Ainsi, la métrique DCG (équation 2.24) nous donne, pour un classement de p morceaux, un score qui tient compte de la position de chaque morceau.

Notons que ce score dépend du nombre d'éléments dans le classement. Si nous voulons comparer deux classements n'ayant pas le même nombre d'éléments, nous pouvons normaliser le DCG par le DCG idéal $IDCG$. Dans un cas où la pertinence est binaire ($rel_i = 0, 1$), nous avons :

$$IDCGp = \sum_{i=1}^{|REL_p|} \frac{1}{\log_2(i+1)} \tag{2.25}$$

Où $|REL_p|$ désigne le nombre d'éléments pertinents.

La normalisation consiste à diviser le score DCG par celui obtenu dans le cas d'un classement parfait (équation 2.26). Nous obtenons un score $IDCG$ compris entre 0 - dans le cas où aucun élément recommandé n'est pertinent - et 1 - dans les cas où tous les éléments pertinents seraient classés en premiers. Cette métrique permet donc de comparer la qualité de 2 classements de tailles différentes.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{2.26}$$

Une étude théorique de la performance de cette métrique a été réalisée par Y. Wang en 2013 dans [42].

Il existe d'autres métriques utilisées pour la recommandation de musique allant au-delà de la simple mesure de précision. Nous pouvons notamment citer l'étalement, la couverture, la nouveauté, la sérendipité ainsi que la diversité. Plusieurs publications donnent un état de l'art détaillé des métriques utilisées pour l'évaluation hors-ligne des algorithmes de recommandation de musique [100, 10].

Bien que l'évaluation soit portée sur un jeu de données hors-ligne, les métriques utilisées restent néanmoins pertinentes vis-à-vis d'un scénario de recommandation.

2.8.2 Evaluation ONLINE

Les métriques « online » sont calculées directement par les plateformes de streaming musical à partir des données d'utilisation de l'application. In fine, pour les plateformes le but des systèmes de recommandation est de fidéliser ses utilisateurs. C'est donc ces grandeurs qui sont mesurées lors du déploiement de nouveaux algorithmes de recommandation, comme me l'a confirmé Manuel Moussallam, directeur de la R&D chez Deezer :

1. le temps passé sur l'application,
2. le ratio écoute/skip,
3. le taux de reconnexion, c'est-à-dire le pourcentage des utilisateurs qui reviennent sur l'application dans les 7 jours.

Dans [102], l'auteur montre que les algorithmes de recommandation se sont petit à petit éloignés de la prédiction de goûts pour tendre davantage vers une sorte de « piège » devant retenir le plus longtemps possible l'auditeur sur la plateforme. Les métriques présentées précédemment sont qualifiées de « captivation metrics », servant à mesurer la rétention des utilisateurs des plateformes.

2.9 Conclusion

Dans ce chapitre, nous avons étudié la recommandation de musique en tant qu'application de l'apprentissage automatique.

Toutes les méthodes employées ne modélisent pas l'humain de la même manière. En filtrage collaboratif le goût modélisé est un goût relatif à celui des autres utilisateurs. Dans les méthodes basées sur le contenu, les simples mesures de similarité entre les morceaux ne permettent pas de modéliser les goûts de l'utilisateur. Certaines approches basées sur des méthodes non supervisées permettent néanmoins de modéliser des goûts variés à travers plusieurs clusters. Par ailleurs, nous avons proposé une expérience permettant de lier des paramètres acoustiques à une catégorisation effectuée par des volontaires. Cette expérience s'inscrit dans la démarche de donner plus de sens aux paramètres acoustiques, qui sont parfois trop éloignés de la perception humaine de la musique.

Les méthodes employées nécessitent des données d'apprentissage, une vérité terrain, qui est bien souvent éloignée du goût des utilisateurs. Les feedback les plus abondants sont en général des indications comportementales. Ainsi les algorithmes de recommandation de musique tendent davantage à prédire des comportements que des goûts. L'action Skip notamment, est interprétée négativement par les plateformes qui cherchent à la prédire afin de la réduire au maximum. Avec les algorithmes de recommandation, les plateformes ont pour objectif de maximiser la durée d'utilisation des applications de streaming tout en maintenant une écoute la plus passive possible.

Les conclusions tirées de cette étude nous guident dans nos choix sur les conditions expérimentales de l'expérience présentée dans la partie suivante, où nous veillons à maintenir le cap vers une véritable prédiction de goût.

Chapitre 3

Etude qualitative d'une prédiction de goûts musicaux

Dans cette partie, nous présentons une expérience de prédiction automatique de goûts musicaux basée sur le signal audio. Ces travaux ont donné lieu à une présentation lors de la conférence CBMI 2021 [37] .

Le chapitre précédent nous a permis de poser des exigences pour l'expérience présentée dans ce chapitre, à savoir :

- Privilégier des feedbacks explicites pour connaître les goûts de volontaires. Ainsi, la tâche de prédiction de goût peut s'apparenter à une tâche de classification à 2 classes : « Like / Dislike ».
- Ne pas représenter le signal audio sous la forme d'un nombre fini de paramètres acoustiques, mais utiliser plutôt une représentation la plus exhaustive possible : le spectrogramme.
- Utiliser des modèles uniques pour chaque utilisateur.

Pour répondre à ces contraintes, nous avons dû constituer un corpus de données.

À partir des informations sur les affinités des volontaires, le but de cette expérience est d'entraîner un système capable de prédire leurs goûts musicaux. Chaque personne ayant des goûts différents, nous avons modélisé les goûts de chaque volontaire par des réseaux de neurones convolutionnels (voir CNN partie 2.1.4) uniques, pré-entraînés pour la reconnaissance de genres. La quantité de données disponible pour chaque volontaire étant faible, nous avons eu recours à une technique de Transfer Learning. Les modèles utilisés ici ont été pré-entraînés sur une tâche de classification en genres de musique. Ainsi, cette expérience nous permettra de vérifier l'hypothèse selon laquelle des paramètres acoustiques appris automatiquement par un CNN pour la classification en genre permettent également de prédire des goûts musicaux. Nous avons recouru à deux méthodes de test différentes afin d'évaluer les performances des modèles : évaluation automatique, et évaluation humaine. Nous avons ensuite analysé les résultats obtenus afin de tirer d'une part des conclusions sur l'efficacité des modèles, et d'autre

part sur la consistance des différentes méthodes de test. Enfin, des pistes d'amélioration ont été proposées pour corriger les défauts des modèles.

3.1 Corpus

Il existe plusieurs corpus fréquemment utilisés en recommandation de musique tels que « Echo Nest taste profile subset » [13] et « LFM-1b » [95]. Néanmoins ces corpus ne proposent ni l'audio des morceaux concernés, ni de feedback *explicite* des utilisateurs sur les morceaux, seulement des statistiques sur leur écoute. Dans notre cas, nous voulons avoir à disposition les enregistrements des morceaux afin de pouvoir extraire les paramètres de notre choix, ainsi qu'une appréciation explicite des utilisateurs sur les morceaux. Nous avons ainsi mené une expérience avec des utilisateurs volontaires de la plateforme Deezer afin de collecter des données correspondant à nos besoins.

Nous avons demandé aux volontaires de naviguer librement sur le site sur une période de plusieurs mois via leur ordinateur et de remplir quand ils le désiraient une des 4 playlists constituées pour l'occasion, correspondant respectivement à 4 affinités différentes :

1. Je déteste, symbolisé par « - - »
2. Je n'aime pas, symbolisé par « - »
3. J'aime, symbolisé par « + »
4. J'adore, symbolisé par « ++ »

Une fois la phase d'utilisation « libre » de la plateforme terminée, nous avons demandé aux volontaires d'écouter une playlist constituée par les musicologues Paul Albenge et Ludovic Florin. Les volontaires pouvaient librement classer les morceaux de cette playlist. Cette playlist a été constituée dans le but de couvrir le plus largement possible les différents critères musicologiques définis précédemment, à travers 100 morceaux. Davantage d'informations sont données sur ces morceaux et sur les critères musicologiques dans la partie 2.6.1.

Nous avons récolté les playlists constituées par 20 volontaires différents, âgés de 18 à 55 ans. Le nombre de morceaux classés par volontaire varie de 77 à 255. Pour chaque volontaire, les playlists « - - » et « - » ont étés fusionnées en une playlist « Dislike » et les playlists « + + » et « + » ont étés fusionnées en une playlist « Like » pour cette tâche de classification. Cette fusion a été effectuée dans le but d'obtenir davantage d'exemples pour chaque classe, et de simplifier la tâche de classification à un problème binaire. Cependant, les 4 catégories d'origine ont été également conservées pour d'éventuelles expérimentations futures.

Nous avons observé que les volontaires ont souvent composé leurs catégories indépendamment de genres musicaux. En effet, pour le même participant, nous pouvons retrouver des morceaux de musique du même genre musical dans les 2 catégories « Like » et « Dislike ». Néanmoins, ces morceaux se distinguent selon d'autres critères, pas nécessairement liés à un genre musical. Nous avons constaté que la playlist musicologique avait été utile pour permettre aux utilisateurs de

trouver des morceaux qui ne leur plaisaient pas, grâce à sa diversité. Toutes les playlists constituées par les utilisateurs sont disponibles sur la page github du projet[1].

3.2 Principe du Transfer Learning

Le Transfer Learning est utilisé dans le cas où des méthodes d'apprentissage profond, qui nécessitent une grande quantité de données d'entraînement, doivent être appliquées à une tâche où peu de données sont disponibles. Un modèle peut être entraîné sur une tâche dite *source* pour laquelle les données sont abondantes. Si cette tâche source est similaire à la tâche *cible*, alors le modèle pourra être réutilisé sur les données cible [81]. Si la tâche source est identique à la tâche cible, alors le modèle pourra directement être réutilisé sur les nouvelles données. Si la tâche cible diffère, alors le modèle devra être réentraîné sur des données cible. Le modèle peut être réentraîné en partie ou bien en totalité. Dans le second cas, le pré-entraînement sur la tâche source servira d'initialisation pour le second entraînement sur la tâche cible.

Un réseau de neurones convolutionnel peut se décomposer en 2 parties (voir section 2.1.4) :

— les couches de convolution, qui extraient automatiquement dans le signal les paramètres pertinents pour la classification.

— les couches denses, qui jouent le rôle de classifieur à partir des paramètres extraits par les couches de convolution.

Si nous faisons l'hypothèse que les paramètres pertinents pour la tâche cible sont identiques à ceux de la tâche source, alors seuls les paramètres de la couche dense sont réentraînés.

Dans notre cas, c'est la faible quantité de données disponibles pour certains volontaires qui a motivé l'utilisation d'une méthode de Transfer Learning. Par la même occasion, cela nous permet de vérifier l'hypothèse suivante : « Les paramètres extraits automatiquement par un réseau de neurones convolutionnel pour la classification en genres sont-ils pertinents pour la prédiction de goûts ? ». Ici, la tâche source est la classification en genres de musique, et la tâche cible est la prédiction de goûts (voir table 3.1). Dans les deux cas, les données utilisées sont des morceaux de musique. Les extraits utilisés en entrée du réseau de neuronnes pour les données source et cible devront avoir la même durée et la même fréquence d'échantillonage.

TABLE 3.1: Données, Tâches, Source, Cible.

	Source	Cible
Données	**GTZAN, FMA**	**Playlists volontaires**
Tâche	Classification en genres	Prédiction de goûts

Nous avons vu précédemment (voir partie 1.3.2) qu'une annotation en genres

1. https://github.com/Nicolas-DBN/Volunteers_playlists

en un nombre limité de classes ne permet pas d'expliquer avec précision les goûts musicaux. Cependant, les paramètres extraits dans les couches intermédiaires d'un réseau de neurones convolutionnel étant de plus bas niveau, ils pourraient constituer une représentation latente pertinente pour la prédiction de goûts.

Ainsi, nous avons entraîné des modèles sur différents corpus de classification en genres, puis réentraîné les couches denses des modèles sur la tâche cible de prédiction de goûts.

3.3 Architecture des CNN

Nous avons utilisé une architecture de CNN similaire à celles présentées dans [129] et [103], en raison de leurs performances au niveau de l'état de l'art pour la classification en genres de musique.

L'entrée du réseau de neurones est une matrice de dimensions 513x128, correspondant à des spectrogrammes de 3 secondes.

Les 256 premiers filtres de convolution ont une hauteur égale au nombre de bandes de fréquences (ici 513) du spectrogramme d'entrée. La longueur de ces filtres est de 4, soit environ 200 ms. Un max-pooling de dimension 2 est ensuite appliqué dans le sens de la longueur (domaine temporel). À nouveau, 128 filtres de convolution (dimension 1×4) sont appliqués, suivis par le même max-pooling que précédemment.

Un pooling moyen et max pooling de dimensions 26 sont ensuite appliqués dans le domaine temporel, suivi par 2 couches denses de dimension 300 et 150. Chaque couche dense a un dropout de 40% pour éviter le sur entraînement, et est suivie par une couche de ReLU (Rectifier Linear Unit).

La sortie du modèle est une couche dense, de longueur égale au nombre de classes à prédire (10 genres pour GTZAN, 8 pour FMA), avec une fonction softmax.

Puisque l'entrée du réseau est un extrait de 3 secondes, un vote majoritaire est effectué sur tous les extraits d'un même morceau afin de donner une prédiction pour le morceau en entier. La prédiction pour le morceau de musique correspond à la classe majoritaire parmi les prédictions sur chacun des extraits.

La figure 3.1 illustre l'architecture utilisée pour la prédiction de goûts. Les couches denses entourées en vert sont réentraînées (training enabled), et les couches convolutionnelles entourées en rouge ne sont pas réentraînées (training disabled). Pour la prédiction de goût, la sortie ne comporte plus que 2 classes.

3.4 Pré-traitement

Chaque morceau de musique (mono, 44100 Hz, 16 bits) a été découpé en extraits de 3 secondes, avec une fenêtre glissante de moitié (1,5s). Nous avons ensuite calculé le spectrogramme (voir partie 2.5.2) de chacun de ces extraits à l'aide de la transformée de Fourier à court terme (Fast Fourier Transform, ou FFT) avec des fenêtres de 2048 points (46,4 ms) avec un recouvrement de

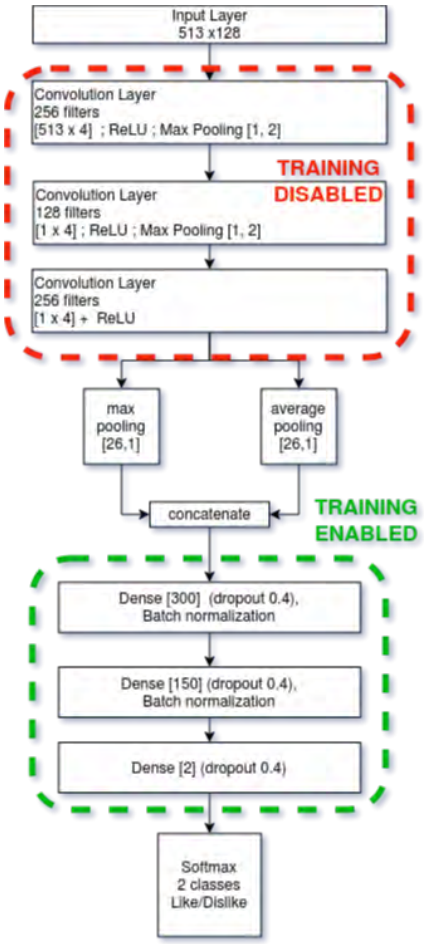

FIGURE 3.1: Illustration du Transfer Learning pour notre prédiction de goûts.

50%. Chaque extrait de 3 secondes est ainsi représenté par un spectrogramme de dimensions 513 x 128. 513 correspond au nombre de bandes de fréquences et 128 au nombre de FFT calculées sur 3 secondes.

3.5 Implementation

Le pré-entraînement et le Transfer Learning ont été implémentés en Python à l'aide de la bibliothèque Keras avec le backend TensorFlow[2]. Un premier entraînement a été effectué d'une part avec le corpus GTZAN et d'autre part avec le corpus FMA, les 2 modèles résultant de ces pré-entraînement ont étés

2. https://keras.io/

sauvegardés. Nous avons ensuite procédé à une nouvelle phase d'entraînement pour chaque volontaire, en utilisant cette fois les données contenues dans les playlists « Like » et « Dislike », et en utilisant les modèles pré-entraînés pour l'initialisation. Différents pas d'apprentissage initiaux ont été testés empiriquement sur les données de validation de chaque volontaire, nous avons conservé le pas présentant les meilleurs résultats moyens : 0,05 pour GTZAN et 0,1 pour FMA.

Durant cette phase d'entraînement, nous avons automatiquement diminué le pas d'apprentissage quand la fonction de coût cessait de diminuer, et l'entraînement était automatiquement stoppé si la fonction de coût continuait de stagner.

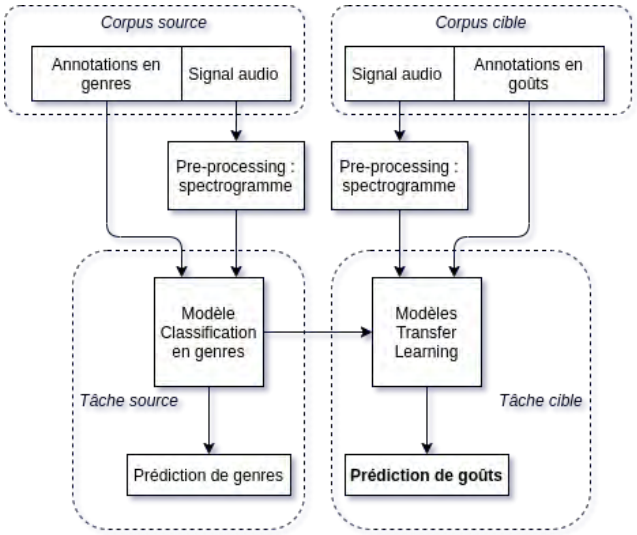

FIGURE 3.2: Prédiction de genre, Transfer Learning et prédiction de goûts.

La figure 3.2 montre un schéma de l'ensemble du processus. Nous avions à disposition un corpus source (GTZAN ou FMA, annotations en genre) et un corpus cible (playlists « Like » et « Dislike »). Après une transformation de l'audio en spectrogramme, un premier modèle source est entraîné à la classification en genre. Le modèle ainsi obtenu est utilisé comme initialisation pour le nouvel entraînement sur la tache cible, la prédiction de goûts.

3.6 Pré-entraînement

Lors du pré-entraînement des modèles, nous avons obtenu des scores similaires à l'état de l'art pour FMA (59%) et GTZAN (85%) ([39] et [129]). Bien que le corpus FMA soit plus volumineux, les scores obtenus sont plus faibles qu'avec GTZAN. Ceci peut s'expliquer par le fait que certains genres de GTZAN étaient représentés par un artiste majoritaire, comme Bob Marley pour le

reggae. Un artiste ayant une forte « empreinte acoustique » à travers sa voix et son style, rend la reconnaissance des genres en question plus aisée. Une analyse menée dans [109] montre d'ailleurs les défauts de ce corpus en termes de répétitions notamment.

Nous avons réalisé des projections des morceaux des corpus FMA (figure 3.3) et GTZAN (figure 3.4) à l'aide de la méthode t-SNE [73].

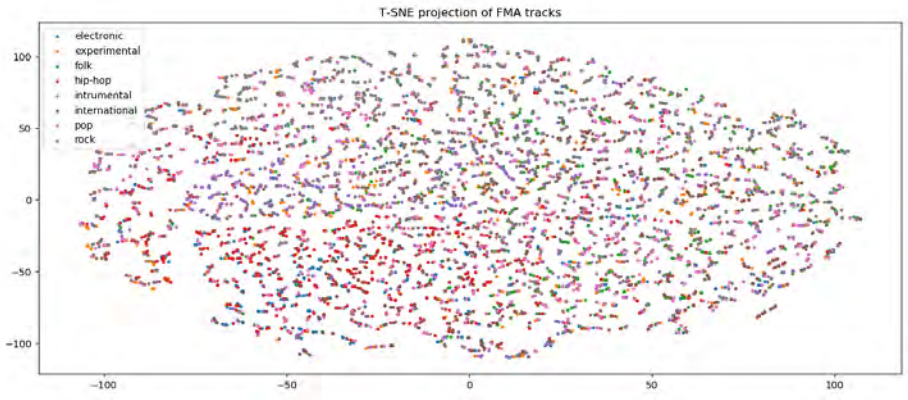

FIGURE 3.3: Projection TSNE des morceaux du corpus FMA.

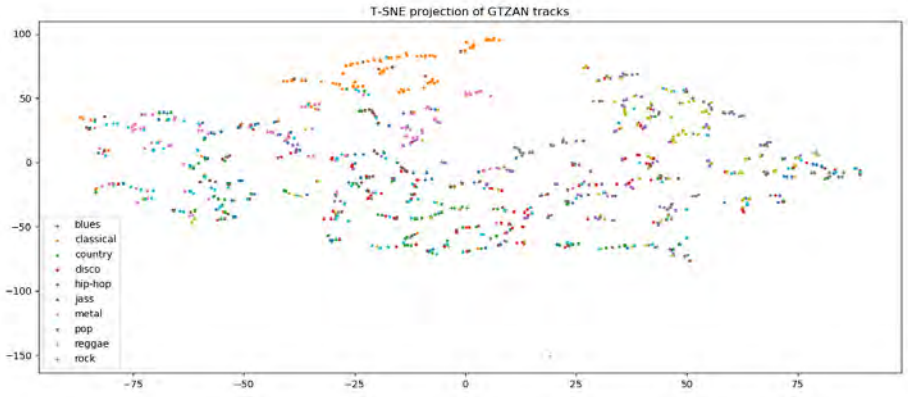

FIGURE 3.4: Projection TSNE des morceaux du corpus GTZAN.

Nous avons réalisé ces projections à partir des statistiques extraites sur les spectrogrammes de ces morceaux. Ces projections témoignent donc de la similarité du point de vue acoustique entre les différents morceaux. En comparant les deux figures, il apparaît que les genres du corpus GTZAN ont l'air plus distinguables acoustiquement parlant, ce qui peut expliquer les meilleures performances au pré-entraînement.

3.7 Prédiction du goût musical

3.7.1 Métriques d'évaluation utilisées pour la prédiction de goûts

Afin d'évaluer les performances de nos modèles, il est nécessaire d'utiliser des métriques pertinentes et adaptées au problème considéré.

Nous pouvons calculer la précision sur chacune des classes, qui correspond au nombre de vrais positifs divisés par le nombre de réponses positives totales : P_l pour la précision sur la classe « Like », P_d pour la classe « Dislike » :

$$P_l = \frac{VP_{\text{Like}}}{VP_{\text{Like}} + FP_{\text{Like}}} \tag{3.1}$$

$$P_d = \frac{VP_{\text{Dislike}}}{VP_{\text{Dislike}} + FP_{\text{Dislike}}} \tag{3.2}$$

La précision globale correspond au nombre total de bonnes réponses divisé par le nombre de réponses, sur les deux classes :

$$P = \frac{VP_{\text{Like}} + VP_{\text{Dislike}}}{VP_{\text{Like}} + VP_{\text{Dislike}} + FP_{\text{Like}} + FP_{\text{Dislike}}} \tag{3.3}$$

Néanmoins, dans un contexte de recommandation, seuls les morceaux prédits comme « like » seront proposés à un utilisateur. Pour cette raison, nous nous concentrerons sur la métrique P_l. De plus, dans le cas de grandes bases de données, tous les morceaux prédits comme « like » ne sont pas recommandés à l'utilisateur, mais seulement ceux ayant les plus fortes probabilités. Ainsi, nous avons choisi d'utiliser la métrique Précision_Like@N ($P_l@N$) pour l'évaluation des modèles. Pour chaque jeu de test, nous avons donc vérifié la validité des prédictions avec les N plus fortes probabilités, pour $N = 1, 3, 5$.

3.7.2 Pré-traitement des données

Dans un premier temps, nous avons équilibré les données de chaque volontaire afin d'avoir 50% de morceaux aimés et 50% de morceaux pas aimés dans le jeu d'entraînement. Pour ce faire, nous avons ignoré une partie des données de la classe sur-représentée. Des classes ainsi équilibrées permettent de ne pas fausser l'apprentissage du réseau de neurones [118]. Les modèles ont étés entraînés sur 90% des données, et nous effectuons des prédictions sur les 10% restants. De plus, nous avons procédé à une validation croisée afin d'améliorer la robustesse de nos résultats. Pour chaque volontaire, nous avons constitué 5 répartitions différentes entre données d'entraînement et de test, afin d'éviter tout biais (avantageux ou non) lié à la répartition des données.

Comme lors de la classification en genres, les morceaux ont été découpés en extraits de 3 secondes et nous avons calculé leurs spectrogrammes.

3.7.3 Prédictions et évaluation hors-ligne

Les métriques précédemment présentées (cf. partie 3.7.1) sont ensuite utilisées pour valider ces prédictions, puis moyennées sur les 5 jeux de validation croisée. Afin de vérifier l'intérêt du Transfer Learning, nous avons également entraîné et testé des modèles sans pré-entrainement. Les distributions de scores $P_l@N$ avec validation croisée sont données sur les figures 3.5 et 3.6.

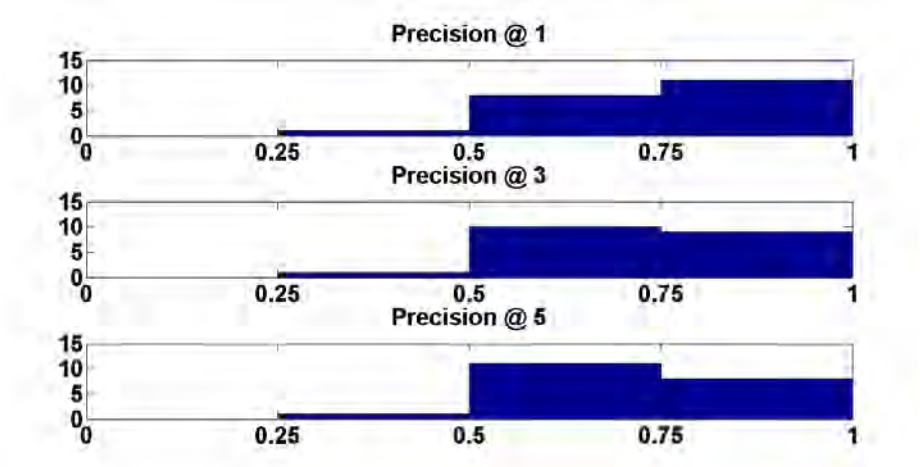

FIGURE 3.5: Répartition des scores des modèles pré-entraînés sur le corpus GTZAN.

Les scores individuels des 20 volontaires sont affichés sur la figure 3.7.

Pour la métrique $P_l@1$, nous voyons que les modèles pré-entrainés avec le corpus FMA ont été meilleurs que ceux pré-entrainés avec GTZAN, pour la plupart des volontaires, exceptés certains cas critiques : #3, #4 et #14. Le système a obtenu une $P_l@1$ de plus de 75% pour 14 volontaires pour le pré-entrainement avec FMA contre 12 avec GTZAN. Nous remarquons que les modèles non pré-entraînés sont toujours compétitifs sur cette métrique.

Avec $P_l@3$, les deux systèmes pré-entraînés ont de meilleurs performances que le système sans pré-entraînement pour la plupart des utilisateurs. Malgré une précision globale plus faible, le système pré-entraîné avec FMA compte plus de scores au-delà de 75%. Enfin, sur la métrique $P_l@5$ les scores des modèles pré-entraînés avec GTZAN et FMA se rapprochent, et sont significativement meilleurs que ceux sans pré-entraînement.

Les scores moyens obtenus sur l'ensemble des volontaires en validation croisée sont donnés dans la table 3.2.

Nous voyons un avantage significatif pour les méthodes utilisant le Transfer Learning. Les scores sont satisfaisants, mais présentent un écart type non négligeable sur la précision globale, qui s'explique par le fait que les systèmes ont obtenu de très bonnes performances pour certains volontaires (voir tables 3.3

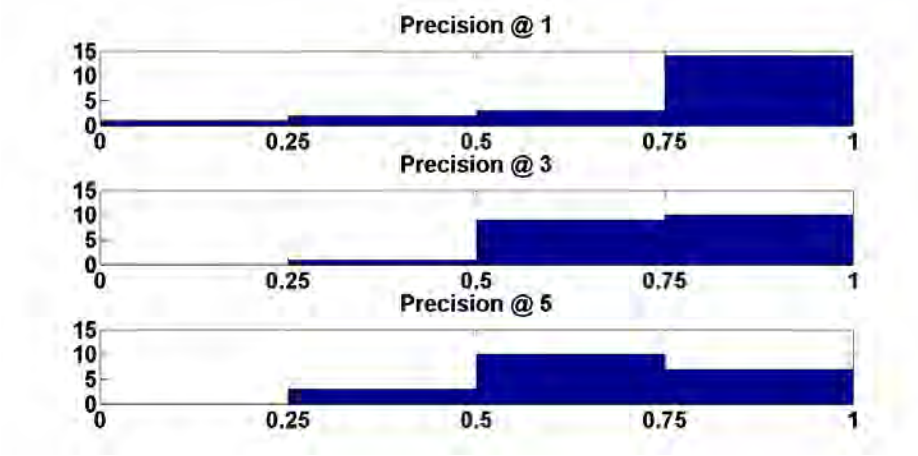

FIGURE 3.6: Répartition des scores des modèles pré-entraînés sur le corpus FMA.

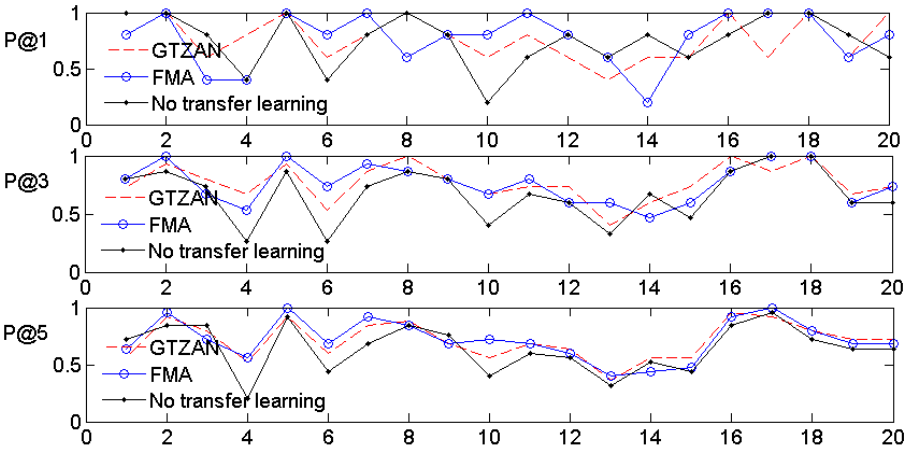

FIGURE 3.7: Scores pour les 20 volontaires.

et 3.4), et de moins bonnes pour d'autres (voir tables 3.5 et 3.6). Nous allons analyser plus finement des exemples pour ces deux cas.

Dans les tables 3.3 et 3.4, nous voyons qu'une bonne précision globale va de pair avec de bonnes Précision@N. Dans ces cas-là, le système est capable de donner de bonnes prédictions sur les deux classes, y compris pour les prédictions avec les probabilités les plus fortes. Pour le volontaire #9, le système a été globalement performant, mais a chuté pour $N = 5$. Cela peut être expliqué par le fait que certains morceaux catégorisés comme « Dislike » par le volontaire pré-

TABLE 3.2: Précision globale P et Précisions @ N $P_l@N$ (%).

Pre-training	P	$P_l@1$	$P_l@3$	$P_l@5$
GTZAN	68 ±12	76	77	71
FMA	60 ±16	77	76	72
no pre training	59 ±13	75	67	64

TABLE 3.3: GTZAN pré-apprentissage : 5 meilleures précisions globales et les métriques $P_l@N$ associées (%).

Volunteer	P	$P_l@1$	$P_l@3$	$P_l@5$
#18	98	100	100	80
#17	90	60	87	92
#7	80	80	87	84
#5	79	100	93	92
#9	77	80	80	68

TABLE 3.4: FMA pré-apprentissage : 5 meilleures précisions globales et les métriques $P_l@N$ associées (%).

Volunteer	P	$P_l@1$	$P_l@3$	$P_l@5$
#18	93	100	100	80
#6	77	80	73	68
#9	77	80	80	68
#8	74	60	87	84
#1	70	80	80	64

TABLE 3.5: GTZAN pré-apprentissage : 5 pires précisions globales et les métriques $P_l@N$ associées (%).

Volunteer	P	$P_l@1$	$P_l@3$	$P_l@5$
#3	44	60	80	80
#12	49	60	73	64
#13	55	40	40	36
#10	60	60	67	56
#4	62	80	67	52

TABLE 3.6: FMA pré-apprentissage : 5 pires précisions globales et les métriques $P_l@N$ associées (%).

Volunteer	P	$P_l@1$	$P_l@3$	$P_l@5$
#19	30	60	60	68
#12	35	80	60	60
#17	37	100	100	100
#13	38	60	60	40
#16	43	100	87	92

sente des caractéristiques acoustiques similaires avec des morceaux catégorisés comme « Like ».

Les tables 3.5 et 3.6 montrent qu'une mauvaise précision globale peut néanmoins toujours donner de bonnes précisions sur les N plus fortes probabilités pour les volontaires #3 et #17. En gardant à l'esprit le fait qu'un bon système de recommandation n'a pas besoin d'être précis sur l'ensemble de ses prédictions - seules celles au niveau de confiance les plus élevées sont critiques - les scores de volontaires #3 et #17 peuvent toujours être considérés comme bons. De plus, le volontaire #3 a annoté 3 fois plus de morceaux comme « Like » que « Dislike », mais le système a produit des prédictions équilibrées entre ces deux classes, ce qui explique la faible précision.

Des résultats plus détaillés pour chaque volontaire sont disponibles sur la page github du projet[3].

Aussi, tous les modèles de deep learning nécessitent un réglage fin de ses hyperparamètres (pas d'apprentissage initial, etc.) pour obtenir des performances optimales. Comme dit précédemment, nous avons constaté que le meilleur pas d'apprentissage initial n'était pas le même pour tous les volontaires. Nous avons donc fait un compromis en conservant celui donnant les meilleures performances globales. Un réglage fin des hyperparamètres, par exemple via des méthodes de « grid search » ou « random search » ([12]) serait envisageable pour chaque volontaire, mais serait probablement gourmand en temps et en ressources à grande échelle.

En conclusion, les scores obtenus sont satisfaisants, avec un avantage pour le modèle pré-appris sur le corpus FMA. Cela peut s'expliquer par sa plus grande taille, mais aussi par la plus grande diversité de genres représentés, comme la catégorie « Electronique » notamment. En utilisant FMA en pré-entraînement, nous pouvons considérer que ce corpus appartient à un « domaine source » plus proche du « domaine cible » que GTZAN. Comme dit précédemment le corpus GTZAN présentait un problème de classes trop homogènes. Ainsi, le bon score obtenu au pré-entraînement ne s'est pas traduit par un bon score pour la prédiction du goût musical. Nous pouvons donc en déduire que le corpus FMA, par sa diversité, amène les modèles à davantage généraliser lors du pré-

3. https://github.com/Nicolas-DBN/Scores

entraînement.

Cette expérience montre qu'en Transfer Learning, le choix du corpus de pré-entraînement ne peut pas se porter uniquement sur les scores donnés par l'état de l'art sur celui-ci. Un corpus plus « difficile » peut se révéler plus utile dans le cadre du Transfer Learning, car il entraîne davantage de généralisation.

Dans certains cas, l'explication des goûts de l'utilisateur se situe ailleurs que dans les caractéristiques acoustiques du morceau. Le contexte, l'heure de l'écoute ou bien encore l'environnement social peuvent être des facteurs déterminants.

3.7.4 Evaluation humaine

Les expériences précédentes ont été réalisées « hors ligne » avec une partie des données destinées à l'entraînement et l'autre au test du système. Nous avons cherché à prédire un résultat qui en réalité a déjà été observé. Bien que ce mode d'évaluation automatique soit largement utilisé dans le domaine de la recherche en recommandation, il reste cependant éloigné du scénario de recommandation de musique dans la réalité. C'est pourquoi nous avons décidé de mener une évaluation supplémentaire des modèles, humaine cette fois. L'idée ici est d'utiliser les modèles entraînés pour chaque volontaire afin de donner des prédictions sur de nouveaux morceaux, absents de l'expérience jusqu'à présent. Nous avons ensuite fait appel aux volontaires afin de mesurer la qualité de ces prédictions.

Pour cela, nous avons dans un premier temps sélectionné 100 morceaux : 5 nouveautés appartenant à 20 catégories proposées dans la section « parcourir » du site Deezer (voir figure 3.8).

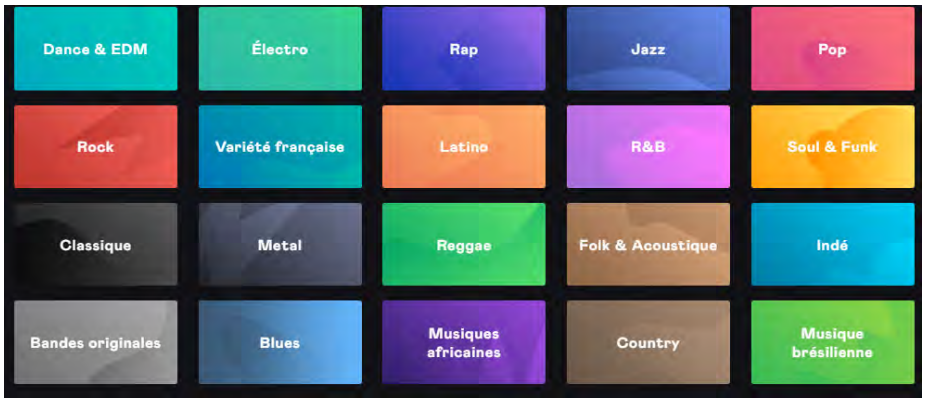

FIGURE 3.8: 20 catégories du nouveau corpus.

À l'aide de l'API de Deezer, nous avons recueilli les genres présents dans la playlist constitué pour cette expérience, voir figure 3.9.

Nous observons que la répartition de ces genres n'est pas homogène. Bien qu'en ayant choisi seulement 5 morceaux de la catégorie Pop, le corpus contient 14 morceaux annotés comme pop. Le genre alternative est également très présent alors qu'aucune des 20 catégories n'en faisait mention.

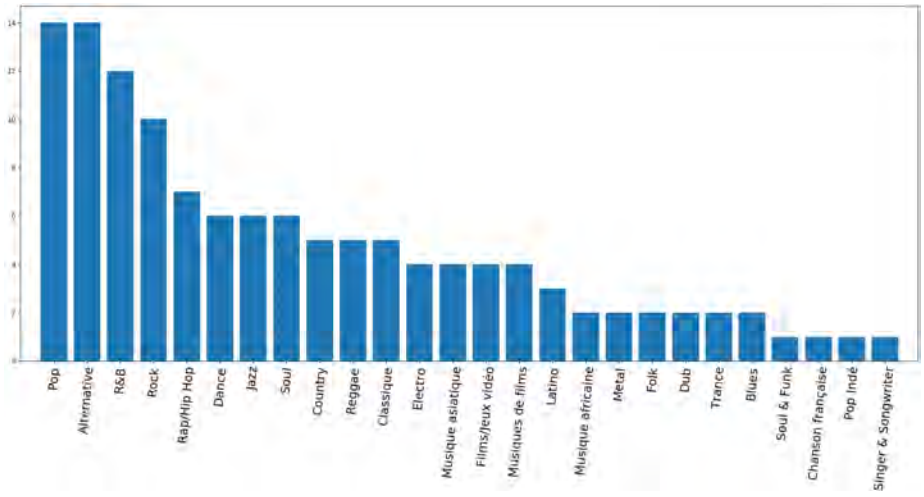

FIGURE 3.9: Genres des morceaux de la playlist source, API Deezer.

Afin d'avoir plus de détails, nous avons procédé de la même manière en utilisant cette fois l'API de Spotify pour obtenir des genres plus détaillés (figure 3.10).

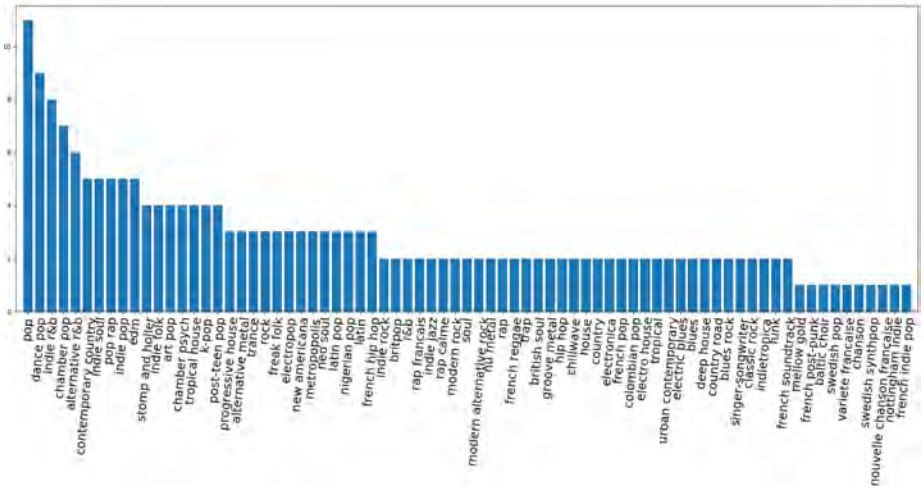

FIGURE 3.10: Genres des morceaux de la playlist source, API Spotify.

Pour rappel, les annotations en genres données par les API de Deezer et Spotify sont multilabel, c'est-à-dire qu'un même morceau peut être annoté avec plusieurs genres différents. Comme vu dans la partie 1.4, certains genres, ont plus de chance d'être associés à d'autres, ce qui explique la prédominance de pop et alternative dans ce corpus.

Pour tous les volontaires, nous avons estimé les probabilités d'aimer chacun des 100 morceaux. Pour chaque participant nous avons utilisé le modèle entraîné précédemment. Nous avons utilisé les modèles pré-entraînés avec le corpus FMA en raison de leurs meilleurs résultats sur les métriques $P_l@1$, $P_l@3$ et $P_l@5$. Les prédictions pour chaque morceau sont différentes pour tous les volontaires puisque des modèles différents ont étés entraînés en amont pour chacun d'entre eux.

Pour un volontaire donné, les morceaux sont ainsi ordonnés dans un ordre décroissant selon ces probabilités, les premiers étant les morceaux ayant le plus de chances d'être appréciés par le volontaire.

Afin d'évaluer les performances du modèle, nous avons sélectionné les 5 premiers et les 5 derniers morceaux. Ces 10 morceaux ont été présentés aux volontaires dans un ordre arbitraire accompagnés de la consigne suivante :
« 1. Classez les morceaux selon votre préférence, du meilleur au moins bon. Vous pouvez directement modifier l'ordre de la playlist sur Deezer (web ou mobile).
2. Indiquez les morceaux que vous aimez vraiment, en les ajoutant à vos coups de coeur. » De plus, des entretiens ont été réalisés avec certains volontaires afin d'obtenir des informations sur leurs goûts musicaux qui permettraient d'expliquer les résultats obtenus.

Les 5 morceaux aux plus fortes probabilités de *Like* sont prédits comme « Like » par l'algorithme. Les 5 premiers morceaux du classement effectué par le volontaire sont considérés comme « Like » dans la vérité terrain, les 5 derniers comme « Dislike ». Comme lors de l'évaluation automatique, nous calculons 3 métriques : $P_l@1$, $P_l@3$, $P_l@5$ à partir de ces prédictions et de la vérité terrain.

Dans un scénario de recommandation musicale, il est évidemment hors de question de proposer des morceaux avec les plus faibles probabilités d'être appréciés à l'utilisateur. Ici ces 5 morceaux font office de « témoins », et permettent d'éliminer 2 biais possibles en ne présentant que les 5 morceaux à plus fortes probabilités. Les morceaux présentés étant extraits d'une pré-sélection de 100 morceaux, il est tout à fait possible que certains volontaires n'aiment aucun morceau de cette pré-sélection (biais 1), ou au contraire les apprécient tous (biais 2). Ainsi, nous évaluons l'appréciation relative des morceaux par les volontaires plutôt que l'appréciation absolue. Ici, le goût est donc considéré comme une préférence entre plusieurs morceaux de musique.

Les résultats de l'évaluation humaine sont détaillés dans la table 3.7.

Les scores sont globalement satisfaisants. Pour des modèles pré-entraînés avec FMA, nous avions obtenu des scores de 77%, 76% et 72% pour les métriques $P_l@1$, $P_l@3$, $P_l@5$. Avec l'évaluation humaine, nous constatons une détérioration sur la métrique $P_l@1$ mais une amélioration sur les métriques $P_l@3$ et $P_l@5$. Pour les scores moyens, les résultats sont donc plutôt cohérents entre les deux métriques. De fortes variations subsistent néanmoins : les goûts de certains volontaires ont pu être modélisés avec succès alors que pour d'autres les résultats demeurent décevants. Une étude approfondie de certains cas sera présentée dans la partie suivante.

Cette deuxième évaluation confirme donc que certains goûts musicaux peuvent être appris à partir de spectrogrammes en utilisant des CNN pré-entraînés à la

TABLE 3.7: Scores obtenus pour chaque volontaire en évaluation humaine.

Volontaire	P_l@1	P_l@3	P_l@5
#1	1/1	3/3	5/5
#2	1/1	3/3	4/5
#4	1/1	3/3	5/5
#5	1/1	2/3	4/5
#6	1/1	3/3	4/5
#7	1/1	3/3	5/5
#9	1/1	3/3	5/5
#10	0/1	2/3	3/5
#11	0/1	1/3	3/5
#12	0/1	1/3	2/5
#13	0/1	2/3	3/5
#14	1/1	2/3	3/5
#16	0/1	2/3	2/5
#18	1/1	3/3	5/5
#20	1/1	2/3	4/5
MOY	66.7%	77.8%	77.3%

reconnaissance de genres. Les paramètres de bas niveau appris automatiquement par le CNN pour la reconnaissance de genre sont donc également pertinents pour la prédiction de goûts musicaux.

Ainsi, bien que les modèles entraînés à reconnaître des genres de musique soient inadaptés à la recommandation de musique et la prédiction de goûts (voir partie 1.3.2), ils sont néanmoins exploitables à ces fins via des méthodes de Transfer Learning.

3.8 Analyse des résultats, étude qualitative

Nous avons vu que le système présenté précédemment a obtenu de bons résultats tant dans l'évaluation automatique que dans l'évaluation humaine. Cependant, il existe des différences de scores non négligeables entre les deux types d'évaluations et entre les volontaires. Nous avons donc mené une étude plus approfondie des données et des résultats.

3.8.1 Corrélation entre l'évaluation humaine et automatique

Pour comparer les résultats obtenus avec une évaluation automatique et une évaluation humaine, nous avons tout d'abord calculé le taux de corrélation entre les scores obtenus de ces deux manières.

Nous avons calculé la corrélation entre les scores obtenus pour chaque volontaire pour l'évalution hors-ligne et pour l'évaluation humaine (voir table 3.8).

TABLE 3.8: Corrélation entre les scores des évaluations humaines et hors-ligne.

	TOP 1	TOP 3	TOP 5
Corrélation	-0.1225	0.3178	0.1669

Il n'apparaît aucune corrélation forte. Ainsi, bien que les scores soient satisfaisants dans les 2 cas, rien ne semble indiquer que de bons scores calculés sur un jeu de données hors ligne amènent à de bons scores sur une évaluation humaine.

Afin d'apporter des explications à ce phénomène, nous allons nous intéresser aux cas pour lesquels les scores sont les moins consistants entre l'évaluation hors-ligne et l'évaluation humaine.

3.8.2 Influence de la quantité de données

En apprentissage automatique, et plus particulièrement en apprentissage profond, la quantité des données utilisées pour entraîner un modèle peut être déterminante. Dans notre cas, il n'apparaît aucun lien entre les bons scores pour certains utilisateurs et la quantité de données utilisée. Le calcul des corrélations entre la quantité de données et les différentes métriques automatiques et humaines n'ont montré aucune corrélation avérée, positive ou négative.

Par exemple, le volontaire #18 est celui pour lequel la quantité de données disponibles pour l'entraînement était la plus faible (« seulement » 6h19). Le modèle permettant de modéliser ce volontaire a obtenu un score parfait sur les 3 métriques. À l'inverse, pour le volontaire #12, pour lequel les résultats obtenus sont les moins bons, nous avions à notre disposition 15h14 de données pour entraîner le modèle. Pour un utilisateur donné, il est raisonnable de penser qu'une quantité d'informations supplémentaires sur ses goûts musicaux nous permettrait d'obtenir de meilleurs résultats. Mais de manière générale, un utilisateur ayant fourni beaucoup de données ne sera pas nécessairement plus facilement modélisable qu'un autre ayant fourni moins d'informations. Des goûts simples, s'appuyant par exemple sur une opposition entre plusieurs genres de musique, peuvent être modélisés à partir de quelques morceaux. À l'inverse, des goûts complexes qui reposent sur des détails plus subtils nécessiteront une quantité d'informations bien supérieure pour être modélisés.

Puisque la quantité de données entre les différents volontaires n'a pas d'influence sur les résultats, nous avons mené une étude qualitative de ces données pour les différents volontaires.

3.8.3 Une prédiction de « dislike » ?

Les résultats du modèle du volontaire #4 sont présents dans la table 3.9.

TABLE 3.9: Scores obtenus pour le volontaire #4.

	TOP 1	TOP 3	TOP 5
Score hors-ligne (5 jeux de validation)	40%	53%	56%
Score humain	100%	100%	100%

Les scores obtenus lors de la première évaluation semblaient indiquer que le système n'était pas parvenu à modéliser correctement les goûts de ce volontaire.

Néanmoins, lors de la deuxième partie de l'expérience, le système a pu distinguer avec une précision de 100% les affinités de ce volontaire. Les prédictions données par le système soumis à une évaluation humaine sont disponibles dans les tables 3.10 et 3.11.

TABLE 3.10: Les 5 morceaux avec les plus fortes probabilités de « Like » pour le volontaire #4.

Artiste	Morceau	Proba. Like	Genre (Spotify)
Sech	Me Olvidé	46%	reggaeton, [...]
Zeb Samuels	Deep Inna Sound	46%	hip hop, [...]
Marcus Gad	Rebel Form of Soul	47%	french reggae, [...]
Whitney	On a oublié	47%	dance pop, [...]
Pasquale Grasso	Parisian Thoroughfare	50%	jazz

TABLE 3.11: Les 5 morceaux avec les plus faibles probabilités de « Like » pour le volontaire #4.

Artiste	Morceau	Proba. Like	Genre (Spotify)
Pyre	Impaler the Redeemer	6%	metal
Crowbar	All I Had (I Gave)	10%	alternative metal, [...]
Sabiendas	The Human Centipede	10%	german death metal
Soulfly	The Summoning (Live)	11%	alternative metal, [...]
Joe Satriani	Shapeshifting	13%	neo classical metal, [...]

En interrogeant l'API de Spotify, nous constatons que les 5 morceaux avec les probabilités de « Like » les plus hautes appartiennent à des genres différents : Rap/Hip hop, Latino, Reggae, Pop, et Jazz. À l'inverse, les 5 morceaux avec les probabilités les plus basses d'être aimés appartiennent tous à des sous-genres de Metal.

De plus, lors d'un entretien passé après l'évaluation, le volontaire a indiqué ne pas aimer particulièrement de morceaux parmi les 5 premiers. Nous constatons également que les probabilités de Like les plus fortes ne dépassent pas 50%.

Nous pouvons donc en conclure que, dans ce cas, le système a correctement identifié des morceaux que le volontaire n'aimerait pas. Les 5 morceaux

en première position appartiennent tous à des genres différents, mais ils ont la caractéristique commune de ne pas appartenir au genre metal.

3.8.4 Une prédiction de « Like »

Le cas précédent du volontaire #4 a été observé à plusieurs reprises sur d'autres volontaires, avec 4 ou 5 morceaux de métal prédits -à raison- comme « Dislike ». Face à de tels résultats, nous pourrions nous demander si ces morceaux ne contiennent pas une caractéristique acoustique qui pousserait les modèles à donner une faible probabilité de « Like » aux morceaux de métal, quel que soit le volontaire, biaisant ainsi les résultats... Mais les données du volontaire #18 nous montrent le contraire (voir table 3.12).

TABLE 3.12: Scores obtenus pour le volontaire #18.

	TOP 1	TOP 3	TOP 5
Score hors-ligne (5 jeux de validation)	100%	100%	80%
Score humain	100%	100%	100%

Les scores obtenus pour le volontaire #18 sont très satisfaisants tant pour l'évaluation automatique que pour l'évaluation humaine, où les scores sont de 100% sur les 3 métriques.

Nous voyons dans la table 3.13 que les 5 morceaux avec les plus fortes probabilités de « Like » appartiennent tous au genre Metal ou à des genres affiliés.

TABLE 3.13: Prédictions pour le volontaire #18.

Prédiction	Artiste	Morceau	% Like	Genre
Top 5	SABIENDAS	The Human Centipede	89	german death metal
Top 5	Caligula's Horse	The Tempest	90	alternative metal
Top 5	Crowbar	All I Had (I Gave)	90	alternative metal
Top 5	Joe Satriani	Shapeshifting	90	neo classical metal
Top 5	Pyre	Impaler the Redeemer	95	metal
Bottom 5	Hamza	Netflix	10	rap/hip-hop
Bottom 5	Teyana Taylor	Made It	13	r&b
Bottom 5	TomE	Cherry Blossom	16	pop
Bottom 5	Morat	Bajo La Mesa	20	latin pop
Bottom 5	Zeb Samuels	Deep Inna Sound	20	rap/hip-hop

Par ailleurs, nous voyons dans la figure 3.11 que ce volontaire a indiqué aimer des morceaux de Metal en grande majorité.

Ainsi, les données de ce volontaire nous ont permis de confirmer que ce système est tout à fait apte à identifier les affinités autant que les aversions des volontaires. Si pour des volontaires tels que le #4 il est possible que les modèles n'aient réussi qu'à identifier leurs aversions, pour le volontaire #18 c'est bien l'affinité avec le Metal qui a été identifiée. La musique Metal a été étudiée dans [78] afin d'expliquer son caractère particulièrement clivant pour les auditeurs.

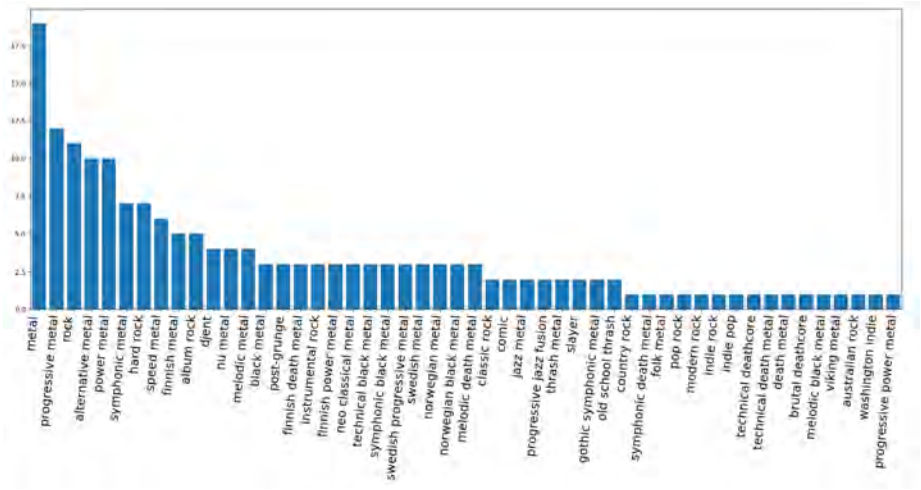

FIGURE 3.11: Genres des morceaux classés « + » et « ++ » par le volontaire #18, API Spotify.

3.8.5 Le problème de la musique commerciale

Lors de l'évaluation automatique, le modèle du volontaire #12 a obtenu des scores corrects (voir table 3.14). En revanche, lors de l'évaluation humaine, c'est pour ce volontaire que le système a le moins bien fonctionné.

TABLE 3.14: Scores obtenus pour le volontaire #12

	TOP 1	TOP 3	TOP 5
Score hors-ligne (5 jeux de validation)	80%	60%	60%
Score humain	0%	33%	40%

Dans la table 3.15 sont affichées les prédictions données par le système pour le volontaire #12. Nous voyons que la plupart de ces prédictions se sont avérées erronées.

Pour expliquer cette différence de résultat, nous nous sommes intéressés aux genres que ce volontaire a indiqué apprécier. Pour cela, nous avons interrogé dans un premier temps l'API de Deezer pour connaître les genres des morceaux que ce volontaire a placés dans les playlists « + » et « ++ ». Sur la figure 3.12, il apparaît que les 3 genres majoritairement représentés sont Electro, Dance et Jazz.

Nous avons procédé de la même manière pour connaître les genres des morceaux que ce volontaire n'a pas appréciés. Il apparaît dans la figure 3.13 que les genres sont très similaires : nous retrouvons Electro, Dance, Jazz, Pop dans les 4 premières positions.

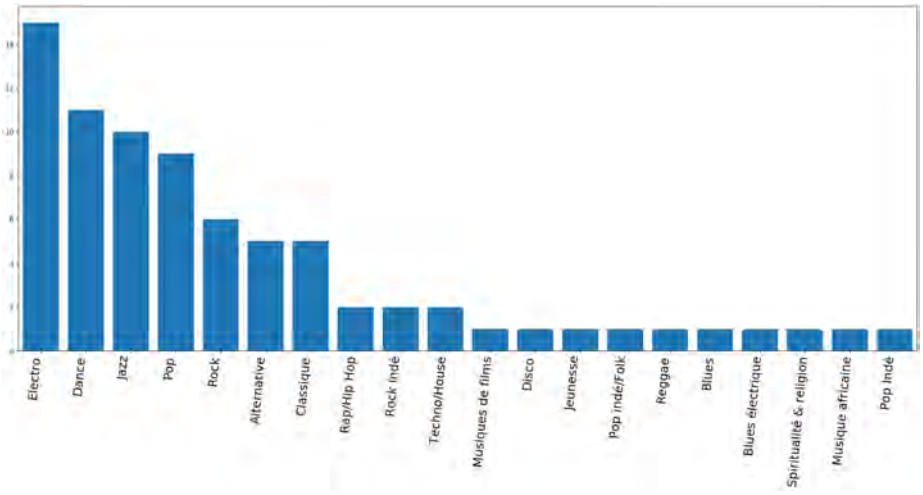

FIGURE 3.12: Genres des morceaux classés « + » et « ++ » par le volontaire #12, API Deezer.

Comme vu précédemment, les genres disponibles via l'API de Deezer peuvent parfois manquer de précision. Nous avons donc utilisé l'API de Spotify pour tenter d'y voir plus clair. Sur la Figure 3.14, les sous-genres donnés par Spotify montrent une prépondérance des sous-catégories d'Electro et de Techno/house dans les affinités de ce volontaire : electronica, microhouse, minimal techno, tech house, float house, deep house, etc.

Comme précédemment, nous retrouvons ces mêmes genres parmi les morceaux que ce volontaire n'a pas appréciés (voir figure 3.15. D'autres sous-genres sont également représentés dans les deux catégories, comme : Indie soul, Rock, ou bien Indie jazz.

Dans ce cas, nous pouvons en conclure que les données fournies au modèle pour son apprentissage ne permettaient pas de généraliser les goûts de ce volontaire.

TABLE 3.15: Prédictions pour le volontaire #12.

Prédiction	Artiste	Morceau	% Like	Genre
Top 5	Omah lay	Damn	52	pop
Top 5	Joe Satriani	Shapeshifting	53	rock
Top 5	Zeb Samuels	Deep Inna Sound	55	hip-hop
Top 5	Pyre	Impaler the Redeemer	62	metal
Top 5	Tory Lanez	Temperature Rising	64	r&b
Bottom 5	The Steeldrivers	Bad For You	24	rock
Bottom 5	Joe Harriott	Shepherd's Serenade	24	jazz
Bottom 5	Tony Succar	Raices Jam	28	jazz
Bottom 5	Marcus Gad	Rebel Form of Soul	29	reggae
Bottom 5	Les Enfoirés	A côté de toi	29	chanson française

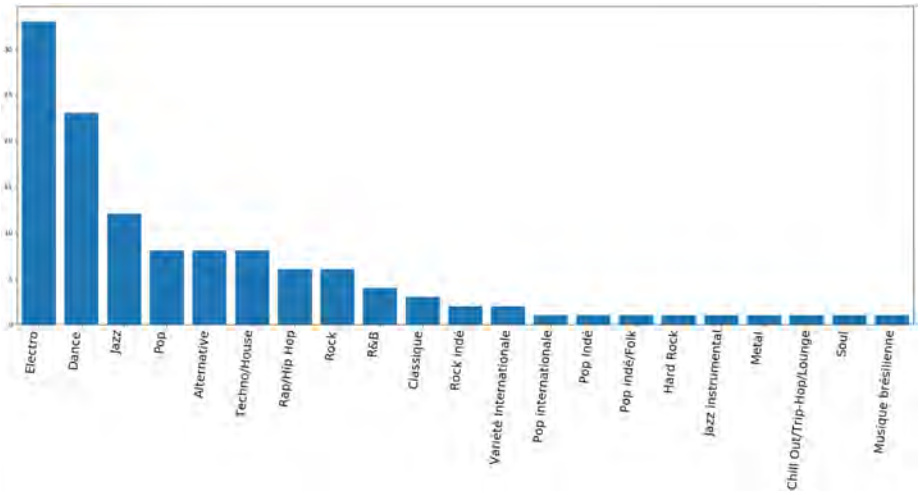

FIGURE 3.13: Genres des morceaux classés « - » et « - - » par le volontaire #12, API Deezer.

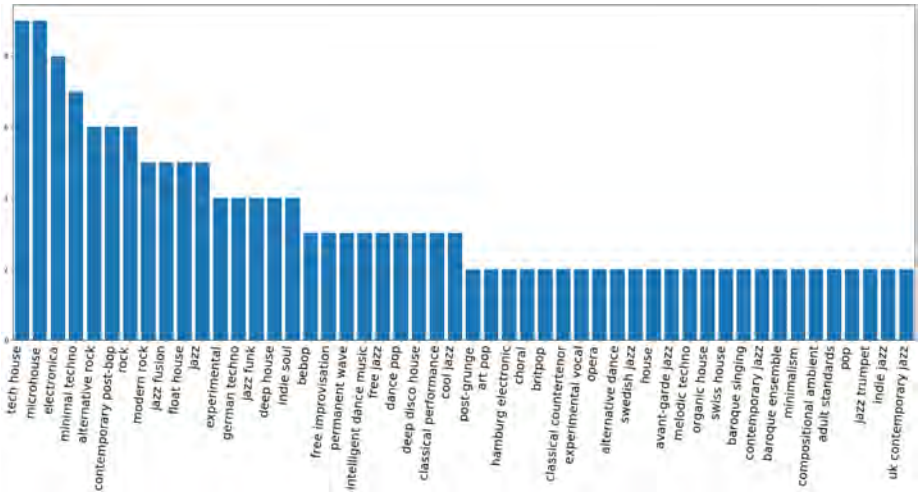

FIGURE 3.14: Genres des morceaux classés « + » et « ++ » par le volontaire #12, API Spotify.

Un entretien avec la personne concernée nous a permis d'en savoir plus sur le fonctionnement de ses goûts, en voici quelques extraits :
« [...] ça fonctionne au niveau des genres : mes préférences sont bossa nova, jazz, metal, rock, techno, house, tech house, drum and bass, dub, reggae. [...] En gros j'aime plutôt la musique instrumentale sans vocale. Je me concentre plutôt [sur] la batterie, si c'est quelque chose super répétitif comme reggaeton [...] je n"écoute pas tout simplement. Et j'ai une allergie à la musique populaire

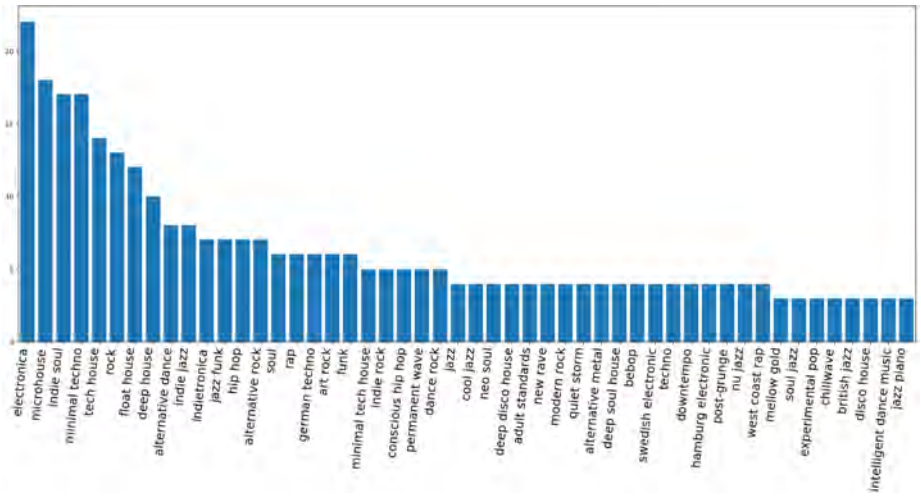

FIGURE 3.15: Genres des morceaux classés « - » et « - - » par le volontaire #12, API Spotify.

de tous les pays. Tout ce qui est Ed Sheran ou Shym etc. [ces] musiques limite me dérangent [...] »

Dans le cas de ce volontaire, le critère qui semble donc être déterminant quant à l'appréciation de certains morceaux semble aller au-delà du genre. Ici, c'est la dimension « commerciale » , « populaire » ou encore « mainstream » de certains morceaux qui provoque l'aversion. Pour répondre à cette problématique, plusieurs solutions peuvent être proposées :

- La première serait d'utiliser une quantité supérieure de données. Avec davantage de données, il est possible que le modèle de ce volontaire s'ajuste suffisamment au point de cerner la nuance entre plusieurs morceaux musicalement similaires, se distinguant principalement par leur aspect « commercial ».
- Une deuxième serait d'utiliser un modèle pré-entraîné spécialement pour détecter - toujours dans le signal audio - le caractère « commercial » de certains morceaux. Des recherches ont déjà été menées dans ce domaine, telles que [127] et [69], où les auteurs cherchent à prédire des statistiques de l'évolution d'un morceau dans le classement « Billboard Hot 100 ». Ce classement, qui fait figure de référence dans le domaine, est élaboré à partir des playlists des radios, des données de ventes ainsi que du streaming. Dans cet article, des paramètres acoustiques sont extraits sur le signal audio des morceaux. À partir des variations de ces paramètres sont calculés différents paramètres de complexité de plus haut niveau, sur lesquels les différents algorithmes de classification et de régression se basent pour prédire la popularité.
- Néanmoins, le caractère « commercial » de certains morceaux de musique peut tout à fait dépendre du contexte, notamment d'effets de mode, etc.

Par exemple, Bauer et. al ont montré dans [9] que cette notion peut varier selon le pays d'apartenance de l'auditeur. De plus, le nombre d'écoutes d'un morceau peut être un très bon indicateur de cette popularité. Les métadonnées pourraient ainsi être prises en compte afin d'améliorer la recommandation de musique pour certains profils particulièrement exigeants. Dans certains cas, les algorithmes basés sur le contenu combinent à la fois des paramètres extraits sur le signal audio, et des métadonnées. Si ce paramètre est discriminant lors de l'apprentissage du modèle, il sera alors pris en compte par l'algorithme d'apprentissage profond. Par exemple, dans [75] les auteurs prennent en compte les paroles des morceaux ainsi que des métadonnées en addition du signal audio afin de prédire sa popularité.

— De plus, le caractère « commercial » peut être également considéré dans un but de recommandation axée sur la découverte. Un paramètre pourrait être commandé par l'utilisateur comme un « filtre » de popularité pour ses recommandations : *« Ne pas me recommander de morceaux avec plus de N écoutes »*. Ici, l'IA perd du terrain pour laisser la main à l'utilisateur sur la recommandation, ce qui pourrait aller à contre-courant de la volonté des plateformes de streaming [14].

3.8.6 L'effet artiste (et l'effet album) : un problème de sur-apprentissage

Pour le volontaire #16, les scores obtenus via une évaluation humaine sont nettement moins bons qu'avec l'évaluation hors-ligne (voir table 3.16).

TABLE 3.16: Scores obtenus pour le volontaire #16.

	TOP 1	TOP 3	TOP 5
Score hors-ligne (5 jeux de validation)	100%	87%	92%
Score humain	0%	66%	40%

Pour alimenter ses playlists « Like », ce volontaire a donné 103 morceaux différents. Néanmoins, il est apparu une grande répétition de certains artistes, et albums. À eux seuls, les artistes Mistura Pura, Gecko Turner, Chlorine Free, mt fujitive et Maria Pomianowska ont accumulé 52 morceaux, répartis sur seulement 6 albums différents.

Chaque artiste possède une signature acoustique particulière à travers sa voix s'il s'agit d'un chanteur, les instruments utilisés ou bien encore le genre de musique joué. Si ces critères peuvent varier au cours de la carrière d'un artiste, cet effet est amplifié au sein d'un même album. La couleur particulière que peut avoir un album est liée aux techniques de production : instruments utilisés, matériel et techniques d'enregistrement et de mixage, effets utilisés...

De plus, dans [48] les auteurs ont mené une expérience sur des outils de

recommandation basés sur le signal audio pour mettre en évidence les phénomènes de « Artist effect » et « Album effect ». Ils ont observé qu'un tiers des premières recommandations de leur algorithme appartenaient au même album que le morceau requête.

Dans notre cas, cette forte proportion de morceaux appartenant aux mêmes albums peut expliquer les très bons scores obtenus avec une évaluation hors-ligne. Si des morceaux du même album se retrouvent à la fois dans les jeux d'apprentissage et de test, le modèle n'a aucun mal à donner une bonne prédiction pour ces morceaux. Avec autant de morceaux appartenant aux mêmes albums, le modèle a probablement « sur-appris » (overfitting) les goûts de ce volontaire : plutôt que de généraliser des caractéristiques récurrentes dans ses affinités, le modèle s'est concentré sur des traits particuliers de ces 5 albums.

En l'absence de données qui lui sont familières, un modèle qui a été sur-entraîné donne ainsi des réponses aléatoires, ce qui peut donc expliquer les mauvais résultats à l'évaluation humaine. Afin d'éviter ce surapprentissage, il aurait fallu s'assurer lors de la répartition des données en jeu d'apprentissage, de validation et de test que les morceaux d'un même album soient bien regroupés dans un même jeu. Une autre manière de procéder pourrait être de veiller à ne pas avoir plus d'un morceau du même album dans toutes les données.

Les résultats de ce volontaire nous montrent donc qu'une grande quantité de données n'apporte pas forcément une bonne généralisation du modèle.

3.8.7 La distinction entre les genres et les sous-genres

Les scores obtenus pour le volontaire #7 sont très satisfaisants tant pour l'évaluation automatique que pour l'évaluation humaine (voir table 3.17).

TABLE 3.17: Scores obtenus pour le volontaire #7.

	TOP 1	TOP 3	TOP 5
Score auto (5 jeux de validation)	100%	93%	92%
Score humain	100%	100%	100%

Ici, nous allons nous intéresser aux genres Deezer et sous-genres Spotify des morceaux de ses playlists « Like » et « Dislike ».

Les figures 3.16 et 3.17 nous montrent les genres représentés parmi les morceaux « Like » et « Dislike » de ce volontaire.

Nous voyons que dans les deux catégories, les genres les plus représentés sont les mêmes : Jazz, Pop, Electro, alternative, Rock. Seul le genre Classique n'est présent que parmi les genres « Dislike ». Ces genres ont été obtenus à partir de l'API de Deezer, qui ne dispose pas d'une nomenclature précise en sous-genres.

Afin d'avoir plus de détails, nous avons interrogé l'API de Spotify afin de connaître les sous-genres des morceaux classés « Like » et « Dislike » par ce

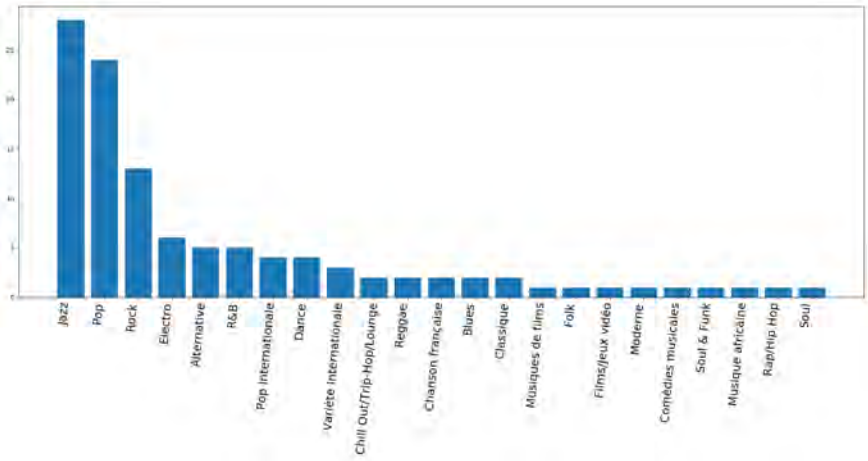

FIGURE 3.16: Genres des morceaux classés « + » et « ++ » par le volontaire #7, API Deezer.

volontaire. Sur les figures 3.18 et 3.19, nous voyons que les sous-genres présents dans les deux catégories sont différents.

Ainsi, nous pouvons penser que le système parvient à modéliser des différences subtiles entre différents sous-genres d'un même genre de musique. Par exemple, le genre jazz était très présent dans les deux catégories « Like » et « Dislike ». Avec les sous-genres de Spotify, nous apprenons que ce volontaire aime le Vocal jazz et le Jazz funk, mais qu'il n'aime pas le Free jazz (free improvisation, experimental, avant garde). Dans ce cas, le système a réussi à modéliser la différence entre les morceaux de ces différents sous-genres. L'aspect vocal d'un morceau de jazz peut être identifié grâce à un réseau de neurones via la signature timbrale de la voix [58]. Par ailleurs, le free jazz se distingue du jazz vocal et du jazz funk par son caractère plus « chaotique », tant au niveau rythmique (temporel) qu'harmonique.

Ce caractère est également identifiable au niveau du spectrogramme. En effet, sur les figures 3.20 et 3.21, nous observons une différence entre les spectrogrammes du morceau de free jazz et le morceau de vocal jazz.

Le morceau de Jazz vocal a un rythme clairement défini. Sur le spectrogramme de « Oye Como Va », la pulsation est clairement définie à travers les traits verticaux régulièrement espacés. Par ailleurs nous voyons également l'évolution des différentes notes chantées et jouées ainsi que leurs harmoniques : traits horizontaux à différentes fréquences. En revanche, sur le morceau de Han Bennink et Willem Breuker, « Mr M.A. de R. in A. »[4], nous n'observons aucune pulsation régulière. Nous pouvons néanmoins observer les notes de la trompette qui font leur apparition de manière déstructurée.

Ainsi, les résultats obtenus pour ce volontaire nous montrent que plusieurs

4. https://www.youtube.com/watch?v=aeupsqy_Pso&t=987s

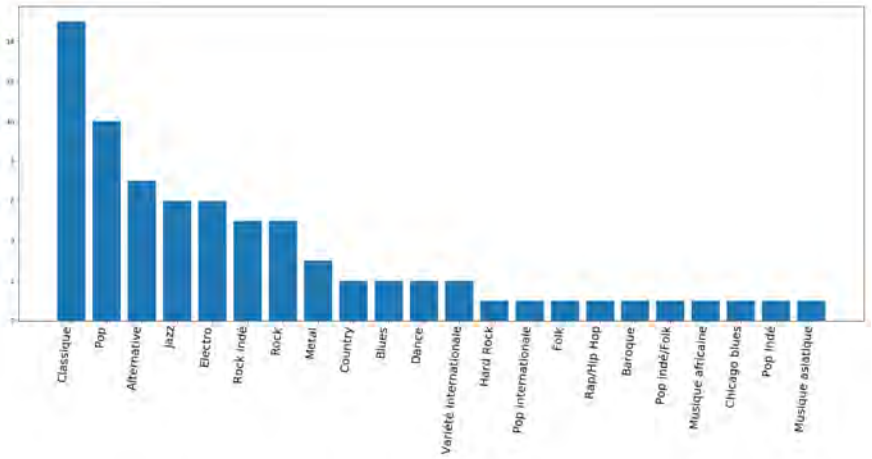

FIGURE 3.17: Genres des morceaux classés « - » et « - - » par le volontaire #7, API Deezer

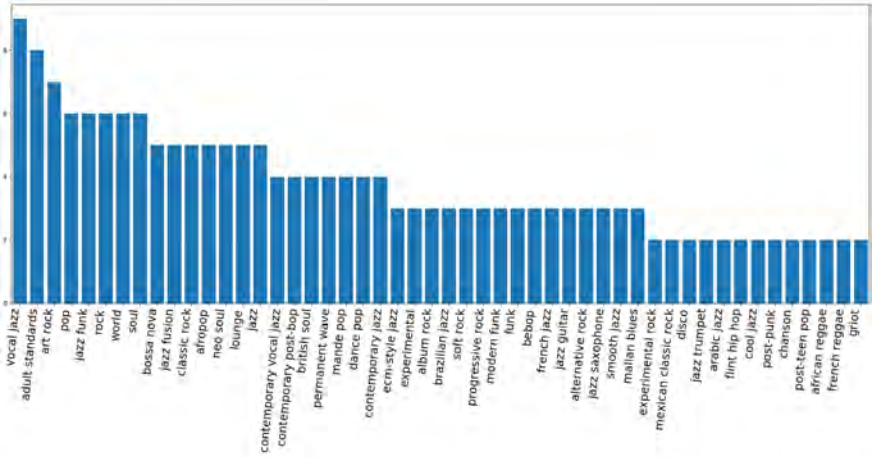

FIGURE 3.18: Genres des morceaux classés « + » et « ++ » par le volontaire #7, API Spotify.

morceaux annotés sous le même genre « grossiers » peuvent comporter de grandes différences. Ces différences se retrouvent dans les affinités des auditeurs et peuvent être explicitées par des annotations en sous-genres, plus fines. Par ailleurs, ces différences sont suffisamment marquées pour être détectées et traitées par des méthodes d'apprentissage automatique. Le cas des différents sous-genres de jazz a par exemple été étudié dans [89] avec de très bons résultats.

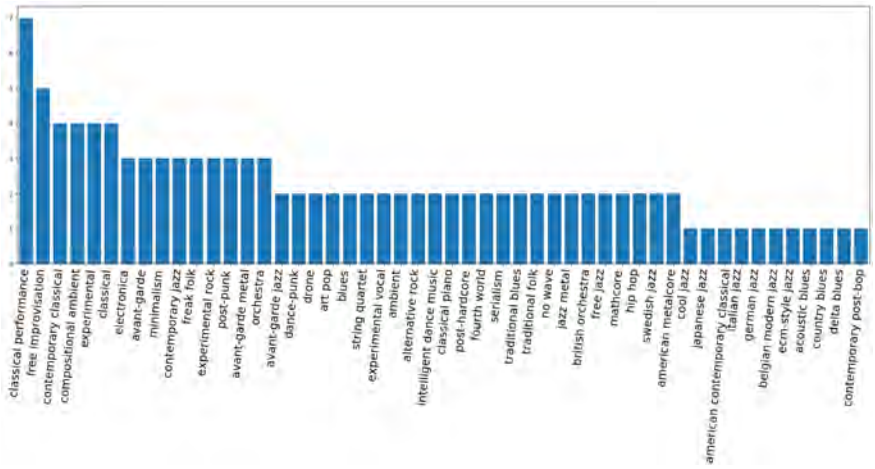

FIGURE 3.19: Genres des morceaux classés « - » et « - - » par le volontaire #7, API Spotify.

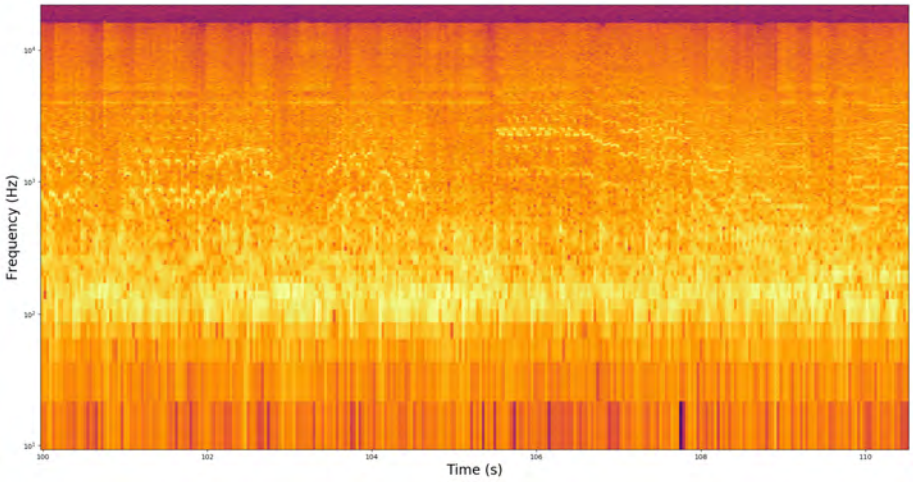

FIGURE 3.20: Spectrogramme d'un extrait du morceau « Mr M.A. de R. in A » par Han Bennink et Willem Breuker.

3.9 Conclusion

Dans cette partie, nous avons proposé une méthode de prédiction de goûts musicaux. Nous avons construit un modèle pour chaque volontaire à partir des données recueillies, et d'un pré-entraînement sur la reconnaissance de genres musicaux. Nous avons testé ces modèles dans un premier temps sur un jeu de test hors-ligne (évaluation « classique » automatique) puis à travers une évaluation humaine.

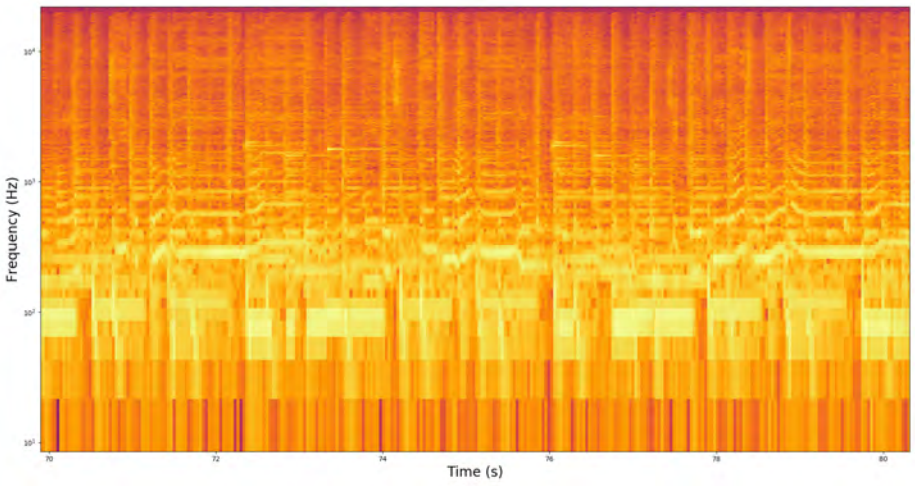

FIGURE 3.21: Spectrogramme d'un extrait du morceau « Oye Como Va » par Eliane Elias.

Les résultats obtenus sont satisfaisants pour une majorité de volontaires. Ces bons résultats montrent que des paramètres appris automatiquement par un réseau de neurones convolutionnel pour la reconnaissance de genres peuvent s'avérer pertinents pour la prédiction des goûts. Nous avons observé une décorrélation entre les tests automatiques et l'évaluation humaine. Par ailleurs, nous n'avons pas observé de lien entre la quantité de données par volontaire et la qualité de la modélisation.

L'étude qualitative des résultats a montré que notre système était bien capable d'apprendre les affinités et aversions des volontaires en se basant sur des modèles pré-entraînés à la reconnaissance en genres. Pour certains volontaires, l'utilisation d'une classification fine en « sous-genres » permet d'expliciter les goûts plus efficacement qu'avec une classification « grossière » en genres. Par ailleurs, les modèles utilisés sont capables de discerner des caractéristiques acoustiques propres à ces sous-genres.

Pour certains volontaires, le critère de genres/sous-genres semble avoir moins d'importance et l'utilisation de modèles pré-entraînés sur la prédiction de popularité pourrait apporter des améliorations.

Les morceaux contenus dans la base d'apprentissage des modèles doivent être aussi variés que possible en termes d'artistes et d'albums afin d'éviter un surapprentissage. De manière générale, un bon score sur une évaluation hors-ligne ne garantit pas de bons résultats sur une évaluation humaine.

Conclusions et perspectives

La recommandation musicale est un outil déterminant pour les plateformes de streaming, car elle permet à leurs utilisateurs de trouver des morceaux de musique à écouter parmi les millions de titres disponibles. Une recommandation plus efficace permettrait notamment aux artistes de niche de toucher un public plus large, et aux auditeurs de découvrir davantage d'artistes.

Dans le domaine académique, la recherche en recommandation de musique s'appuie sur l'intelligence artificielle et sur la recherche d'information musicale. Au cours de cette thèse, nous nous sommes focalisés sur les problématiques de la prédiction des goûts musicaux, dans un domaine où les incertitudes autour de l'explicabilité de ces goûts sont nombreuses :

- Quels éléments influencent les goûts musicaux ? Quels sont les liens entre les genres musicaux et les goûts des auditeurs en matière de musique ?
- La recommandation de musique est-elle une prédiction de goûts ? Comment les données comportementales sont-elles interprétées par les plateformes de streaming ? Le comportement traduit-il vraiment les goûts musicaux ?
- Les paramètres acoustiques appris par un réseau de neurones convolutionnels pour la reconnaissance automatique de genres sont-ils pertinents pour la prédiction de goûts musicaux ? Quels facteurs déterminants dans le goût musical un réseau de neurones convolutionnels est-il capable d'identifier dans un spectrogramme ?

Pour répondre à ces problématiques, nous avons donc développé notre recherche autour des goûts musicaux, du fonctionnement de l'IA, des données utilisées et de la place de l'humain dans ce processus.

Dans le chapitre 1, nous avons étudié les goûts musicaux, leurs liens avec les genres et les paramètres contextuels qui peuvent influencer l'écoute de musique. Après une étude bibliographique, nous avons analysé des données massives en provenance du site Deezer. D'une part, les données étudiées vont dans le sens des théories sociologiques sur les goûts musicaux. En particulier, nos observations sont en adéquation avec la théorie de Peterson sur les goûts univores et omnivores : les musiques populaires d'aujourd'hui (rap, hip-hop, pop, etc.) se caractérisent par un auditoire concentré sur peu de titres, contrairement aux musiques dites « savantes » (jazz, classique, etc.), dont l'auditoire se porte sur une quantité plus variée de morceaux. D'autre part, nos observations ont montré une accumulation des écoutes par un nombre très restreint d'artistes et de mor-

ceaux, confirmant la nécessité d'une amélioration des outils de recommandation de musique.

Dans le chapitre 2, nous nous sommes intéressés aux données, aux méthodes et aux modes d'évaluation employées par les plateformes de streaming pour la recommandation. Il est apparu que ces outils prédisent davantage des comportements que des goûts musicaux. En effet, les données comportementales (skip, durée d'écoute) sont bien plus abondantes que les données d'affinités (like, etc.), pourtant bien plus significatives et pertinentes pour la recommandation. Dans une volonté de replacer l'humain dans le processus de recommandation, nous avons mené une expérience de catégorisation. En utilisant une méthode de classification hiérarchique ascendante, nous avons représenté le résultat de cette catégorisation par un dendrogramme (voir partie 1.4.3). L'analyse de ce dendrogramme par des musicologues a montré que les critères dominants pour la catégorisation des extraits sonores reposaient principalement sur des critères de genres, d'instruments de musiques et sur l'appartenance ou non à la musique occidentale. Pour chacune des sous-parties du dendrogramme obtenu, nous avons pu estimer la catégorisation effectuée à partir de paramètres acoustiques. Pour ce faire, nous avons dû sélectionner ces paramètres indépendamment pour chaque sous-partie. Ainsi, nous avons obtenu un ensemble de paramètres acoustiques corrélés avec les résultats d'une catégorisation humaine. Pour la recommandation de musique, où des paramètres acoustiques sont parfois employés, la sélection de paramètres acoustiques d'après une catégorisation humaine amènerait à des paramètres plus pertinents.

Dans le dernier chapitre, en nous appuyant sur les conclusions des chapitres précédents, nous avons implémenté une prédiction de goûts musicaux basée sur des réseaux de neurones convolutionnels profonds pré-entraînés sur une prédiction de genres. Les résultats obtenus sur les 20 volontaires sont satisfaisants, avec près de 80% de précision moyenne sur les 3 premières et 5 premières prédictions ($P@3$ et $P@5$). Les résultats de cette prédiction ont été étudiés qualitativement afin de proposer des pistes d'amélioration pour la technique employée. Pour limiter le sur-apprentissage, il faudrait supprimer de la base d'apprentissage des morceaux provenant d'un même album ou artiste. Par ailleurs, les caractéristiques de la musique dite « commerciale » se sont révélées déterminantes pour un volontaire dans ses affinités, mais le système employé n'a pas été en mesure d'isoler des paramètres permettant de les discriminer. Pour résoudre ce problème, des informations sur la popularité des morceaux pourraient être utilisées, ou bien des modèles spécifiquement conçus pour estimer la popularité des morceaux.

Les liens entre les genres de musique et les affinités musicales ont été étudiés tout au long de cette thèse : durant l'expérience présentée en préambule, dans l'analyse des logs de Deezer, dans l'expérience de catégorisation libre et dans l'expérience de prédiction de goût du chapitre 3. Les résultats obtenus au cours de ces différentes expériences montrent que les genres peuvent être une première étape pour expliquer des affinités, mais ne sont en général pas suffisants. Des critères plus précis comme des sous-genres, des instruments de musique, le caractère « mainstream » ou commercial de morceaux sont régulièrement apparus

dans les témoignages et les donnés issues de nos expérimentations. Ainsi, des systèmes de classification automatique en genres ne sont pas suffisants pour prédire des goûts musicaux. Néanmoins, nous avons vu dans le dernier chapitre que ces modèles peuvent être utilisés comme une base robuste pour un apprentissage de goûts musicaux.

Au cours des travaux de cette thèse, nous avons vu que les algorithmes de recommandation musicale sont encore en voie d'amélioration. Ce qui est en cause ici c'est leur explicabilité, mais aussi l'explicabilité des goûts musicaux, qui dépendent de multiples facteurs. Par ailleurs, nous avons vu que les algorithmes employés sont davantage focalisés sur une prédiction du comportement des utilisateurs que sur la prédiction de leurs goûts. Ceci s'explique par une abondance de données comportementales contrairement aux données d'affinités. Les expériences menées en troisième partie nous ont montré que des affinités données explicitement par des volontaires sont pourtant prédictibles en utilisant une IA basée sur des spectrogrammes extraits du signal audio.

Perspectives

Des systèmes qui inciteraient davantage l'utilisateur à expliciter ses affinités et ses aversions seraient en mesure de disposer de données pertinentes pour une meilleure recommandation de musique personnalisée. Pour expliciter leurs affinités, les utilisateurs ont déjà à disposition un bouton « like » sur la majorité des plateformes de streaming. L'ajout d'un bouton « dislike » ou d'un équivalent permettrait aux plateformes de mieux connaître les aversions de leurs utilisateurs, et constituerait un gain d'information comparé au seul bouton « skip » qui n'est pas suffisamment explicite. Par ailleurs, cela conduirait à un échange plus transparent entre l'utilisateur et l'algorithme : en donnant explicitement ses goûts à un algorithme, l'utilisateur se sentirait davantage en contrôle des mécanismes sous-jacents, contrairement au fonctionnement « boîte noire » actuel. Ce gain de confiance associé à des recommandations plus pertinentes inciterait les utilisateurs à communiquer davantage leurs goûts. L'immense catalogue dont disposent les plateformes de streaming peut être utilisé afin de pré-entraîner un modèle de classification automatique en genres très robuste, en vue d'une utilisation des paramètres en « transfer learning ». Les paramètres appris à partir de millions de titres seraient capables d'identifier des nuances acoustiques plus précises que les modèles basés sur les corpus tels que GTZAN ou FMA utilisés durant cette thèse. Par ailleurs, en ne ré-entraînant qu'une partie du modèle pour chaque utilisateur, nous conservons ainsi l'autre partie inchangée qui est commune à tous les modèles. Le transfer learning permettrait ainsi une grande économie de stockage dans le cas d'un déploiement à grande échelle. La méthode proposée dans le chapitre 3 serait donc applicable à un cadre industriel : la base de données musicale dont dispose la plateforme serait utilisée pour le pré-entraînement et les modèles personnalisés de chaque utilisateur seraient construits à partir de leurs interactions via les boutons « like » et « dislike ».

Davantage d'études à grande échelle pourraient être menées dans le but

d'expliquer les comportements des utilisateurs face aux recommandations de musiques. C'est le cas de la recherche présentée dans [52], où les auteurs ont combiné des questionnaires à grande échelle avec les données comportementales des utilisateurs de Spotify interrogés. Ces études pourraient inclure davantage de données contextuelles (âge, activité, données sociologiques) qui permettraient de donner un complément d'explication aux données comportementales.

Durant cette thèse, nous avons mené des expérimentations sur les annotations musicologiques du corpus présenté dans la partie 2.6.1 du chapitre 2. Nous avons mis en oeuvre différentes méthodes de classification et de régression pour tenter d'estimer les différents critères musicologiques à partir de paramètres acoustiques. Les résultats obtenus n'ont pas été suffisamment concluants pour être présentés en détail dans ce mémoire. Le faible nombre d'extraits (100) ainsi que la complexité des critères musicologiques peuvent expliquer les mauvais résultats de ces expérimentations (proches de l'aléatoire). La plupart des critères annotés par les musicologues proviennent d'une transcription quantitative d'éléments décrits habituellement de manière qualitative, ce qui était la demande faite aux musicologues. De plus, ces critères s'expriment à différents niveaux de granularité : certains s'appliquent à un morceau entier et d'autres en quelques secondes seulement. La base de données ainsi constituée ne nous a pas permis de lier des paramètres acoustiques à des critères musicologiques par manque de robustesse des annotations, et par leur faible nombre.

Cependant, des bases de données, comme celle de Pandora[5], qui comportent une quantité massive d'annotations musicologiques, peuvent être mises à contribution. La grande quantité de morceaux disponibles ainsi que leur protocole d'annotation standardisé devrait permettre l'apprentissage robuste d'algorithmes d'intelligence artificielle. Ces IA ainsi entraînées permettraient d'annoter automatiquement un plus grand nombre de morceaux, en vue d'une recommandation musicale basée sur des critères musicologiques. Il serait ainsi possible de construire un corpus multidisciplinaire comportant des données comportementales, musicologiques, sociomusicologiques et acoustiques. L'analyse de ces données pourrait s'effectuer à travers différentes grilles de lectures, mais aussi en combinant ces différents champs disciplinaires notamment grâce à des outils d'intelligence artificielle. L'inclusion des sciences humaines dans la construction de modèles de comportements et d'affinités aurait pour intérêt principal d'améliorer leur explicabilité, qui constitue bien souvent l'inconvénient majeur de l'intelligence artificielle.

Dans un autre contexte, ces résultats pourraient s'appliquer à la musique générée automatiquement via l'intelligence artificielle. En effet, celle-ci demeure toujours un challenge de taille, mais a connu des progrès ces dernières années [21]. À plus long terme, nous pourrions imaginer des systèmes qui, au lieu de choisir un morceau dans une base de données, généreraient de manière automatique de la musique en accord avec le contexte et le goût de l'utilisateur. Par exemple, un modèle pourrait à partir de l'activité, de l'heure de la journée et des goûts musicaux d'un utilisateur déterminer un ensemble de paramètres pour

5. https://www.pandora.com/about/mgp

générer une musique en adéquation avec le contexte. Au-delà de simplement générer une musique énergique lorsque l'utilisateur fait du sport ou une musique calme lorsqu'il travaille, ces algorithmes pourraient respecter des critères précis selon les affinités de l'utilisateur vis-à-vis du contexte, tels que le bon timbre de voix, une instrumentation particulière, etc.

Annexe A

Liste des publications

Dauban, N., Albenge, P., Florin, L., Pinquier, J., Sénac, C., Gaillard, P., & Guyot, P. (2018). Catégorisation libre d'extraits musicaux et analyse automatique. Dans *15e Conférence en Recherche d'Information et Applications (CORIA 2018)* [36]

Dauban, N., Sènac, C., Pinquier, J., Gaillard, P., Florin, L., & Albenge, P. (2019, April). Automatic Analysis and Musicological Interpretation of Human Free Sorting of Musical Excerpts. Dans *11th International Conference on Advances in Multimedia (MMEDIA 2019)* (pp. 42-48). [38]

Dauban, N., Sénac, C., Pinquier, J., & Gaillard, P. (2021, June). Towards a content-based prediction of personalized musical preferences using transfer learning. Dans *2021 International Conference on Content-Based Multimedia Indexing (CBMI)* (pp. 1-6). IEEE. [37]

Annexe B

Genres dans les morceaux du top 1000

TABLE B.1: Genres : Nombre d'occurences dans les morceaux du top 1000

Genre	Occurences
Rap/Hip Hop	554
Pop	145
Rap français	106
Dance	71
R&B	45
Rock	35
Alternative	32
Electro	32
Films/Jeux vidéo	13
Musiques de films	12
Pop internationale	12
Techno/House	10
Chanson française	10
Singer & Songwriter	9
Latino	9
Rock indé	5
Variété Internationale	5
Hard Rock	4
Metal	4
Reggae	3
Pop Indé	2
Soul	2
Soul & Funk	2
Pop indé/Folk	1
Musique asiatique	1
Musique brésilienne	1
R&B contemporain	1
Rock français	1
Soul contemporaine	1
Comédies musicales	1

TABLE B.2: Genres : Nombre d'écoutes dans les morceaux du top 1000

Genre	Ecoutes
Rap/Hip Hop	600223
Pop	167032
Rap français	105159
Dance	72434
R&B	56101
Rock	30893
Electro	29945
Alternative	22973
Pop internationale	15399
Films/Jeux vidéo	12193
Singer & Songwriter	11697
Musiques de films	10952
Latino	9635
Chanson française	9397
Techno/House	9254
Rock indé	4522
Reggae	3755
Metal	3439
Variété Internationale	3198
Pop Indé	2243
Hard Rock	2039
Soul & Funk	1214
Comédies musicales	1155
Soul	943
Pop indé/Folk	873
Rock français	873
Musique brésilienne	717
R&B contemporain	610
Musique asiatique	516
Soul contemporaine	497

TABLE B.3: Nombre d'occurences, d'écoutes, et nombre d'écoutes par occurence, pour tous les genres (partie 1).

Genre	**Occurences**	**Ecoutes**	**Ratio**
Pop	46376	542694	11,7
Rap/Hip Hop	35913	1272009	35,4
Rock	24905	169384	6,8
Electro	17998	135641	7,5
Dance	17537	205998	11,7
Alternative	15668	128410	8,2
Films/Jeux vidéo	8873	60750	6,8
Musiques de films	8106	55426	6,8
R&B	6883	123516	17,9
Reggae	4794	34071	7,1
Classique	4453	11889	2,7
Rap français	4443	233260	52,5
Metal	4289	19369	4,5
Chanson française	4181	39781	9,5
Pop internationale	4029	44802	11,1
Jazz	3706	10726	2,9
Variété Internationale	3388	24902	7,4
Techno/House	3184	30539	9,6
Rock indé	2775	19820	7,1
Latino	2604	34174	13,1
Hard Rock	2337	12730	5,4
Singer & Songwriter	1979	24071	12,2
Soul	1927	12161	6,3
Musique africaine	1831	6766	3,7
Jeunesse	1588	7839	4,9
Rock Indé/Pop Rock	1586	8280	5,2
Pop Indé	1529	9612	6,3
Folk	1282	6426	5,0
Disco	1135	4998	4,4
Musique arabe	1024	3582	3,5
Trance	921	2423	2,6
Dancehall/Ragga	895	4922	5,5
Pop indé/Folk	853	7100	8,3
Blues	843	2595	3,1
Soul & Funk	819	5869	7,2
Musique asiatique	777	5580	7,2
Country	757	2968	3,9
Musique brésilienne	643	3345	5,2
Rock & Roll/Rockabilly	506	1776	3,5
Chill Out/Trip-Hop/Lounge	494	2133	4,3
Bandes originales	447	1631	3,6
Dub	446	4326	9,7
R&B contemporain	434	4519	10,4
Comédies musicales	377	3362	8,9
Livres audio	299	999	3,3
Comédie	269	609	2,3
BO TV	250	976	3,9

TABLE B.4: Nombre d'occurences, d'écoutes, et nombre d'écoutes par occurence, pour tous les genres (partie 2).

Genre	**Occurences**	**Ecoutes**	**Ratio**
Opéra	239	454	1,9
Soul contemporaine	230	1226	5,3
Rock français	214	3243	15,2
East Coast	213	1383	6,5
Musique indienne	195	723	3,7
Dubstep	179	457	2,6
Jazz instrumental	103	284	2,8
Bollywood	101	266	2,6
Sports	85	144	1,7
Ska	82	247	3,0
Baroque	80	149	1,9
Musiques de jeux vidéo	70	205	2,9
Pop française	66	879	13,3
Dirty South	65	479	7,4
Bolero	63	246	3,9
Comptines/Chansons	63	264	4,2
Jazz vocal	58	149	2,6
Electro Pop/Electro Rock	55	231	4,2
Spiritualité & religion	53	158	3,0
Dancefloor	47	874	18,6
Jazz Hip Hop	43	180	4,2
Electro Hip-Hop	37	149	4,0
Tropical	37	139	3,8
Moderne	23	88	3,8
Blues acoustique	22	46	2,1
Grime	21	140	6,7
Chicago blues	20	36	1,8
Période classique	19	21	1,1
West Coast	14	74	5,3
R&B vieille école	13	23	1,8
Soul vieille école	12	17	1,4
Blues électrique	11	18	1,6
Old School	6	31	5,2
Médieval	5	11	2,2
Romantique	5	11	2,2
Ranchera	4	5	1,3
Country blues	3	49	16,3
Corridos	1	1	1,0
Histoires	1	1	1,0
Renaissance	1	1	1,0
Blues classique	1	1	1,0
Éducation	1	1	1,0
Enfants & famille	1	1	1,0

Annexe C

Extrait du log Deezer


```
[{"anon_user_id":"????","media_id":460260772,"info_conn":"mobile","platform_name":"ios
{"anon_user_id":"????","media_id":437148202,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":454423012,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":408868862,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":440529852,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":408868832,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":464920482,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":408868832,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":481339432,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":128716605,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":482719392,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":381546731,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":482719392,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":381546731,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":471750652,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":381546731,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":456247992,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":381546731,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":482719392,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":447098092,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":7180013,"info_conn":"mobile","platform_name":"ios","
{"anon_user_id":"????","media_id":142082325,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":137036530,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":366148531,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":461704022,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":1097647,"info_conn":"mobile","platform_name":"ios","
{"anon_user_id":"????","media_id":461704022,"info_conn":"mobile","platform_name":"ios"
{"anon_user_id":"????","media_id":1097647,"info_conn":"mobile","platform_name":"ios","
{"anon_user_id":"????","media_id":438850292,"info_conn":"mobile","platform_name":"ios"
```


FIGURE C.1: Extrait du log de Deezer : identifiant anonyme de l'utilisateur (masqué ici), identifiant du morceau, type de connexion, os de l'utilisateur.

```
","click_next":1,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1528034422},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":9,"ts_listen":1527953173},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":120,"ts_listen":1528034259},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":189,"ts_listen":1527953187},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":190,"ts_listen":1528031495},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1527935012},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":179,"ts_listen":1528035457},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1527955417},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":40,"ts_listen":1528028100},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1527952385},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":148,"ts_listen":1528023829},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":111,"ts_listen":1527938129},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":158,"ts_listen":1528024060},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":2,"ts_listen":1527937756},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1528034395},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":111,"ts_listen":1527939221},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":217,"ts_listen":1528035637},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":3,"ts_listen":1527937897},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":48,"ts_listen":1528024857},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":31,"ts_listen":1527937565},
click_next":1,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1528034384},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":157,"ts_listen":1527960590},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1528034403},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":17,"ts_listen":1527953155},
,"click_next":0,"click_loved":0,"click_banned":0,"listening_time":173,"ts_listen":1528034429},
click_next":0,"click_loved":0,"click_banned":0,"listening_time":217,"ts_listen":1527928232},
,"click_next":1,"click_loved":0,"click_banned":0,"listening_time":1,"ts_listen":1528034424},
```


FIGURE C.2: Extrait du log de Deezer : skip, like, ban, durée d'écoute, moment d'écoute.

Annexe D

MIR Toolbox

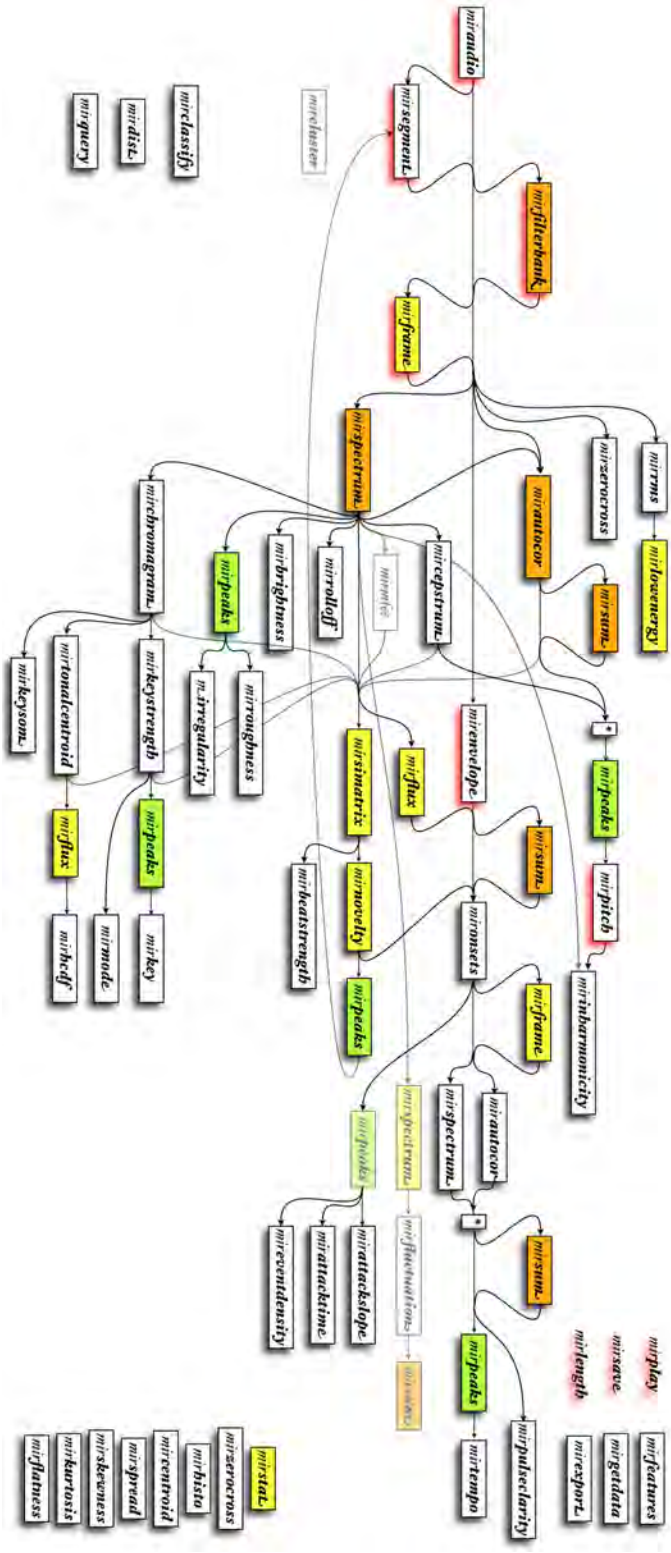

FIGURE D.1: Schéma représentant les différentes fonctions de la toolbox MIR.

Annexe E

26 morceaux PERMUSES

TABLE E.1: Les 26 morceaux dont les extraits (1 extrait par morceau) sont à trier.

Morceau	**Artiste**
Le feu de joie	Bénabar
La fabuleuse histoire des compagnons	Ptits t'hommes
Les pêcheurs de perles	Bizet
Les Hébrides	Mendelssohn
Glory Box	Portishead
La mort sur le dancefloor	Vitalic
It's Only a Paper Moon	Ella Fitzgerald
Mr PC	John Coltrane
Waldschrein	Equilibrium
Wanderer	Ensiferum
Accroche à ma terre	Frangines (Les)
Crystal Clear	Pharrell Williams
Tout brûle déjà	La Rumeur
Logo dans le ciel	Orelsan
Good Morning Midnight	Biga Ranx
Johnny Was	Bob Marley
Ride	Lolo Zouai
Sweet Time	Raveena
Disorder	Joy Division
Lithium	Nirvana
Next to You	Ben l'Oncle Soul
Baby It's You	Smith
In the Halls of the Usurper	Jake Kaufman
Let the Games Begin	Harry Gregson-Williams
Musique traditionnelle japonaise	Inconnu
Chant arabo-andalou	Inconnu

Bibliographie

[1] Charu C Aggarwal. Content-based recommender systems. In *Recommender Systems*, pages 139–166. Springer, 2016.

[2] Chris Anderson. *The long tail : Why the future of business is selling less of more*. Hachette Books, 2006.

[3] Chris Anderson and Mia Poletto Andersson. Long tail. 2004.

[4] Will Atkinson. The context and genesis of musical tastes : Omnivorousness debunked, bourdieu buttressed. *Poetics*, 39(3) :169–186, 2011.

[5] Jean-Julien Aucouturier, Francois Pachet, et al. Music similarity measures : What's the use ? In *ISMIR*, pages 13–17, 2002.

[6] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.

[7] Linas Baltrunas and Xavier Amatriain. Towards time-dependant recommendation based on implicit feedback. In *Workshop on context-aware recommender systems (CARS'09)*, pages 25–30. Citeseer, 2009.

[8] Jean Pierre Barthélémy and Alain Guénoche. Les arbres et les représentations de proximité. *Paris. Dunod (english translation : Trees and Proximity Representations, New York, Wiley, 1991)*, 1988.

[9] Christine Bauer and Markus Schedl. Global and country-specific mainstreaminess measures : Definitions, analysis, and usage for improving personalized music recommendation systems. *PloS one*, 14(6) :e0217389, 2019.

[10] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. paper recommender systems : a literature survey. *International Journal on Digital Libraries*, 17(4) :305–338, 2016.

[11] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription : An overview. *IEEE Signal Processing Magazine*, 36(1) :20–30, 2018.

[12] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1) :281–305, 2012.

[13] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[14] Jean-Samuel Beuscart, Samuel Coavoux, and Sisley Maillard. Les algorithmes de recommandation musicale et l'autonomie de l'auditeur. *Reseaux*, (1) :17–47, 2019.

[15] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Emilia Gómez, and Perfecto Herrera. Content-based music recommendation based on user preference examples. In *ACM conf. on recommender systems. Workshop on music recommendation and discovery (Womrad 2010)*, 2010.

[16] Ludovico Boratto, Salvatore Carta, Gianni Fenu, and Roberto Saia. Semantics-aware content-based recommender systems : Design and architecture guidelines. *Neurocomputing*, 254 :79–85, 2017.

[17] Pierre Bourdieu. *La distinction : critique sociale du jugement*. Minuit, 2016.

[18] Marc Bourreau, Sisley Maillard, and François Moreau. Une analyse économique du phénomène de la longue traîne dans les industries culturelles. *Revue française d'économie*, 30(2) :179–216, 2015.

[19] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[20] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[21] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. *Deep learning techniques for music generation*. Springer, 2020.

[22] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. The music streaming sessions dataset. In *The World Wide Web Conference*, pages 2594–2600, 2019.

[23] Jiajun Bu, Shulong Tan, Chun Chen, Can Wang, Hao Wu, Lijun Zhang, and Xiaofei He. Music recommendation by unified hypergraph : combining social media information and music content. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 391–400, 2010.

[24] Erion Çano and Maurizio Morisio. Hybrid recommender systems : A systematic literature review. *Intelligent Data Analysis*, 21(6) :1487–1524, 2017.

[25] Òscar Celma Herrada et al. *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra, 2009.

[26] Tak Wing Chan and John H Goldthorpe. Social stratification and cultural consumption : Music in england. *European sociological review*, 23(1) :1–19, 2007.

[27] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv :1910.05446*, 2019.

[28] Parag Chordia, Mark Godfrey, and Alex Rae. Extending content-based recommendation : The case of indian classical music. In *ISMIR*, pages 571–576, 2008.

[29] Szu-Yu Chou, Yi-Hsuan Yang, Jyh-Shing Roger Jang, and Yu-Ching Lin. Addressing cold start for next-song recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 115–118, 2016.

[30] Manuel Pacheco Coelho and José Zorro Mendes. Digital music and the "death of the long tail". *Journal of Business Research*, 101 :454–460, 2019.

[31] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[32] Yandre MG Costa, Luiz S Oliveira, and Carlos N Silla Jr. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied soft computing*, 52 :28–38, 2017.

[33] Philippe Coulangeon. La stratification sociale des goûts musicaux. *Revue française de sociologie*, 44(1) :3–33, 2003.

[34] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1) :21–27, 1967.

[35] Sally Jo Cunningham, David Bainbridge, and Annette Falconer. " more of an art than a science" : Supporting the creation of playlists and mixes. 2006.

[36] Nicolas Dauban, Paul Albenge, Ludovic Florin, Julien Pinquier, Christine Sénac, Pascal Gaillard, and Patrice Guyot. Catégorisation libre d'extraits musicaux et analyse automatique. 2018.

[37] Nicolas Dauban, Christine Sénac, Julien Pinquier, and Pascal Gaillard. Towards a content-based prediction of personalized musical preferences using transfer learning. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2021.

[38] Nicolas Dauban, Christine Sènac, Julien Pinquier, Pascal Gaillard, Ludovic Florin, and Paul Albenge. Automatic analysis and musicological interpretation of human free sorting of musical excerpts. In *11th International Conference on Advances in Multimedia (MMEDIA 2019)*, pages 42–48, 2019.

[39] Michaël Defferrard, Sharada P. Mohanty, Sean F. Carroll, and Marcel Salathé. Learning to recognize musical genre from audio. In *WWW '18 Companion : The 2018 Web Conference Companion*, 2018.

[40] Ricardo Dias and Manuel J Fonseca. Improving music recommendation in session-based collaborative filtering by using temporal context. In *2013 IEEE 25th international conference on tools with artificial intelligence*, pages 783–788. IEEE, 2013.

[41] Paul DiMaggio and John Mohr. Cultural capital, educational attainment, and marital selection. *American journal of sociology*, 90(6) :1231–1261, 1985.

[42] Consistent Distinguishability. A theoretical analysis of normalized discounted cumulative gain (ndcg) ranking measures. 2013.

[43] Anthony William Fairbank Edwards. *Likelihood*. CUP Archive, 1984.

[44] Andres Ferraro, Dmitry Bogdanov, Xavier Serra, and Jason Yoon. Artist and style exposure bias in collaborative filtering based music recommendations. *arXiv preprint arXiv :1911.04827*, 2019.

[45] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. Personality traits and music genre preferences : how music taste varies over age groups. In *1st Workshop on Temporal Reasoning in Recommender Systems (RecTemp) at the 11th ACM Conference on Recommender Systems, Como, August 31, 2017.*, volume 1922, pages 16–20. CEUR-WS, 2017.

[46] Arthur Flexer and Thomas Grill. The problem of limited inter-rater agreement in modelling music similarity. *Journal of new music research*, 45(3) :239–251, 2016.

[47] Arthur Flexer and Taric Lallai. Can we increase inter-and intra-rater agreement in modeling general music similarity ?. In *ISMIR*, pages 494–500, 2019.

[48] Arthur Flexer and Dominik Schnitzer. Effects of album and artist filters in audio similarity computed for very large music databases. *Computer Music Journal*, 34(3) :20–28, 2010.

[49] Vreixo Formoso, Diego FernáNdez, Fidel Cacheda, and Victor Carneiro. Using profile expansion techniques to alleviate the new user problem. *Information processing & management*, 49(3) :659–672, 2013.

[50] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *IEEE transactions on multimedia*, 13(2) :303–319, 2010.

[51] Pascal Gaillard. Laissez-nous trier ! tcl-labx et les tâches de catégorisation libre de sons. pages 189–210, 2009.

[52] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 55–64, 2018.

[53] Hervé Glevarec and Michel Pinet. La «tablature» des goûts musicaux : un modèle de structuration des préférences et des jugements. *Revue française de sociologie*, 50(3) :599–640, 2009.

[54] Kuldeep Gurjar and Yang-Sae Moon. Comparative analysis of music similarity measures in music information retrieval systems. *Journal of Information Processing Systems*, 14(1), 2018.

[55] Christopher Harte, Mark Sandler, and Martin Gasser. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 21–26. ACM, 2006.

[56] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1) :1–12, 2004.

[57] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1) :5–53, 2004.

[58] Yuanbo Hou, Frank K Soong, Jian Luan, and Shengchen Li. Transfer learning for improving singing-voice detection in polyphonic instrumental music. *arXiv preprint arXiv :2008.04658*, 2020.

[59] Kristoffer Jensen. Pitch independent prototyping of musical sounds. In *IEEE 3rd Workshop on Multimedia Signal Processing*, pages 215–220. IEEE, 1999.

[60] Olivier Jeunen and Bart Goethals. Predicting sequential user behaviour with session-based recurrent neural networks. 2019.

[61] Mohsen Kamalzadeh, Dominikus Baur, and Torsten Möller. A survey on music listening and management behaviours. 2012.

[62] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8) :5455–5516, 2020.

[63] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8) :30–37, 2009.

[64] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 208–211, 2008.

[65] Olivier Lartillot. Mirtoolbox 1.6.1 user's manual, 2014.

[66] Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, and Jose Fornari. Multi-feature modeling of pulse clarity : Design, validation and optimization. In *ISMIR*, pages 521–526, 2008.

[67] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10) :1995, 1995.

[68] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553) :436–444, 2015.

[69] Junghyuk Lee and Jong-Seok Lee. Music popularity : Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11) :3173–3182, 2018.

[70] Ning-Han Liu. Comparison of content-based music recommendation using different distance estimation methods. *Applied intelligence*, 38(2) :160–174, 2013.

[71] Omar Lizardo and Sara Skiles. After omnivorousness. *Routledge International Handbook of the Sociology of Art and Culture. New York : Routledge*, pages 90–103, 2015.

[72] Alex Lopez-Suarez and M Kamel. Dykor : a method for generating the content of explanations in knowledge systems. *Knowledge-Based Systems*, 7(3) :177–188, 1994.

[73] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov) :2579–2605, 2008.

[74] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[75] David Martín-Gutiérrez, Gustavo Hernández Peñaloza, Alberto Belmonte-Hernández, and Federico Álvarez García. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8 :39361–39374, 2020.

[76] Mark Mulligan. The death of the long tail : The superstar music economy. *Music Industry Blog*, 4, 2014.

[77] Andrew Y Ng et al. Preventing" overfitting" of cross-validation data. In *ICML*, volume 97, pages 245–253. Citeseer, 1997.

[78] Rosalie Ollivier, Louise Goupil, and Jean-Julien Aucouturier. Enjoy the violence : Is appreciation for extreme music the result of higher-order cognitive control over the threat response system ? 2018.

[79] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. A deep multimodal approach for cold-start music recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, pages 32–37, 2017.

[80] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. One-class collaborative filtering. In *2008 Eighth IEEE International Conference on Data Mining*, pages 502–511. IEEE, 2008.

[81] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10) :1345–1359, 2009.

[82] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. Novel audio features for music emotion recognition. *IEEE transactions on affective computing*, 2018.

[83] Vilfredo Pareto. Essai sur la courbe de la répartition de la richesses. In *Faculté de droit à l'occasion de l'exposition nationale suisse, Genève, Université de Lausanne*, 1896.

[84] Etienne Parizet and Vincent Koehl. Application of free sorting tasks to sound quality experiments. *Applied Acoustics*, 73(1) :61–65, 2012.

[85] Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 21–28, 2009.

[86] Richard A Peterson. Understanding audience segmentation : From elite and mass to omnivore and univore. *Poetics*, 21(4) :243–258, 1992.

[87] Martin Pichl and Eva Zangerle. Latent feature combination for multi-context music recommendation. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2018.

[88] Nick Prior. Bourdieu and the sociology of music consumption : A critical assessment of recent developments. *Sociology Compass*, 7(3) :181–193, 2013.

[89] R. J. M. Quinto, R. O. Atienza, and N. M. C. Tiglao. Jazz music subgenre classification using deep learning. In *TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 3111–3116, 2017.

[90] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741, 2009.

[91] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[92] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3) :211–252, 2015.

[93] Shaghayegh Sahebi and William W Cohen. Community-based recommendations : a solution to the cold start problem. In *Workshop on recommender systems and the social web, RSWEB*, page 60, 2011.

[94] Thomas Schäfer, Fee Auerswald, Ina Kristin Bajorat, Nika Ergemlidze, Katharina Frille, Jonas Gehrigk, Anastasia Gusakova, Bernadette Kaiser, Rosi Anna Pätzold, Ana Sanahuja, et al. The effect of social feedback on music preference. *Musicae Scientiae*, 20(2) :263–268, 2016.

[95] Markus Schedl. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 103–110, 2016.

[96] Markus Schedl. Deep learning in music recommendation systems. *Frontiers in Applied Mathematics and Statistics*, 5 :44, 2019.

[97] Markus Schedl and Arthur Flexer. Putting the user in the center of music information retrieval. In *ISMIR*, pages 385–390. Citeseer, 2012.

[98] Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3) :523–539, 2013.

[99] Markus Schedl and David Hauger. Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*, pages 947–950, 2015.

[100] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2) :95–116, 2018.

[101] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.

[102] Nick Seaver. Captivating algorithms : Recommender systems as traps. *Journal of Material Culture*, 24(4) :421–436, 2019.

[103] Christine Senac, Thomas Pellegrini, Florian Mouret, and Julien Pinquier. Music feature maps with convolutional neural networks for music genre classification. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, pages 1–5, 2017.

[104] Nesma Settouti, Mohammed El Amine Bechar, and Mohammed Amine Chikh. Statistical comparisons of the top 10 algorithms in data mining for classification task. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(1) :46–51, 2016.

[105] Bo Shao, Dingding Wang, Tao Li, and Mitsunori Ogihara. Music recommendation based on acoustic features and user access patterns. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8) :1602–1611, 2009.

[106] Diego F Silva, Chin-Chia M Yeh, Yan Zhu, Gustavo EAPA Batista, and Eamonn Keogh. Fast similarity matrix profile for music analysis and exploration. *IEEE Transactions on Multimedia*, 21(1) :29–38, 2018.

[107] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks : Visualising image classification models and saliency maps. *arXiv preprint arXiv :1312.6034*, 2013.

[108] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1) :1929–1958, 2014.

[109] Bob L Sturm. An analysis of the gtzan music genre dataset. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 7–12, 2012.

[110] Nava Tintarev and Judith Masthoff. Explaining recommendations : Design and evaluation. In *Recommender systems handbook*, pages 353–382. Springer, 2015.

[111] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5) :293–302, 2002.

[112] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265. IEEE, 2017.

[113] Andreu Vall, Hamid Eghbal-Zadeh, Matthias Dorfer, Markus Schedl, and Gerhard Widmer. Music playlist continuation by learning from hand-curated examples and song features : Alleviating the cold-start problem for rare and out-of-set songs. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, pages 46–54, 2017.

[114] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013.

[115] Koen Van Eijck. The impact of family background and educational attainment on cultural consumption : A sibling analysis. *Poetics*, 25(4) :195–224, 1997.

[116] Charles Van Loan. *Computational frameworks for the fast Fourier transform*. SIAM, 1992.

[117] Thomas Völkel, Jakob Abeßer, Christian Dittmar, and Holger Großmann. Automatic genre classification of latin american music using characteristic rhythmic patterns. In *Proceedings of the 5th Audio Mostly Conference : A Conference on Interaction with Sound*, pages 1–7, 2010.

[118] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen. Balanced distribution adaptation for transfer learning. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1129–1134, 2017.

[119] Xinxi Wang, David Rosenblum, and Ye Wang. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 99–108, 2012.

[120] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.

[121] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3) :37–52, 1987.

[122] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1) :1–37, 2008.

[123] Baolin Yi, Xiaoxuan Shen, Hai Liu, Zhaoli Zhang, Wei Zhang, Sannyuya Liu, and Naixue Xiong. Deep matrix factorization with implicit feedback embedding for recommendation system. *IEEE Transactions on Industrial Informatics*, 15(8) :4591–4601, 2019.

[124] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. Challenging the long tail recommendation. *arXiv preprint arXiv :1205.6700*, 2012.

[125] Eva Zangerle and Martin Pichl. Content-based user models : Modeling the many faces of musical preference. In *19th International Society for Music Information Retrieval Conference*, 2018.

[126] Eva Zangerle, Martin Pichl, and Markus Schedl. User models for culture-aware music recommendation : Fusing acoustic and cultural cues. *Transactions of the International Society for Music Information Retrieval*, 3(1), 2020.

[127] Eva Zangerle, Michael Vötter, Ramona Huber, and Yi-Hsuan Yang. Hit song prediction : Leveraging low-and high-level audio features. In *ISMIR*, pages 319–326, 2019.

[128] Bingjun Zhang, Jialie Shen, Qiaoliang Xiang, and Ye Wang. Compositemap : a novel framework for music similarity measure. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 403–410, 2009.

[129] Weibin Zhang, Wenkang Lei, Xiangmin Xu, and Xiaofeng Xing. Improved music genre classification with convolutional neural networks. In *Interspeech*, pages 3304–3308, 2016.

Table des figures

Liste des tableaux