Mémoire d'Habilitation à Diriger des Recherches
Université Pierre et Marie Curie (Université Paris VI)

par
Geoffroy Peeters

Indexation automatique de contenus audio musicaux

Soutenue le 12 avril 2013,
devant le jury composé de :

Mme	Régine	André-Obrecht	Rapporteur
M.	François	Pachet	Rapporteur
M.	Xavier	Serra	Rapporteur
M.	Jean-Dominique	Polack	Examinateur
M.	Gaël	Richard	Examinateur

Travail effectué à STMS-IRCAM-CNRS-UPMC — 1, Place Igor Stravinsky — 75004 Paris — France
Version : 1.2 — 16 avril 2013

Remerciements

Je remercie Régine André-Obrecht, François Pachet, Xavier Serra, Jean-Dominique Polack et Gaël Richard d'avoir accepté d'être membres du jury pour évaluer mon travail de recherche pour l'obtention d'une Habilitation à Diriger des Recherches.

Je suis très reconnaissant à Hugues Vinet pour avoir soutenu ces recherches pendant toutes ces années.

Je remercie l'ANR, Oseo et la Commission Européenne pour avoir financé les projets ayant permis le développement de ces recherches.

Je remercie toutes les personnes avec qui j'ai pu collaborer au cours des recherches de projets, que j'ai pu encadrer lors de thèses ou de stages de Master ; ainsi que tous les développeurs qui ont permis le passage de ces recherches dans des logiciels. Ils se reconnaitront dans ce document.

Je remercie également l'ensemble des membres de l'équipe Analyse/Synthèse des sons de l'IRCAM sans qui ces travaux n'auraient pu être menés. Je remercie toutes les personnes internes ou externes à l'IRCAM avec qui j'ai pu échanger, amorcer et développer ma réflexion sur l'indexation musicale.

Je remercie tout le personnel de l'IRCAM, en particulier Sylvie Benoit, pour m'avoir aidé dans la réalisation des évènements que j'ai pu organisés ou des missions scientifiques que j'ai éffectuées. Je remercie Carlos Agon pour m'avoir aidé dans les tâches adminsitratives relatives à ce dossier de HDR.

Enfin, je remercie mes amis, ma famille et en particulier Kim, Jérôme et Christel pour m'avoir motivé et aidé dans la relecture de ce document.

Table des matières

Index

Dictionnaire des acronymes

ANR	Agence National de la Recherche
ASAD	Algorithme de Sélection Automatique de Descripteurs
ARM	Modèle Auto-Régressif Multivarié
BS	Beat-Synchronous
CQT	Constant Q-Transform
DTW	Dynamic Time Warping
ERB	Equivalent Rectangular Bandwidth
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IHM	Interface Homme-Machine
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
MDS	Multi-Dimensional Scaling analysis
MIR	Music Infomation Retrieval
MFCCs	Mel Frequency Cepstral Coefficients
NMF	Non Negative Matrix Factorization
PCA	Principal Component Analysis
SVM	Support Vector Machine
TECC	True Envelope Cepstral Coefficient
TFD	Transformée de Fourier Discrète
TFCT	Transformée de Fourier à Court Terme
UBM	Universal Background Model

CHAPITRE 1

Introduction

Dans ce document je décris les recherches que j'ai effectuées ou encadrées depuis ma soutenance de thèse en juillet 2001 [1]. J'y adjoins également celles réalisées à partir de 1999, effectuées en dehors de ma thèse, puisque celles-ci en constituent le démarrage.

Contexte. Ces recherches débutent en octobre 1999, date à laquelle je suis envoyé en mission à Melbourne pour participer au meeting du comité de normalisation ISO/MPEG et introduire les descripteurs audio dits « de timbre » [2] dans la norme alors naissante MPEG-7 (ISO/IEC 15938) [118]. En **1999**, la communauté ISMIR n'existe pas [3], les recherches sur l'indexation audio se concentrent sur la parole, sur la segmentation parole/musique (comme celles de Scheirer [199] ou [196] [30] [193]), sur la reconnaissance des instruments de musique par approche CASA [4] (comme celles de Martin [105]) ou sur les premiers algorithmes d'estimation des battements à partir de l'audio (comme ceux de Goto [65] ou Klapuri [87]). Les évaluations s'effectuent alors sur quelques fichiers, les représentations « objets » de sources audio suscitent un intérêt croissant (MPEG-4 SAOL [201] ou AudioBIFS [200]), seules quelques applications reposent sur des techniques d'indexation [5], mais elles servent à décrire essentiellement des échantillons. Napster est alors en plein essor et l'idée que chaque mélomane se retrouve vite perdu dans une collection de milliers de morceaux justifie l'intensification des recherches en indexation musicale. **Aujourd'hui**, ISMIR est devenue une communauté bien établie attirant autant les chercheurs qu'une partie de l'industrie ; IEEE-ICASSP et ACM-Multimedia intègrent maintenant l'indexation musicale dans leurs thématiques ; les évaluations s'effectuent sur un million de morceaux [16] ; une partie des techniques d'indexation musicale est intégrée dans la vie de tous les jours (Shazam ou Midomi) ; mais les mélomanes n'ont pas des milliers de morceaux puisque la musique s'écoute aujourd'hui en « streaming » à travers des bases de données (Last-FM, Spotify ou Deezer voire YouTube) alimentées en métadonnées au travers de web services (EchoNest ou BMat).

Mes travaux s'inscrivent le long du développement de ce qui est appelé l'indexation automatique de contenus audio, avec pour une large partie des contenus audio représentant des morceaux de musique mais également des sons d'instruments ou des sons environnementaux. **Je désigne ici par « indexation audio », l'ensemble des recherches permettant de repérer des éléments significatifs dans des documents audio ou dans des collections de documents audio**. Mes travaux incluent donc des recherches relatives à la création de technologies permettant de repérer automatiquement ces éléments significatifs (voir parties 2 et 3) mais également des recherches visant à définir ces éléments, à créer des données annotées en ces éléments (voir partie 4.1) et des recherches relatives à l'évaluation des performances des technologies repérant ces éléments (voir partie 4.2). Finalement ces technologies sont utilisées dans des applications (nécessitant donc des optimisations), aux travers de scénarii utilisateurs, accessibles à travers des interfaces, le tout évalué lors de tests utilisateurs. Je décrirai ces derniers éléments au fil des présentations de mes recherches.

1. Les recherches effectuées pendant ma thèse, sous la direction de Xavier Rodet, portaient sur la modélisation des signaux audio monophoniques par décomposition en formes d'ondes élémentaires périodiques et sinusoïdes à variations lentes, dans l'objectif de permettre des modifications sonores de hautes qualités.
2. Issus des expériences perceptives de Stephen McAdams
3. La conférence ISMIR est crée en 2000 à Plymouth.
4. Computational Auditory Scene Analysis
5. Citons par exemple SoundFisher de MuscleFish [218] , FindSound de Comparisonics et Studio-OnLine de l'IRCAM [14].

Éléments significatifs. L'indexation audio vise à reconnaitre automatiquement des éléments significatifs dans un document audio. Ces éléments sont généralement significatifs dans un **contexte** donné. Le contexte désigne ici l'application et son utilisateur (grand public, musicien, compositeur, chercheur). Même si actuellement la majorité des applications de l'indexation audio concerne l'accès aux documents audio-numériques (moteurs de recherche), un intérêt croissant se développe pour son utilisation dans la « création ». C'est ce qu'a montré le récent workshop « MIR and Creation » [6]. Ces éléments sont également significatifs pour un **type de contenu** donné. Par exemple, les éléments significatifs d'un contenu de type flux radiophonique pourront être la connaissance de ses plages de parole, de musique et de publicité [169]. Pour les plages de parole, il pourra s'agir du locuteur, de la langue ou de la transcription de sa parole en texte [168] ; pour les plages de musique, il pourra s'agir de la connaissance des segments de couplets et refrains, de sa suite d'accords, de son tempo ou de son appartenance à une classe d'humeur ou genre musical ; pour un son isolé, il pourra s'agir de la description de son attaque ou de sa rugosité. Finalement les éléments significatifs se distinguent par leurs **portées intra ou inter-documents**. Un élément peut être :

- localisé en temps (accord, tempo, label d'un segment) voire être le temps lui-même (position d'un battement, début ou fin d'un segment),
- valide pour l'ensemble des temps d'un item, on dit qu'il a une portée « globale » en temps (tonalité, humeur ou genre musical),
- transversal aux items, il ne fournit pas de description par lui-même mais permet la comparaison entre items (similarité, identification audio).

Algorithmes d'estimation. Les algorithmes utilisés pour l'estimation des éléments significatifs dépendent du type d'élément, du type de contenu et des propriétés du signal audio. Pour un même élément (par exemple l'estimation de hauteur) on utilisera des algorithmes différents selon les propriétés du signal audio (signal monophonique ou polyphonique). De même, les modèles de connaissance sous-jacents aux algorithmes dépendent du langage sous-jacent à celui ayant généré le contenu : pour un contenu de type « musique » on utilisera des modèles de connaissance de « notes » ou « battements », pour un contenu de type « parole », on utilisera des modèles de connaissance de « grammaire » ou « syllabes ». Cette connaissance peut être introduite de deux manières :

- Elle peut être inférée automatiquement par un algorithme (dit « algorithme d'apprentissage ») sur la base d'exemples de données annotées (approche dite « **machine-learning** »). C'est souvent le cas lorsque la description à obtenir est difficile à définir mais peut être illustrée par des exemples. Etant donné que l'algorithme ne se concentre pas directement sur le problème mais sur l'inférence de règles permettant de résoudre les problèmes sur base d'exemples, le même algorithme d'apprentissage peut être utilisé pour permettre l'estimation de différents éléments significatifs.
- Elle peut être introduite manuellement par le créateur de l'algorithme d'estimation (approche que j'appelle « **human-learning** »). C'est souvent le cas lorsque la description à obtenir peut être définie. Dans ce cas, chaque description nécessite le développement d'un algorithme spécifique.

Notons que ces deux approches peuvent cohabiter au sein d'un même algorithme.
Finalement, un algorithme peut effectuer l'estimation d'un élément :

- directement à partir du signal audio (par exemple l'estimation de l'élément « genre musical » par modélisation statistique du comportement des MFCCs extraits du signal audio),
- par utilisation d'autres éléments préalablement estimés (par exemple l'estimation de l'élément « suite d'accords » par utilisation de l'estimation préalable des éléments « hauteurs de notes »),
- de manière conjointe avec celle d'autres éléments interdépendants (par exemple l'estimation conjointe des éléments interdépendants « premiers temps » et « changements d'accord »).

Organisation du document. Ce document est organisé comme suit. Les parties 2 et 3 décrivent mes recherches relatives à la création de technologies permettant de repérer automatiquement les éléments significatifs. La séparation en parties 2 et 3 distingue le fait que les éléments significatifs soient relatifs ou non à la notation musicale. La partie 4, quant à elle, décrit mes recherches concernant la création de corpora annotés en éléments significatifs, la définition de ces éléments et les campagnes d'évaluation des technologies d'indexation.

La rédaction de ce document repose sur l'utilisation d'acronymes généralement utilisés dans le domaine MIR. Afin de ne pas alourdir le corps de ce document, j'ai choisi de regrouper leurs significations dans un « Dictionnaire des acronymes » donné en préambule. Dans celui-ci, je fournis également un index par collaborations, par projets et par logiciels développés, de manière à permettre un autre parcours de ce document.

6. `http://recherche.ircam.fr/anasyn/peeters/pub/workshop_mir_creation/`. Des exemples d'applications utilisant les techniques d'indexation pour la « création » sont : transformation du signal par descripteurs [208], recomposition par superposition d'éléments [32] [48], par concaténation d'éléments – EchoNest Remix API ou système CataRT [203] –, transformation du swing [67], transformation de rythme – LoopMash –, suivi temporel [38] ou interaction ordinateur/musicien – système OMax [8] ou VirtualBand [125] –.

CHAPITRE 2

Indexation par descripteurs audio et apprentissage machine

Dans cette partie je résume mes travaux concernant l'estimation de paramètres non relatifs à la notation musicale. Ces travaux reposent pour la plupart sur l'extraction de descripteurs audio (génériques ou spécifiques) et sur des techniques d'apprentissage machine.

2.1 Description du timbre

La perception des sons a été étudiée de façon systématique depuis Helmholtz. Il est maintenant bien admis que les sons peuvent être décrits par leur hauteur, sonie, durée subjective, et ce qu'on appelle le « timbre ». Le « timbre » fait référence aux caractéristiques permettant de distinguer deux sons de même hauteur, sonie et durée subjective. Les mécanismes sous-jacents à sa perception sont assez complexes et impliquent la prise en compte simultanée de plusieurs dimensions. Le « timbre » est donc un attribut multidimensionnel incluant, parmi d'autres, l'enveloppe spectrale, temporelle, et les variations de chacune d'elle.

2.1.1 Quantification du timbre

État de l'art. Pour mieux comprendre cet aspect multi-dimensionnel, de nombreuses expériences perceptives ont été réalisées dont celles de Grey [69], McAdams [108] ou Lakatos [95]. Dans celles-ci il est demandé aux sujets de juger la similarité/dissimilarité entre paires de sons égalisés en hauteur, sonie et durée perceptive et de source méconnaissable. L'ensemble des jugements obtenus est ensuite analysé à l'aide de techniques de type Multi-Dimensional-Scaling (MDS) [108]. Cette analyse MDS vise à représenter les stimuli/sons dans un espace de dimension faible (généralement deux ou trois dimensions) tel que leur inter-distance dans ce nouvel espace représente au mieux les dissimilarités moyennes. L'espace résultant est généralement désigné par le terme « espace de timbre », ses axes par « dimensions perceptives » et les distances entre les sons dans cet espace par « distances perceptives ». L'interprétation **qualitative** vise à expliquer les dimensions perceptives par différentes représentations acoustiques. L'interprétation **quantitative** vise à prédire (quantifier) la position des stimuli/sons le long de ces dimensions à l'aide de différentes représentations acoustiques. Les premières interprétations quantitatives datent de 1978 [69]. En 1994, Krimphoff [90] propose un ensemble plus large d'attributs acoustiques (aujourd'hui appelés « descripteurs audio ») pour quantifier ces axes. Cette étude est poursuivie par Misdariis [114] dans le cadre du projet `Studio-OnLine`. Les descripteurs sont améliorés et un premier « modèle de prédiction de distances perceptives » (modèle mathématique permettant d'approximer les distances perceptives à l'aide de descripteurs audio) est proposé.

Contributions. Nos recherches s'inscrivent dans la continuité de celles de Misdariis et ont été effectuées dans le cadre du projet `Cuidad`. A travers ce projet, l'Europe souhaite permettre à des centres de recherche en musique de participer à l'établissement de la norme internationale ISO `MPEG-7`. Cette norme vise à établir un standard de description des documents multimédia, donc également de l'audio. De 1999 à 2002, en collaboration avec l'équipe Perception et Cognition Musicale de l'IRCAM et le Music Technology Group de Barcelone (en particulier Perfecto Herrera), nous effectuons plusieurs propositions de normalisation : une proposition d'organisation générale des descriptions multimédia (Multimedia Descriptor Scheme), une relative aux descripteurs audio de

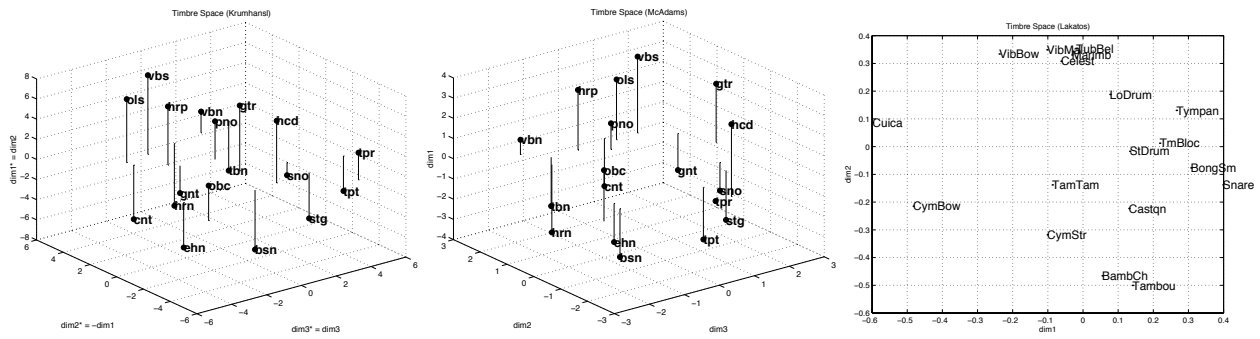

FIGURE 2.1 – Espaces de timbre de Krumhansl, McAdams et al. et Lakatos.

timbre, et une relative aux descripteurs audio dits « dynamiques » (voir partie 2.2.1.1). Les descripteurs audio de timbre font aujourd'hui partie intégrante de la norme MPEG-7 Audio.

Dans la suite nous notons d_n^k la valeur d'un descripteur audio k pour un son n, p_n^x la valeur de la projection sur l'axe x de l'espace de timbre d'un son n, $\hat{p}_n^x$ sa prédiction, $dist(n,n')$ la distance entre les sons n et n' dans l'espace de timbre et $\hat{dist}(n,n')$ sa prédiction.

Notre premier objectif est d'obtenir des descripteurs d^k permettant une meilleure prédiction acoustique des positions des sons dans les espaces de timbre tout en étant robuste à la présence de bruits et d'artefacts [1]. Pour cela, nous partons des descripteurs proposés par Krimphoff et Misdariis, y adjoignons ceux de Susini, Houix, Pressnitzer et Smith, créons de nouveaux descripteurs (comme l'`harmonic spectral spread`) ou améliorons l'estimation de descripteurs existants (comme le `log-attack-time`).

Notre deuxième objectif est d'élargir le nombre de catégories de sons considérées. Pour la norme MPEG-7, nous distinguons quatre catégories : sons harmoniques, inharmoniques, non cohérents et percussifs. Comme Misdariis, lorsque plusieurs espaces de timbre sont disponibles pour une même catégorie de sons, leurs résultats sont combinés afin d'améliorer la qualité de la prédiction. Pour la catégorie des sons harmoniques, nous utilisons donc, après ajustement (normalisation et rotation), les espaces de Krumhansl [91] et McAdams [108]. Pour la catégorie des sons percussifs, nous utilisons l'espace de Lakatos [95]. Ces trois espaces de timbre sont illustrés à la Figure 2.1.

Notre troisième objectif est d'améliorer les prédictions de position $\hat{p}_n^x$ et de distance $\hat{dist}(n,n')$. L'approche classiquement utilisée pour la prédiction $\hat{p}_n^x$ reposait sur la recherche du descripteur k tel que le vecteur $\mathbf{d}^k$ soit le plus corrélé avec $\mathbf{p}^x$. Le modèle de prédiction $\hat{p}_n^x$ est ensuite obtenu par régression linéaire sur le meilleur descripteur : $\hat{p}_n^x = f(d_n^k)$. Dans l'approche classique, le modèle de prédiction $\hat{dist}(n,n')$ est alors obtenu en supposant les dimensions $i = \{x,y,z\}$ orthogonales et en combinant les distances par axe : $\hat{dist}(n,n')^2 = \sum_{i=\{x,y,z\}}(\hat{p}_n^i - \hat{p}_{n'}^i)^2$. Mes deux contributions ont été d'une part l'utilisation de régressions multiples pour la prédiction de la position sur chaque axe : $\hat{p}_n^x = f(d_n^k, d_n^{k'})$; d'autre part la prédiction des distances $\hat{dist}(n,n')$ directement par modélisation des distances observées $dist(n,n')$ donc sans utilisation de la position des sons sur les axes perceptifs. Pour cela, nous estimons les meilleurs descripteurs ainsi que leurs poids associés par minimisation d'une erreur quadratique moyenne : $\epsilon = \sum_{n,n'}(dist(n,n') - \hat{dist}(n,n'))^2$ dans laquelle $\hat{dist}(n,n') = f(d_n^k, d_n^{k'}, d_n^{k''}, d_{n'}^k, d_{n'}^{k'}, d_{n'}^{k''})$.

Cette recherche conduit à la sélection de 7 descripteurs pour la caractérisation des deux espaces de timbre (harmonique et percussif) : `Log-Attack-Time, Temporal Centroid, Harmonic Spectral Centroid/Spread/Deviation/Variation, Spectral Centroid`. Nous avons intégré dans la norme ISO MPEG-7 Audio [118] ces descripteurs ainsi que les formules de distance $\hat{dist}$ correspondant aux espaces de timbre harmoniques et percussifs.

La procédure de normalisation ISO/MPEG nécessite la validation expérimentale des propositions (étape appelée « Core Experiment »). Pour cela nous avons validé le fait que l'utilisation combinée de tous les descripteurs de timbre et de la formule de distance fournit une meilleure similarité perceptive que celle obtenue à l'aide d'un seul de ces descripteurs ou qu'un son pris aléatoirement. Pour cela, le protocole expérimental suivant est utilisé. Pour un son « cible » n, nous présentons à l'utilisateur trois propositions de sons similaires : — m_1 : similarité obtenue à l'aide de tous les descripteurs et de la formule de distance, — m_2 : similarité obtenue à l'aide d'un seul des descripteurs, — m_3 : un tirage aléatoire. Nous montrons pour un ensemble de 27 sons cibles et sur un panel de 25 participants que la similarité obtenue en utilisant notre proposition (m_1) est préférée de manière statistiquement significative.

1. Les stimuli utilisés pour les expériences perceptives étant des sons de synthèse sans bruit ni artefact, nous cherchons à obtenir des descripteurs applicables sur des sons réels pouvant contenir du bruit et des artefacts.

Tous ces travaux sont décrits dans l'article [162]. Les descripteurs de timbre ainsi que les formulations de distances perceptives ont fait l'objet du brevet international [163]. Le travail sur la description du timbre se poursuit toujours et a donné lieu récemment à l'article de journal [160] réalisé en collaboration avec l'université Mc-Gill de Montréal.

2.1.2 Transformation des sons par descripteurs de timbre

Les descripteurs de « timbre » visent à représenter les propriétés perceptives les plus importantes d'un son n à l'aide d'un ensemble de paramètres scalaires d_n^k : $n \overset{extraction}{\longrightarrow} \{d_n^k\}$. L'objectif du stage de Master de Damien Tardieu [208], que j'ai encadré, est d'utiliser ces paramètres pour modifier un son. Ces paramètres étant perceptivement significatifs les modifications obtenues sont supposées l'être aussi. Cette recherche se différencie donc de recherches précédentes sur la synthèse par descripteurs de haut niveau (initialement proposée par Serra [204] et étudiée préalablement à l'IRCAM par Jean-Philippe Lambert) par l'utilisation des descripteurs de « timbre ». Le schéma suivant est utilisé : $n \overset{extraction}{\longrightarrow} \{d_n^k\} \overset{modification}{\longrightarrow} \{d_n^{'k}\} \overset{re-synthese}{\longrightarrow} n'$. Dans cette recherche nous étudions le sous-ensemble des descripteurs de timbre relatifs à l'enveloppe spectrale. Afin de limiter le problème d'indétermination (les descripteurs n'étant pas une bijection, il existe une multitude de sons n' possédant les mêmes valeurs de descripteurs $d_n^{'k}$), nous adjoignons au centroïde et à l'étendue spectrale les moments spectraux de troisième et quatrième ordre. Une couche intermédiaire est également introduite : l'enveloppe spectrale du signal est représentée par un modèle de type B-Spline adjoint d'un résiduel. Une formulation mathématique permet alors de relier les 4 moments spectraux aux paramètres des B-Splines et donc d'obtenir une bijection. Les résultats expérimentaux montrent le potentiel de cette approche qui reste malgré tout instable du fait de l'absence de contrainte sur la forme du spectre.

2.1.3 Orchestration automatique

Les descripteurs de timbre permettent de retrouver le son n' perceptivement le plus proche d'un son cible n selon le schéma : $n \overset{extraction}{\longrightarrow} \{d_n^k\} \overset{?}{\simeq} \{d_{n'}^k\} \overset{extraction}{\longleftarrow} n'$.

Suite à nos recherches sur le modèle de prédiction $\hat{dist}(n, n')$, des compositeurs de l'IRCAM nous ont posé la problématique suivante : retrouver le sous-ensemble de sons m_i tel que leur combinaison soit perceptivement la plus proche d'un son cible n_c : $n_c \simeq \sum_i n_i$. Cette problématique est relative à celle de l'orchestration dans laquelle un compositeur cherche la meilleure combinaison d'instruments de l'orchestre en vue d'obtenir un timbre spécifique. Cette recherche a fait l'objet de la thèse de Damien Tardieu [209], que j'ai co-encadré. En particulier, cette thèse étudie le problème de la reconstruction d'un échantillon synthétique n_c à l'aide d'instruments de l'orchestre $\{n_1, n_2 \cdots n_N\}$ devant être joués. n_i désigne donc la combinaison d'un instrument, d'une hauteur, d'un mode de jeu et d'une nuance. Une taxinomie très fouillée des instruments de musique et des modes de jeu est proposée. A l'inverse d'études précédentes [172] [192] dont le but est de reconstituer un son à partir d'échantillons, il s'agit ici de reproduire un son à partir d'instruments joués, donc avec une variabilité inhérente due à l'instrumentiste, à l'instrument et aux conditions de jeu (effet de salle). Cette variabilité est prise en compte par l'utilisation de modèles génératifs. Un son d'instrument est considéré comme le résultat de l'intersection d'un sous-modèle génératif d'instrument et d'un ou plusieurs sous-modèles génératifs de mode de jeu, de hauteur et de nuance. Ces modèles sont appris de manière indépendante à partir de l'analyse d'un large corpus d'apprentissage d'échantillons audio. L'utilisation de sous-modèles permet de pallier le problème du manque de données pour certains modes de jeu. Pour la représentation des sons, les descripteurs audio ayant des propriétés d'additivité [2] sont favorisés. Ceci permet de réduire drastiquement le coût de calcul. La recherche s'effectue alors par optimisation multicritères (parcours du front de Pareto).

Cette thèse effectuée en collaboration avec celle de Grégoire Carpentier [31] a donné lieu à de nombreux débouchés pour les compositeurs à l'IRCAM. Ce sujet a été poursuivi lors de la thèse de Philippe Esling [47] étudiant (entre autre) l'aspect temporel de l'orchestration.

2. La valeur du descripteur d'un son résultant de la somme de deux sons peut être obtenue à partir des descripteurs des sons individuels : $d_{n+n'}^k = d_n^k + d_{n'}^k$.

2.2 Descripteurs audio génériques

Les descripteurs de timbre présentés dans la partie 2.1 appartiennent à une famille plus générale appelée « descripteurs audio ». Nous désignons par « descripteur audio » une valeur quantitative ou qualitative décrivant une caractéristique acoustique particulière d'un signal audio. Ces descripteurs sont le résultat de l'application d'un opérateur mathématique sur le signal, sur une représentation dérivée du signal (spectre, modèle sinusoïdal, modèle perceptif ...), ou d'une cascade de ces opérateurs, voire encore le résultat d'un algorithme d'estimation (estimation de la fréquence fondamentale). Alors que les descripteurs de timbre sont généralement des scalaires (monodimensionnels) de sémantique directe (comme le « temps d'attaque »), les descripteurs audio peuvent prendre une forme multidimensionnelle (vecteur ou matrice) sans sémantique directe (comme les « MFCCs »).

En 2002, lors du commencement de cette étude, il m'est apparu que de nombreux descripteurs audio existaient dans la littérature mais restaient souvent isolés, chacun dans leur domaine (reconnaissance de la parole, reconnaissance des instruments de musique, études perceptives, identification audio). Ces descripteurs étaient/sont souvent des variations (changements d'échelle d'amplitude ou de fréquence, changements de représentation, variations de banc de filtres, prétraitements ...) autour d'un même concept. Il m'est apparu également que leurs différents auteurs se contredisaient quant aux justifications du choix d'une variation particulière. J'ai donc décidé de regrouper ces descripteurs, d'uniformiser et de formaliser leur description. Pour cela j'ai proposé avec Perfecto Herrera [77] une première organisation des descripteurs audio selon différents axes :

Le concept : Le concept désigne ce que le descripteur cherche à décrire. Ainsi les coefficients d'auto-corrélation, un cepstre, les MFCCs ou les moments spectraux sont différentes représentations s'attachant à décrire un même concept : l'enveloppe spectrale.

L'extraction : Il s'agit du mode de calcul du concept donnant le descripteur. Celui-ci regroupe autant l'implémentation générale du concept (auto-corrélation, cepstre, MFCCs ou moments spectraux) que les prétraitements pouvant être appliqués ou les représentations utilisées (TFD, séparation en partie sinusoïdale harmonique et bruits, banc de filtres de type Mel ou ERB) et les paramètres spécifiques au mode de calcul (nombre de filtres, nombre d'harmoniques).

Les hypothèses relatives au contenu : Certains descripteurs reposent sur l'hypothèse que le signal sur lequel ils sont appliqués possède certaines caractéristiques. Ainsi, le calcul du descripteur « rapport de l'énergie des harmoniques impaires sur paires » suppose la présence d'un contenu « mono source, monophonique et signal harmonique » ; à l'inverse le descripteur « MFCCs » ne repose sur aucune hypothèse relative au contenu du signal. On parlera dans la suite de descripteurs **« spécifiques »** et de descripteurs **« génériques »**.

La validité temporelle : Elle désigne l'étendue temporelle décrite par le descripteur. Nous distinguons différents niveaux. Les descripteurs dits **« instantanés »** (comme les MFCCs) décrivent le signal sur une étendue courte (correspondant généralement à la trame d'une TFCT). A l'inverse, les descripteurs dits **« globaux »** décrivent la totalité d'un fichier : — soit que conceptuellement il n'y a qu'une valeur pour l'ensemble du morceau (la durée d'un morceau, le nom de l'artiste, le temps d'attaque d'une note d'un échantillon), — soit qu'une valeur globale puisse être déduite de la succession de valeurs instantanées (tempo moyen). Entre ces deux niveaux, se trouve le niveau **« segmental »** correspondant aux intégrations temporelles (voir partie 2.2.1).

Cependant, même si l'uniformisation du calcul des descripteurs est possible, chaque application peut nécessiter des spécificités de calcul. Ainsi pour une tâche de reconnaissance, on souhaitera des descripteurs insensibles au volume global du son (insensibilité au niveau d'enregistrement) et au taux d'échantillonnage (la même classe doit être reconnue quel que soit le volume ou le taux d'échantillonnage). La formalisation des descripteurs devra donc permettre cette insensibilité. A l'inverse, pour une description perceptive, on cherchera une formulation rendant compte de cette sensibilité.

Le sous-ensemble des descripteurs de « timbre » est détaillé dans l'article de journal [160] joint en annexe de ce document. Cet article propose également une étude de la redondance d'information apportée par ces descripteurs. Leur schéma d'extraction est donné à la Figure 2.2. L'ensemble global des descripteurs est décrit dans le document en ligne [141]. Ces descripteurs ont fait l'objet de plusieurs développements à l'IRCAM, dont le logiciel `ircamdescriptor`, développé par Patrice Tisserand, Carmine Emanuele Cella, Alessandro Saccoia et maintenant Frédéric Cornu. Ce logiciel est aujourd'hui la base de nombreuses applications à l'IRCAM.

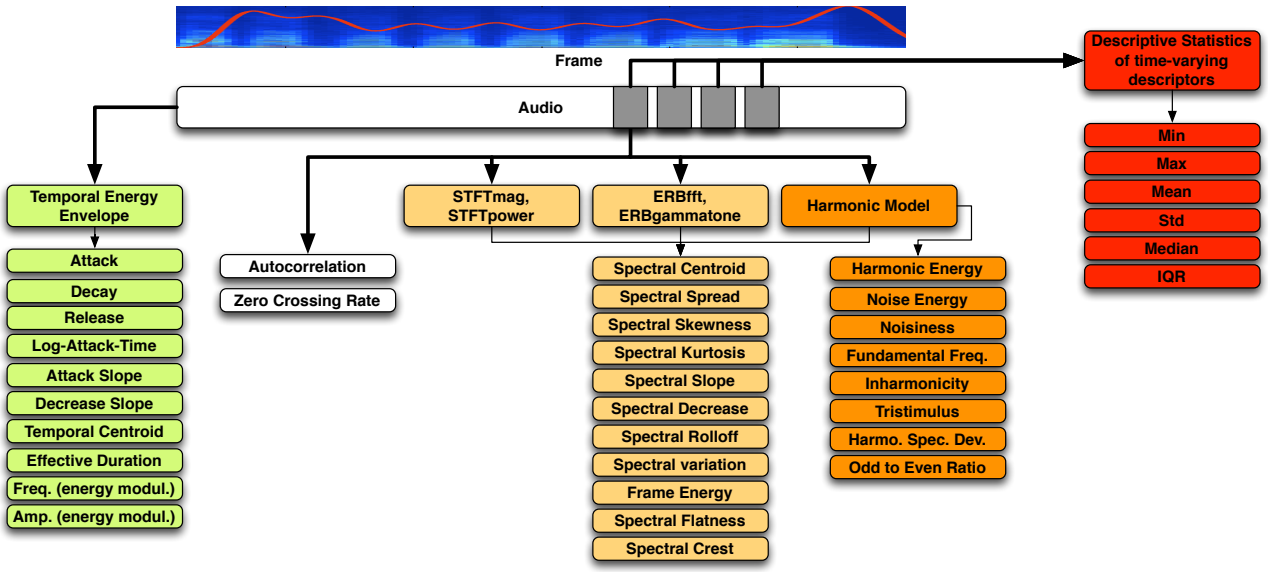

FIGURE 2.2 – Schéma d'extraction du sous-ensemble des descripteurs audio dits de « timbre ».

2.2.1 Intégration temporelle

L'intégration temporelle s'applique aux descripteurs « instantanés » (i.e. calculés à chaque instant du signal). Elle vise principalement deux objectifs : — permettre une réduction du nombre de données (un morceau de musique de 4 min conduit à 12.000 vecteurs de descripteurs pour une extraction toutes les 20 ms), — mettre en évidence des propriétés relatives à l'ordonnancement temporel des données à travers un ensemble limité de paramètres. L'intégration temporelle de descripteurs audio est souvent désignée sous le terme de « texture window ». Nous renvoyons le lecteur à l'article de Joder [81] pour un bon aperçu de l'état de l'art dans ce domaine. Au cours des années, nous avons utilisé ou proposé différentes techniques d'intégration temporelle. Nous les présentons ici en les divisant en deux catégories :

- celles ne reposant pas sur un modèle appris mais sur un opérateur (une suite d'opérateurs) mathématique(s) appliqué(s) au flux de données instantanées (par exemple l'opérateur moyenne),
- celles reposant sur un modèle appris (par exemple un dictionnaire de mots).

2.2.1.1 Intégration temporelle sans apprentissage de modèles

Nous notons $\mathbf{d}_m$ le vecteur de descripteurs audio à la trame m et d_m^k sa $k^{\text{ième}}$ dimension.

Opérateurs simples. Dans ce cas, le comportement temporel du descripteur sur un horizon L_m (pouvant être le signal entier ou une fenêtre glissante) est résumé par des opérateurs mathématiques simples de type moyennes, écart-types, dérivées premières et secondes. Nos contributions principales ont été les propositions d'utilisation des valeurs moyennes, écart-types et dérivées pondérées par la sonie temporelle [141] ainsi que des valeurs médianes et écarts interquartiles afin de permettre une meilleure résistance à la présence de données périphériques (« outliers ») [160].

Modèle Auto-Régressif Multivarié (ARM). L'application des modèles auto-regressifs multivariés a été proposée par [19] afin de modéliser le comportement temporel des descripteurs. Comme dans le cas d'un signal monodimensionnel, l'objectif est de représenter la dépendance des valeurs au cours du temps par un filtre tout-pôles d'ordre P : $d_m^k = \sum_{p=1}^{P} \alpha_p^k d_{m-p}^k + \epsilon_m^k$ dans lequel α_p^k désigne les coefficients du filtre et ϵ_k le résiduel de modélisation. Dans le cas d'un modèle « multivarié », nous considérons également la dépendance entre les différentes dimensions k. Le résultat est une matrice de coefficients $\mathbf{A} = \{\alpha_p^{k,k'}\}$. Notre contribution principale a résidé dans l'utilisation du résiduel de modélisation AR multivarié [210].

Descripteurs dynamiques, spectre de modulation. En 1998, Worms et Rodet [219, 189] proposent de modéliser les évolutions temporelles de l'énergie dans différentes bandes de fréquences par leur spectre d'amplitude. Le résultat est utilisé comme signature d'un signal audio pour une tâche d'identification audio (voir partie 2.6). Notre contribution a été d'étendre cette modélisation à tout type de descripteurs tels les MFCCs ou Chromas [140]. Je donne le formalisme général de cette modélisation ci-dessous. Nous notons $w(m)$ une fenêtre

de pondération de taille L_m et R_m un pas d'avancement, la TFCT de d_m^k s'exprime (convention passe-bande) :

$$Y(k,k',o) = \sum_{m=1}^{L_m} d^k_{m+oR_m}\, w_{L_m}(-m) \exp\left(-j2\pi \frac{k'}{M} m\right) \tag{2.1}$$

Dans le cas où d_m^k représente le carré du module de la TFCT, $d_m^k = |X(k,m)|^2$, nous retrouvons la méthode proposée par [219]. Son calcul est illustré à la Figure 2.3. Cette méthode est cependant plus connue aujourd'hui sous le nom « modulation spectrum » [13] [9]. Elle peut également être rapprochée des méthodes « Auditory filterbank temporal envelopes » de [111], des « Penny Cepstral Features » de [217] ou encore du « multiscale scattering » de [6]. Comme nous le verrons dans la suite, cette représentation s'est montrée particulièrement efficace pour les tâches d'identification audio (partie 2.6) et d'estimation de structures musicales (partie 3.3.2.1).

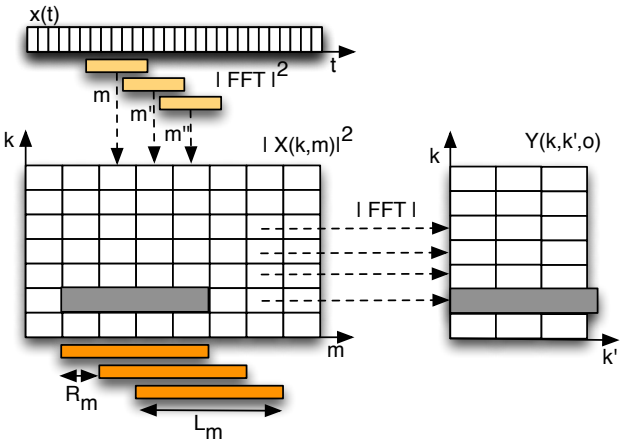

FIGURE 2.3 – Calcul des descripteurs dynamiques.

2.2.1.2 Intégration temporelle avec apprentissage de modèles

L'intégration temporelle de descripteurs instantanés avec apprentissage de modèles se divise elle-même en deux catégories selon que l'apprentissage s'effectue sur la suite de données elle-même ou sur un corpus d'entraînement externe.

Méthodes par dictionnaire. Dans un premier temps un dictionnaire de C mots $\{c_1 \dots c_i \dots c_C\}$ est créé à partir de l'observation d'un ensemble de descripteurs (ceux-ci peuvent provenir de l'ensemble des descripteurs du flux à traiter ou d'un corpus d'apprentissage externe à ce flux). Ce dictionnaire peut être appris par des techniques de regroupement (« clustering ») classiques de type K-moyenne (« K-means ») ou par une modélisation Markovienne cachée [3]. Une donnée $\mathbf{d}_m$ peut ensuite être représentée par le mot c_i le plus proche (dans ce cas la représentation est scalaire) ou par le vecteur de dimension C représentant la probabilité d'appartenance à chacun des mot c_i. Dans [145] je propose une variante de cette technique dans laquelle les mots du dictionnaire c_i correspondent à des modèles de classes (classes de genre musical) apprises de manière supervisée. Une extension de ces méthodes de dictionnaire consiste à prendre en compte l'ordonnancement temporel des descripteurs. Dans cette extension, les descripteurs successifs $\mathbf{d}_m, \mathbf{d}_{m+1} \dots$ sont représentés par leurs mots $c_i(m), c_i(m+1) \dots$; une matrice représente ensuite les transitions temporelles entre ces mots.

Multi-probe histogram. Bien que ne reposant pas sur un modèle, nous indiquons également le « multi-probe histogram », récemment introduit par [223] et [83], puisqu'il constitue une autre forme de modélisation « quantifiée » de cet enchaînement temporel. Dans celui-ci, chaque vecteur $\mathbf{d}_m$ est représenté par son vecteur de rang. Une matrice représente ensuite les transitions temporelles entre rangs proches (les rangs élevés reçoivent un score élevé).

Dans ces différentes méthodes, les décisions sont cumulées sur un horizon L_m.

Modèle du monde, « Universal Background Model ». A l'inverse des techniques précédentes, celle dite du « Universal Background Model (UBM) » [185] [29] nécessite un préapprentissage sur un corpus externe au flux. L'objectif de cette technique est de représenter le monde des descripteurs par un modèle de mélange Gaussien et d'ensuite déformer ce monde afin de représenter les données observées sur l'horizon L_m [4]. La représentation résultante est la concaténation des C vecteurs moyennes μ_{c_i} du modèle de mélange gaussien. La taille de ce vecteur dépend donc à la fois de la dimension K des descripteurs et du nombre C de composantes du modèle de mélange gaussien. Le vecteur concaténé résultant est appelé « super-vecteur ».

3. Par exemple [1] utilise les HMMs pour obtenir le dictionnaire représentatif d'un morceau dans le cadre de l'estimation de structures musicales.

4. Cette déformation est obtenue en adaptant les vecteurs moyennes μ_{c_i} $c \in C$ du GMM par un algorithme de type Expectation/Maximization.

2.2.1.3 Choix des durées et pas d'avancement pour l'intégration temporelle

Quelle que soit la technique choisie, l'intégration temporelle se fait sur une durée L_m et autour d'instants m. L'approche habituelle consiste à effectuer cette intégration par une méthode dite « de fenêtre glissante », i.e. une fenêtre de taille L_m est « glissée » le long de d_m^k avec un pas d'avancement fixe : $m = oR_m,\ o \in \mathbb{N}$ dans lequel R_m est le pas d'avancement. Par exemple, dans le cas de l'identification audio, nous utilisons L_m et R_m correspondant à une durée de 2 s et un pas d'avancement de 0.5 s. Le choix de ces paramètres n'est cependant pas anodin. Même s'il permet une réduction du nombre de données, l'intégration temporelle peut provoquer un « lissage » des données. Dans le cas de la musique, une approche utilisée est de rendre L_m et O_m dépendant du tempo et de la position des battements (intégration dite « beat-synchrone »). Dans le cadre de la thèse d'Hélène Papadopoulos sur l'estimation de la suite d'accords, nous avons effectué une étude fouillée [130] de ces intégrations « beat-synchrone ». Dans le cadre de l'estimation de structure, nous avons proposé avec Florian Kaiser [84] une approche permettant d'adapter automatiquement la taille et la position de ces fenêtres afin de limiter ce « lissage ».

2.2.2 Sélection automatique de descripteurs

L'objectif d'un Algorithme de Sélection Automatique de Descripteurs (ASAD) est de déterminer, parmi un ensemble de descripteurs, le sous-ensemble minimal (non redondant) le plus informatif pour résoudre un problème de prédiction. Pour la prédiction de données *quantitatives*, nous avons étudié différentes stratégies de sélection de descripteurs dans nos recherches sur le timbre (voir partie 2.1). Je présente ici mes travaux concernant les ASADs dans le cas de la prédiction de données *qualitatives* (problèmes de classification). Dès le début de mes recherches sur la classification automatique, mon approche a consisté à extraire un grand nombre de descripteurs et à laisser un ASAD sélectionner les plus pertinents. De ce fait, j'ai étudié les ASADs dès le début. Même si certains algorithmes de classification sont relativement peu sensibles à la présence de descripteurs non informatifs (c'est le cas des SVM), d'autres le sont beaucoup plus (c'est le cas des KNN). Dans tous les cas, l'utilisation d'un ASAD permet de réduire la dimensionnalité des données, donc de réduire la malédiction de la dimension, de réduire le coût de calcul et de stockage. Mon objectif n'était pas de développer un nouvel ASAD mais de pouvoir utiliser un ASAD efficace dans nos systèmes de classification. Après de multiples essais infructueux avec des ASADs existants (comme ReliefF [89] ou CFS [73]) j'ai néanmoins décidé d'en développer un nouveau.

Selon Molina [117], les ASADs peuvent être rangés en trois catégories : les algorithmes de type *Filter* où l'ASAD est distinct de l'algorithme de classification et est utilisé en amont de celui-ci, de type *Embedded* où l'ASAD fait partie intégrante de l'algorithme de classification et de type *Wrapped* où l'ASAD utilise le résultat de l'algorithme de classification pour effectuer la sélection. Mes propositions d'ASADs font partie de la catégorie « Filter » et visent à satisfaire les deux critères suivants :

Critère A : Choix d'un sous-ensemble de descripteurs informatifs vis-à-vis des classes,

Critère B : Choix de descripteurs non redondants.

J'ai effectué trois propositions d'ASADs. Le premier, décrit dans [165], utilise les poids de la matrice de transformation des descripteurs de l'Analyse Linéaire Discriminante (LDA). Chaque axe p de la LDA visant à maximiser la séparation entre les classes, le poids α_p^k (du descripteur k sur l'axe p) est considéré comme représentatif de l'importance du descripteur pour séparer les classes. Cependant du fait que les α_p^k ne sont pas nécessairement positifs, leur utilisation s'est avérée problématique. Le second, également décrit dans [165], utilise l'information mutuelle conditionnelle entre les classes et les descripteurs. Cependant le calcul de l'information mutuelle conditionnelle n'est pas possible d'un point de vue analytique et nécessite une approximation problématique. Je présente ici mon troisième algorithme, appelé **Inertia Ratio Maximization with Feature Space Projection (IRMFSP)** décrit dans [138]. Cet algorithme sélectionne les descripteurs de manière à satisfaire itérativement les critères A et B.

Critère A (Inertia Ratio Maximization). Le critère A cherche les descripteurs les plus informatifs. L'information apportée par un descripteur est mesurée par la discrimination entre classes fournie par la connaissance d'un descripteur. Celle-ci est mesurée au travers du rapport r^k de l'inertie interclasse sur l'inertie totale selon le descripteur $\mathbf{d}^k$ (également appelé discriminant de Fisher). A chaque itération, nous cherchons le descripteur maximisant ce rapport.

Critère B (Feature Space Projection). Le critère B vise à empêcher la redondance des descripteurs sélectionnés. Alors que d'autres ASADs, comme le CFS de [73] utilisent une pénalité proportionnelle à la corrélation entre descripteurs candidats et descripteurs déjà sélectionnés, dans l'algorithme IRMFSP, nous appliquons une orthogonalisation de Gram-Schmidt de manière à rendre les descripteurs restants orthogonaux aux descripteurs sélectionnés.

Ce processus (maximisation du rapport et projection) est répété jusqu'à ce que l'adjonction d'un nouveau descripteur n'apporte qu'un gain faible. Ce gain est mesuré par comparaison du rapport obtenu à l'itération courante à celui obtenu à l'initialisation.

A la Figure 2.4, nous illustrons les résultats obtenus par l'algorithme IRMFSP pour la sélection des descripteurs d'un problème à deux classes : séparation entre instruments tenus et non-tenus. Les trois axes représentent les trois premiers descripteurs sélectionnés : le taux de décroissance temporelle (1ère dimension), le centroïde spectral (2ème), le taux de croissance temporel (3ème). Dans [138], nous comparons les résultats obtenus par l'algorithme IRMFSP à ceux obtenus par l'algorithme CFS [73]. Les deux sont utilisés en amont du même algorithme de classification (un modèle Gaussien multidimensionnel). Pour un problème de classification en 27 instruments de musique, l'utilisation de CFS conduit à un taux de reconnaissance moyen de 60.9%, IRMFSP à 95.1%. De par sa simplicité et son coût de calcul très faible, l'algorithme IRMFSP a fait l'objet de plusieurs utilisations à l'extérieur de l'IRCAM. Il a également fait l'objet de modifications dont celles de Essid [49]. Nos modifications récentes (non encore publiées) visent à réduire sa dépendance envers la distribution des classes. Cet algorithme possède cependant des limites comme le résume Ramona dans [173] : « ... la phase d'orthogonalisation introduite nécessite un certain nombre d'exemples pour être statistiquement fiable. De plus cette fiabilité décroît fortement à mesure que les effets d'orthogonalisations successives se cumulent. »

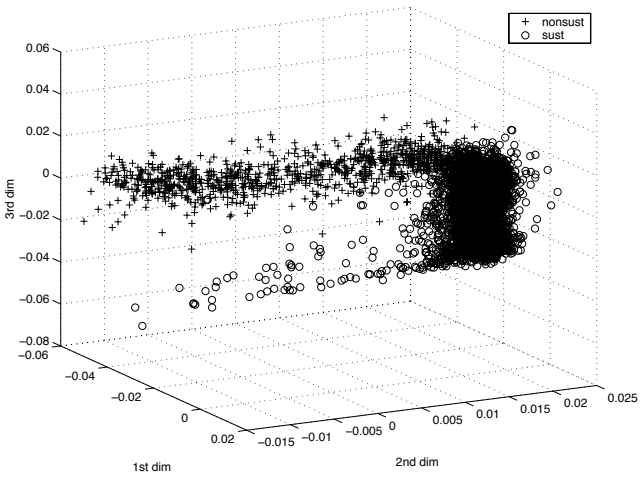

FIGURE 2.4 – Trois premiers descripteurs sélectionnés par l'algorithme IRMFSP pour la séparation entre instruments tenus et non-tenus.

2.3 Classification et segmentation automatique

Nous désignons sous le terme « classification » l'action consistant à attribuer une (ou plusieurs) étiquettes c_i à une donnée ; étiquette choisie parmi un ensemble fini et connu d'étiquettes $\{c_1 \dots c_i \dots c_C\}$. Dans notre cas, cette donnée peut être un échantillon audio, un extrait ou la totalité d'un morceau de musique. Lorsque les classes c sont mutuellement exclusives, on parle de classification « *single-label* » (une donnée appartient à une classe et à une seule classe, généralement la plus vraisemblable) ; lorsqu'elles ne le sont pas on parle de classification « *multi-label* » (une donnée peut appartenir simultanément à plusieurs classes, voire à aucune classe, il s'agit donc également d'un problème de détection). La « segmentation » désigne généralement le fait d'identifier dans un flux des instants de changement, ou des instants entre lesquels certaines propriétés sont homogènes. Ces propriétés peuvent correspondre à une classe (classe de « parole » ou de « musique »). Dans ce cas, la segmentation comporte également une phase d'étiquetage des segments en classe. La segmentation peut donc s'opérer par identification de changements dans le flux temporel (à l'aide de critères d'information de type BIC ou autres) suivi d'un étiquetage en classes des segments, ou, à l'inverse, en étiquetant en classes les différents instants d'un flux temporel et en regroupant les instants adjacents de classes identiques en segments. Notre choix s'est porté sur cette seconde approche et il n'y a donc pas dans notre cas de différence fondamentale (hormis des post-traitements) entre classification et segmentation.

Un **système de classification** est généralement constitué de deux sous-systèmes :

- un sous-système permettant l'apprentissage (supervisé) de la relation entre données n (observées à travers leurs descripteurs audio d_n^k) et étiquettes c_n (phase d'entraînement du modèle M_c à partir d'exemples) : $\forall n \ (c_n, d_n^k) \overset{apprentissage}{\longrightarrow} M_c$
- un sous-système permettant l'attribution d'une étiquette à une donnée à partir de ses descripteurs et d'un modèle entraîné (phase d'évaluation) : $\forall c \ (M_c, d_n^k) \overset{evaluation}{\longrightarrow} c_n$

Un « système » de classification regroupe l'ensemble des éléments nécessaires à ces deux tâches. Dès le départ la stratégie choisie pour ce système est l'extraction d'un grand nombre de descripteurs audio (voir partie 2.2). Les plus informatifs sont « filtrés » par utilisation d'un algorithme de « sélection automatique de descripteurs » (partie 2.2.2). Les descripteurs ainsi sélectionnés peuvent ensuite être transformés afin de vérifier les propriétés de l'algorithme de classification (utilisation d'une transformée de type Box-Cox [23] pour « gaussianiser » la distribution des descripteurs si l'algorithme de classification est de type Gaussien) ou de maximiser la séparation

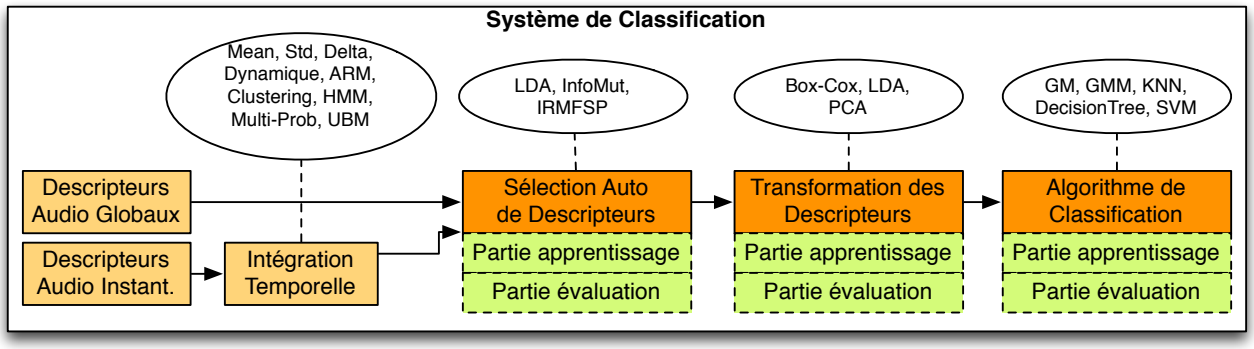

FIGURE 2.5 – Schéma des interactions des éléments constitutifs d'un système de classification type.

des classes (Analyse Linéaire Discriminante). Le résultat de ces transformations sert finalement d'entrée à un « algorithme de classification » comportant une phase d'apprentissage et d'évaluation. La Figure 2.5 représente le schéma type de nos systèmes.

Algorithmes de classification. Au cours des années, nous avons proposé plusieurs systèmes de classification, mais pas de nouvel algorithme. Pour ce dernier nous nous sommes reposés sur l'état de l'art. Au cours de nos recherches (voir [138] [26] [156] [151]), nous avons comparé différents types d'algorithmes de classification : – génératifs (modèle Gaussien, GMM, HMM), – discriminants (arbre de décision binaire, SVM), – « par exemples » (KNN), – « ensemble d'algorithmes » (Adaptive Boosting, RandomForest). Même s'il est tentant de tirer de ces comparaisons une conclusion quant à un meilleur algorithme, notre expérience nous a montré que chaque problème peut conduire à un choix différent, et que le choix des paramètres d'un algorithme (tel le choix du noyau pour les SVM et de ses paramètres) est également déterminant. Malgré ces considérations, l'algorithme SVM amène souvent à de très bons résultats et est dès lors devenu notre algorithme de prédilection. Nous avons également comparé différentes topologies d'organisation des classes : « à plat » (toutes les classes sont considérées au même niveau) ou hiérarchique (les classes sont regroupées manuellement en méta-classes [5]). Lorsqu'une hiérarchie peut être trouvée, le gain obtenu peut être important. Nous montrons cela dans [138] dans le cas de la reconnaissance de 27 instruments par algorithme de classification Gaussien. Le passage d'une topologie « à plat » à une hiérarchique accroît la reconnaissance de 54% à 64%. Notons que lorsque cette hiérarchie ne peut être trouvée manuellement, elle peut l'être automatiquement comme l'a proposé Essid [49].

Evaluation des performances. Dès nos premiers travaux, nous avons été confronté au problème de l'évaluation des performances et du sur-apprentissage. Étant donnée la difficulté à obtenir des données, nos premiers résultats sur la reconnaissance d'instruments [165] ont été effectués par validation à plis croisés (« N-fold cross validation ») sur un seul corpus, `Studio-OnLine`. Ces résultats très encourageants n'ont cependant pas pu être étendus à d'autres corpora de test du fait d'un sur-apprentissage [6]. Ceci remet quelque peu en cause les conclusions de notre article [165]. Pour cette raison, nous avons rapidement adopté les validations croisées sur des bases indépendantes d'entraînement et de test. Celles-ci sont désignées par « Minus-1 DB » dans [101] ou par « Leave-One-Database-Out » dans [138]. Nous avons également proposé l'utilisation du Rappel moyen (« mean-Recall ») (moyenne à travers les classes) pour mesurer les performances d'un système de classification puisque celui-ci, à l'inverse de la F-Measure, est indépendant de la distribution des classes.

2.3.1 Systèmes génériques de classification

A partir de 2006, de nouveaux projets ont nécessité de nouvelles recherches sur la classification. Le développement de systèmes dédiés à chaque problème de classification étant une tâche lourde, j'ai décidé de créer un système de classification générique [145]. Celui-ci doit pouvoir être appliqué à tout problème formalisable sous forme de classe (classification single-label, multi-label, segmentation). La conception du système est basée sur la séparation entre la partie **applicative** (l'algorithme qui sera réellement utilisé pour classer ou segmenter une donnée et qui sera distribué) et la partie **recherche** (dans laquelle le chercheur définit les paramètres permettant d'optimiser l'apprentissage des classes et d'évaluer les performances du système en cours de développement). Ceci a donné lieu à trois générations de systèmes de classification génériques.

Notons que l'utilisation de systèmes génériques est devenue courante dans la communauté MIR : système EDS de Sony CSL [126], jAudio/ACE de l'Université de McGill [109] [110], Marsyas [212], M2K de l'IMIR-

5. Dans [138], les classes d'instruments de musique sont ainsi regroupées manuellement en cordes frottées, pincées, vents ...
6. L'effet de différence de niveaux d'enregistrement de certains instruments de `Studio-OnLine`, combiné à un sur-apprentissage dû à une utilisation aveugle de l'algorithme de LDA, conduit à des conclusions erronées dans l'article.

SEL [78], Yaafe de Télécom ParisTech ou encore Essentia du MTG de Barcelone. Ces systèmes diffèrent par les descripteurs audio utilisés, leur algorithme de classification, leur langage de programmation, leurs optimisations. Leurs performances respectives sont cependant difficiles à évaluer puisqu'ils sont généralement appliqués à des tâches différentes. Le système de l'IRCAM se nomme `ircamclassification`.

ircamclassification1 [145]. Il s'agit du système de base reposant sur des descripteurs génériques, une sélection de paramètres par l'algorithme IRMFSP et le choix d'algorithmes de classification de type Gaussien, GMM, KNN ou HMM. Ce système permet la classification single-label et la segmentation.

ircamclassification2 [26]. Le système précédent est ensuite étendu par Juan José Burred aux problèmes multi-labels : une donnée peut potentiellement appartenir à plusieurs classes, voire à aucune. De ce fait les classes ne sont plus mutuellement exclusives et nous ne pouvons plus utiliser l'apprentissage discriminant de manière directe. Nous l'utilisons au travers d'une approche de type un-contre-tous (« one-versus-all »). Dans cette approche, un problème à C classes est ramené à C problèmes à 2 classes. Ces dernières sont les classes c_i et leurs opposés. Nous créons donc C systèmes de classification (incluant chacun un algorithme de sélection de descripteurs et un algorithme de classification). L'algorithme de classification utilisé est de type SVM. Dans [26], nous proposons un nouveau critère d'optimisation pour la détermination des paramètres des noyaux des SVMs permettant de réduire l'effet des déséquilibres entre classes. En fonction du nombre de fausses réjections et faux positifs demandés par une application, nous entraînons finalement les seuils τ_i à appliquer à l'affinité de chacun des SVMs pour décider de l'appartenance d'un item à une classe c_i.

ircamclassification3 [211]. La dernière génération, étudiée par Damien Tardieu et Christophe Charbuillet, se base sur des techniques utilisées en reconnaissance du locuteur. L'étage de modélisation temporelle des descripteurs est remplacé par deux techniques issues de la reconnaissance du locuteur : l'adaptation d'Universal Background Model et les modèles AR multivariés [36] [211] (voir partie 2.2.1). Le système générique a également été transformé en un système entièrement modulaire (système de plug-ins) par l'utilisation d'abstractions. Des abstractions correspondant aux concepts de descripteurs, d'algorithmes de sélections de descripteurs, de classification, de fusions de décision ou d'évaluateurs sont ainsi créées. Dans ce système, l'ajout d'une instance particulière d'une abstraction (comme un nouvel algorithme de classification) est très facile.

Ces systèmes génériques sont programmés en `matlab`. L'extraction des descripteurs utilise la librairie C++ `ircamdescriptor`. Dans le cas d'une intégration auprès d'un industriel, une fois les paramètres optimaux trouvés pour résoudre le problème particulier (comme par exemple la prédiction des tags de genres musicaux spécifiques à l'application MSSE d'Orange), le système génère le sous-ensemble de code C++ nécessaire pour cette application (incluant le sous-ensemble de descripteurs, les paramètres de transformation de ceux-ci, ceux des modèles SVM ...). Ce code sert ensuite à la compilation d'une application optimisée dédiée.

Dans le projet `3DTVs`, nous étendons actuellement avec Laurent Benaroya le système `ircamclassification` au traitement de l'audio multi-canal dans l'objectif de tirer profit de l'information de localisation spatiale pour la classification.

2.3.2 Performances

Afin de permettre la comparaison de nos systèmes de classification génériques à ceux de l'état de l'art, nous utilisons l'ensemble des résultats obtenus lors des évaluations `MIREX` de 2008 à 2012 pour les 4 tâches single-label et les 2 tâches multi-labels. Bien que les protocoles ainsi que les corpora de tests `MIREX` peuvent être sujets à discussion, ces résultats indiquent cependant une tendance. Nous indiquons ces résultats à la Table 2.1 [7]. Chaque cellule indique le résultat obtenu par un des systèmes IRCAM (partie supérieure) et le meilleur résultat obtenu parmi l'ensemble des participants (partie inférieure). Un de nos résultats souligné indique une première place. Les auteurs d'un algorithme sont donnés en indice des scores : « 49.8_{TCCP} » indique que le score 49.8 a été obtenu par les auteurs [Tardieu, Charbuillet, Cornu, Peeters]. Pour le lecteur de la version pdf de ce document, chaque cellule est un hyper-lien vers les résultats détaillés. Ces résultats montrent de très bonnes performances de nos systèmes pour toutes les tâches hormis pour la tâche « Classical Composer ». Ils montrent également une constante augmentation de nos résultats au cours des années. En particulier les résultats que nous avons obtenus pour la tâche « Multi-Label : Tag MajorMiner » ne sont toujours pas surpassés même par l'approche reposant sur les « Deep Believe Networks » (PH).

7. Il est à noter que les résultats obtenus par « PH » en 2011 ont été retirés de cette table étant donné le non-respect du protocole d'évaluation. Pour cette même raison, nous n'avons pas participé à ces tâches `MIREX` en 2012.

Tâche MIREX	Mesure d'Eval.	2008	2009	2010	2011	2012
Single-Label : Mood	Mean Accuracy	$\underline{63.7}_{\text{GP}}$ $\rightarrow 63.7_{\text{GP}}$	63.7_{GP} $\rightarrow 65.7_{\text{CL}}$	63.2_{GP} $\rightarrow 64.1_{\text{WLJW}}$	67.2_{TCCP} $\rightarrow 69.5_{\text{JR}}$	$\rightarrow 67.8_{\text{PP}}$
Single-Label : Genre Mixed	Mean Accuracy	63.9_{GP} $\rightarrow 66.4_{\text{GT}}$	70.6_{BP} $\rightarrow 73.3_{\text{CL}}$	70.7_{BRPC} $\rightarrow 73.6_{\text{SSPK}}$	75.3_{TCCP} $\rightarrow 75.6_{\text{WR}}$	$\rightarrow 76.1_{\text{WJ}}$
Single-Label : Genre Latin	Mean Accuracy	-	67.3_{BP} $\rightarrow 74.7_{\text{CL}}$	70.7_{BRPC} $\rightarrow 79.9_{\text{SSPK}}$	74.9_{TCCP} $\rightarrow 75.8_{\text{SSPK}}$	$\rightarrow 77.0_{\text{RW}}$
Single-Label : Classical Comp.	Mean Accuracy	49.0_{GP} $\rightarrow 53.3_{\text{ME}}$	55.7_{BP} $\rightarrow 61.0_{\text{CL}}$	55.2_{BRPC} $\rightarrow 65.3_{\text{WLB}}$	57.9_{TCCP} $\rightarrow 68.8_{\text{JR}}$	$\rightarrow 69.7_{\text{LBLJK}}$
Multi-Label : Tag Mood	Av. Tag F-Meas	-	19.5_{BP} $\rightarrow 21.9_{\text{LWW}}$	$\underline{46.6}_{\text{BRPC}}$ $\rightarrow 46.6_{\text{BRPC}}$	47.8_{TCCP} $\rightarrow 49.1_{\text{SSKS}}$	$\rightarrow 49.1_{\text{SSKSS}}$
Multi-Label : Tag MajorMiner	Av. Tag F-Meas	-	29.0_{BP} $\rightarrow 31.1_{\text{LWW}}$	$\underline{47.8}_{\text{BRPC}}$ $\rightarrow 47.8_{\text{BRPC}}$	$\underline{49.8}_{\text{TCCP}}$ $\rightarrow 49.8_{\text{TCCP}}$	$\rightarrow 49.5_{\text{PH}}$

TABLE 2.1 – Résultats obtenus par les algorithmes `ircamclassification1`, 2 et 3 pour les tâches single et multi-label `MIREX`.

GP	Peeters	(ircamclassification1)
BP	Burred, Peeters	(ircamclassification2)
BRPC	Burred, Ramona, Peeters, Cornu	(ircamclassification2 revised)
TCCP	Tardieu, Charbuillet, Cornu, Peeters	(ircamclassification3)
CL	Coa, Li	(Chinese Academy of Sciences)
WLJW	Wang, Lo, Jeng, Wang	(Academia Sinica, Taipei, Taiwan)
LWW	Lo, Wang, Wang	(Academia Sinica, Taipei, Taiwan)
JR	Ren, Wu	(Tsing Hua University, Hsinchu, Taiwan)
WR	Wu Ren	(Tsing Hua University Hsinchu, Taiwan)
WJ	Wu, Jang	(Tsing Hua University, Hsinchu, Taiwan)
RW	Ren, Wu, Jang	(Tsing Hua University, Hsinchu, Taiwan)
LBLJK	Lim, Byun, Lee, Jang, Kim	(University, Seoul, Korea)
GT	Tzanetakis	(University of Victoria)
SSPK	Seyerlehner, Schedl, Pohle, Knees	(Johannes Kepler University, Linz, Austria)
SSKSS	Seyerlehner, Schedl, Knees, Sonnleitner, Schlüter	(Johannes Kepler University, Linz, Austria)
ME	Mandel, Ellis	(Columbia University, NY)
WLB	Wack, Laurier, Bogdanov	(Universitat Pompeu Fabra, Barcelona, Spain)
PP	Panda, Paiva	(University of Coimbra)
PH	Hamel	(Université Montréal, Canada)

2.3.3 Applications

Notre premier système de classification fut développé dans le cadre du projet `Cuidado` pour la reconnaissance des échantillons d'instruments de musique [138]. Le système permettait la reconnaissance de 27 instruments à l'aide d'un algorithme de classification Gaussien hiérarchique. Herrera considère dans [76] ce système comme une bonne représentation de l'état de l'art des performances dans ce domaine. `ircamclassication1` a servi dans le projet `Ecoute` à la création des systèmes dédiés à la reconnaissance de l'humeur et du genre musical pour la compagnie WMI et pour la segmentation parole/musique pour la compagnie Dalet [145]. `ircamclassification2` a servi dans le projet `SampleOrchestrator` à la création des systèmes dédiés de reconnaissance des matériaux et des onomatopées pour la compagnie Univers-Sons [214] ; et dans le projet `Quaero` à la reconnaissance de l'humeur et du genre musical pour les compagnies Orange et Exalead. `ircamclassification3` a servi dans le projet `Quaero` à la création des systèmes dédiés à la reconnaissance de l'instrumentation, de l'humeur et du genre musical ainsi qu'à la segmentation voix chantée/instrumental pour les compagnies Orange [152] et Exalead [97]. La Figure 2.6 représente l'interface des moteurs de recherche MSSE d'Orange (à gauche) et MUMA d'Exalead (à droite). Le volet de droite (respectivement de gauche) de cette interface représente les nuages de « tags » permettant la navigation multi-label. Ces tags sont estimés à l'aide d'`ircamclassification3`. La conception de l'IHM de MSSE a été faite en collaboration avec l'IRCAM. Notons que les techniques UBM et ARM (voir partie 2.2.1.2 et 2.2.1.1) n'ont pas uniquement été choisies pour leurs très bonnes performances. En effet, ces techniques sont également utilisées dans notre système de recommandation par similarité musicale (voir partie 2.5). Le partage de ces descripteurs entre ces deux systèmes permet donc de diminuer sensiblement le coût de calcul total.

Remerciements. Si c'est une chose de construire un système générique atteignant de bonnes performances à `MIREX`, c'en est une autre de construire un système de classification réellement fonctionnel, pouvant être intégré auprès d'un industriel. L'article [152] décrivant le travail effectué pour l'intégration des technologies IRCAM dans le moteur de recherche MSSE Orange rend en cela bien compte de la lourdeur de ce travail. Je voudrais donc remercier ici Juan José Burred, Frédéric Cornu, Mathieu Ramona, Damien Tardieu et Christophe Charbuillet pour leur courage dans ces tâches souvent ingrates en publications.

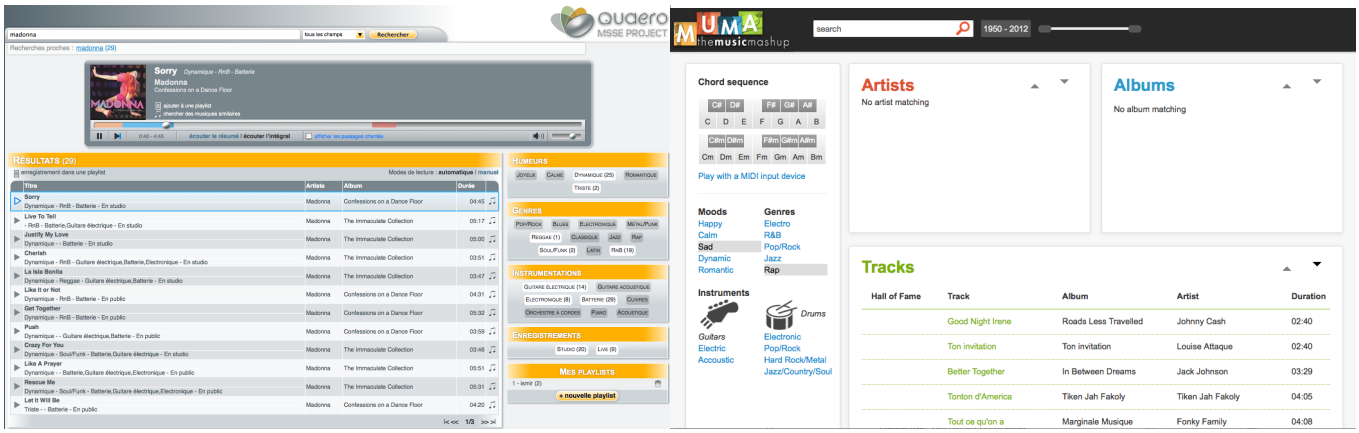

FIGURE 2.6 – Interface de la partie musique du moteur de recherche MSSE d'Orange et MUMA d'Exalead.

2.4 Descripteurs audio spécifiques

Les systèmes de classification présentés jusqu'à présent reposaient tous sur le même schéma : extraction d'un grand nombre de descripteurs génériques, sélection automatique de descripteurs, transformation, algorithme de classification. Il existe cependant des problèmes de classification pour lesquelles l'utilisation de descripteurs génériques ne permet pas une description adéquate des propriétés à mettre en évidence. Lorsqu'une description sémantique du problème est possible, il est alors avantageux de développer des descripteurs spécifiques pour le problème à résoudre. J'en donne deux exemples ci-après.

2.4.1 Descripteurs morphologiques

Alors que la description des échantillons vise généralement la reconnaissance de leur source (par exemple le nom de l'instrument de musique ayant produit le son), le projet `Ecrins` [8] vise à permettre la description de sons abstraits, effets sonores, sons non-naturels ou synthétiques dont la source est généralement inconnue ou non reconnaissable. A l'exception de [186] ou plus récemment de [48], ce type de description a fait l'objet de peu de recherches. Dans ce projet nous proposons de décrire ces sons à l'aide d'un ensemble de profils morphologiques dérivés des propositions de Pierre Schaeffer [197]. Ces profils morphologiques décrivent l'évolution temporelle du contenu d'un son selon plusieurs points de vue. Ces évolutions sont rangées dans des profils « types » illustrés à la Figure 2.7. Dans l'article de journal [153], nous étudions l'estimation automatique de deux de ces points de vue. Nous les considérons comme deux problèmes de classification, chacun à cinq classes : les profils dynamiques (stable, montant, descendant, montant/descendant, impulsif) et mélodiques (stable, montant, descendant, montant/descendant, descendant/montant). Chacune des classes est illustrée par un ensemble de sons choisis par Emmanuel Deruty. Pour chacun des deux problèmes, nous comparons ensuite la prédiction de leurs classes à l'aide de descripteurs génériques et à l'aide de descripteurs spécifiquement développés. Les deux ensembles de descripteurs sont utilisés dans le même système de classification, `ircamclassification1`. Pour ces deux problèmes, les descripteurs génériques conduisent à des rappels-moyens de 76% et 48% , les descripteurs spécifiques à 97% et 73%.

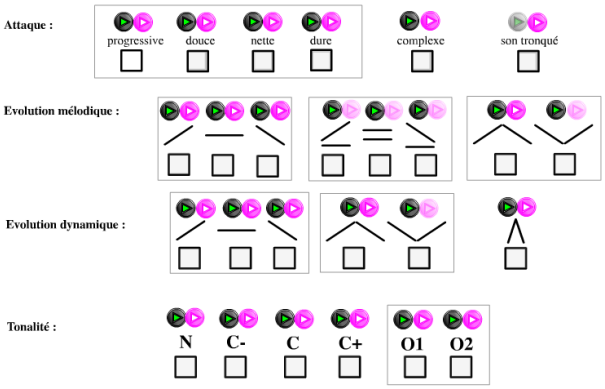

FIGURE 2.7 – Profils morphologiques du projet `Ecrins`.

8. Projet en collaboration avec le GRM.

2.4.2 Description de la voix chantée

La description de la voix chantée constitue un deuxième exemple de l'utilisation de descripteurs spécifiques. La voix chantée étant l'élément focalisant le plus l'attention de l'auditeur (du moins en musique populaire et lyrique), je décide en 2008 de démarrer une recherche sur ce sujet au travers du stage de Master [179] et de la thèse de doctorat [180] de Lise Régnier que j'ai encadrés. Cette recherche vise non pas à décrire l'aspect mélodique de la voix [9] mais les caractéristiques permettant d'une part d'isoler dans un morceau les segments de voix chantée, d'autre part de permettre l'identification du (de la) chanteur(euse) et d'établir des similarités entre chanteurs(euses).

Etat de l'art. La description de la voix chantée a fait l'objet de nombreuses recherches : soit dans le but de localiser temporellement les segments vocaux (ou de séparer la source vocale), soit dans le but d'identifier le(la) chanteur(se). Notons que dans ce dernier domaine une certaine confusion règne entre l'identification de l'artiste (The Beatles), reposant souvent sur une description incluant l'accompagnement, et l'identification du(de la) chanteur(se) (Paul McCartney ou John Lennon). Parmi les nombreuses techniques proposées (reconnaissance de phonèmes [41], modèle MFCC/GMM [99], descripteurs génériques/SVM [178], critères d'harmonicité [37] [224], méthode TIFCT [104], re-synthèse des sinusoïdes de la mélodie principale [58] [113]), le travail de Lise Régnier se situe dans la continuité de celui de Lachambre [94] et l'utilisation de descriptions de type vibrato. Le modèle utilisé par Lise Régnier est un modèle intonatif, incluant le vibrato, et la « thèse soutenue » est que cette intonation du(de la) chanteur(se) fait partie de sa signature vocale.

Descripteurs intonatifs. Le point de départ de cette recherche est le résultat étonnant obtenu lors du stage de Master, publié dans l'article [181], montrant que l'utilisation d'un seul descripteur spécifique, le vibrato, conduit à des résultats de segmentation en partie chantée/instrumentale équivalants à ceux obtenus par un système de classification reposant sur des descripteurs génériques. Ce résultat donnera lieu pendant la thèse au développement de descripteurs « intonatifs » (décrivant le contour intonatif de la mélodie vocale par caractérisation du vibrato, tremolo, portamento et legato). Pour l'estimation de ces descripteurs une estimation sinusoïdale est d'abord effectuée sur le signal polyphonique à l'aide du logiciel `Pm2` de l'IRCAM [190]. Les descripteurs intonatifs sont issus de l'analyse de l'évaluation temporelle de ces sinusoïdes. A l'inverse des sinusoïdes elles-mêmes, les paramètres de ces sinusoïdes sont des signaux basse fréquence observés sur des durées courtes. Dans la thèse, différents algorithmes d'estimation des paramètres de fréquence et d'amplitude de ces signaux sont étudiés. Les meilleurs résultats sont obtenus à l'aide de la méthode « haute résolution » proposée dans [12]. Ce modèle intonatif peut également être utilisé pour regrouper les composantes appartenant à la voix chantée. Dans [182], un algorithme de clustering hiérarchique par agglomération reposant sur des critères CASA (Computational Auditory Scene Analysis, incluant des relations d'harmonicité et d'étendue temporelle) est proposé afin de regrouper les partiels en classes. Nous montrons que ces classes correspondent (souvent) aux sources, dont une est la voix chantée.

Descripteurs de timbre. En complément de la description de l'intonation de la voix, la description du « timbre » vocal est également étudiée. Dans [183], nous proposons pour cela l'utilisation des True-Envelope Cepstral Coefficients ou TECC (coefficients cepstraux dérivés de la True Envelope [191]) et montrons que ceux-ci sont plus performants que les représentations de type MFCC et LPC.

Combinaison d'algorithmes de classification. Les représentations de type intonatif et TECC orbitant dans des domaines différents (global pour une note et local en trame, local en harmonique et global en fréquence), nous proposons dans [183] une méthode originale pour les combiner tout en préservant leur performances relatives. La méthode repose sur un schéma de classification à mi-chemin entre la cascade d'algorithmes de classification et leur parallélisation. Ce schéma de classification est utilisé pour une tâche d'identification du chanteur. Une large évaluation des différents descripteurs et différents algorithmes de classification (SVM, GMM, KNN) est effectuée en s'appuyant sur deux corpora représentant le chant lyrique et le chant en musique populaire. Nous montrons que le système utilisant la combinaison des algorithmes SVM, des descripteurs TECC et intonatifs permet d'améliorer les résultats sur les deux bases par rapport à l'utilisation d'un MFCC/SVM (de 73.4% à 89.1%).

Vérification du chanteur. Dans [184], les descripteurs TECC et intonatifs sont utilisés pour une tâche de vérification de chanteur. Nous montrons que pour cette tâche, l'EER (Equal Error Rate) obtenu par les descripteurs intonatifs est meilleur (inférieur) que celui obtenu par les TECC. La combinaison des deux descripteurs diminue encore davantage l'EER.

Cette recherche montre également la pertinence de la création de descripteurs spécifiques à une tâche.

9. L'estimation de la mélodie est traitée dans d'autres recherches de l'équipe Analyse/Synthèse de l'IRCAM. Nous l'avons également ébauchée lors du co-encadrement de Justin Salomon [194].

2.5 Recommandation musicale par similarité acoustique

Alors que la classification vise à poser une étiquette c_i sur un document n (un segment ou un morceau entier), cette partie ainsi que la suivante concerne des recherches effectuées dans le cadre d'applications inter-documents (nécessitant l'accès à un ensemble de documents).

La recommandation musicale par similarité acoustique vise à recommander à un utilisateur une liste de morceaux de musique $n_i,\ i \in I$ de contenu acoustique similaire à un morceau cible n_c [10]. Cette similarité peut évidemment être effectuée par recherche d'items de « qualité » (classes) équivalentes comme le propose [205]. Nous étudions ici la similarité reposant sur une description « quantitative ».

Contrairement aux recherches faites sur la similarité des sons instrumentaux (voir partie 2.1), celles sur la similarité entre morceaux ne reposent pas sur le résultat d'expériences perceptives. Les corrélats acoustiques proposés pour exprimer la proximité entre morceaux restent donc hypothétiques : l'instrumentation ? le rythme ? les caractéristiques vocales ? Il est en effet difficile d'effectuer des expériences de similarité/dissimilarité sur des morceaux du fait de l'aspect sans doute encore beaucoup plus multidimensionnel du « timbre musical » et du fait de phénomènes quasi inévitables d'identification et/ou de catégorisation en musique. Signalons cependant les expériences effectuées au MTG par [70], qui par des techniques de transformations du signal ouvrent de nouvelles voies dans la compréhension du timbre musical. Malgré ce flou, l'ensemble des caractéristiques expliquant cette similarité est souvent (abusivement) désigné sous le nom de « timbre ». Les performances d'un algorithme de similarité musicale sont évaluées a posteriori : soit par comparaison des métadonnées entre le morceau cible et les morceaux recommandés [11], soit par des tests perceptifs effectués [12]. Dans le projet `Quaero`, nous avons proposé de comparer les recommandations fournies par un algorithme à celles fournies par le site Pandora [13]. Outre la nécessité qu'elle soit pertinente, la recommandation musicale basée sur la similarité acoustique possède d'autres défis. Premièrement, elle doit pouvoir passer à l'échelle (« scalability ») [14], ensuite le système de recommandation doit permettre que l'ensemble de la base puisse être recommandée (éviter les items orphelins) et enfin il doit permettre d'éviter qu'un item ne focalise toutes les recommandations (éviter les items attracteurs).

Etat de l'art. De par son importance applicative, ce sujet a fait et fait toujours l'objet de nombreuses recherches. Les premières datent du début des années 2000 avec les travaux de Logan [103], Aucouturier et Pachet [10] ou encore Ellis et Whitman [45]. Ces travaux seront poursuivis plus tard par (entre autre) l'école dite « autrichienne » : Pampalk, Pohle, Seyerlehner, Schnitzer [128] [205] [202]. L'approche traditionnelle pour calculer la distance entre deux morceaux n et n' est l'approche dite « bag of frames ». Dans celle-ci, le contenu d'un morceau n est représenté par des MFCCs, eux-mêmes représentés par un modèle Gaussien ou un GMM. Le terme « paquet »/« bag » dénote le fait dans ces modèles la notion de temporalité est perdue. La distance entre deux morceaux est calculée par la divergence symétrisée de Kullback-Leibler (SKLD) entre leurs modèles respectifs (ou par l'intermédiaire d'une méthode de type Earth-Moving-Distance ou Monte-Carlo dans le cas des GMMs). L'ensemble des distances $\hat{dist}(n, n')$ sert ensuite à la construction d'une matrice. Dans le cas de la recommandation, différents procédés de normalisation de cette matrice peuvent être utilisés afin de limiter la présence d'orphelins et d'attracteurs [171]. Parmi les autres techniques proposées, signalons celle de Casey et Slaney [33] reposant sur une représentation de type « shingles » (concaténation vectorielle de trames temporelles) et permettant l'utilisation d'une norme L_2 entre ces représentations.

Contributions. Le travail de Christophe Charbuillet, que j'ai encadré, s'est focalisé sur le problème du passage à l'échelle.

L'objectif du projet `DISCO` (en collaboration avec le laboratoire CEDRIC du CNAM et le LAMSADE de Paris Dauphine) est l'étude des structures d'index les plus pertinentes pour l'accès optimisé aux données en très grand volume. Dans [35] nous étudions le passage à l'échelle des approches reposant sur la SKLD. Pour de très grandes bases de données, le calcul exhaustif des distances n'est pas envisageable. Des techniques optimisées de recherche doivent être utilisées, le choix de celles-ci dépend des propriétés des données. Dans le cas où la norme L_2 peut être utilisée, de nombreuses techniques d'optimisation de recherche existent (comme le Local Sensitive Hashing). Cependant SKLD n'est ni une norme L_2 ni une métrique (non respect de l'inégalité triangulaire). De ce fait, les structures d'index optimisées pour les métriques (comme le M-tree) ne sont pas non plus utilisables.

10. Notons qu'il n'est pas possible aujourd'hui d'établir le point i dans cette liste avant lequel les morceaux sont similaires à la cible et au-delà duquel ils ne le sont plus. La liste est donc fournie par ordre décroissant de similarité et affichée de la sorte. De même, peu de recherches prennent en compte les goûts de l'utilisateur dans cette recommandation.

11. A défaut d'autres informations les morceaux d'un même album, artiste ou genre musical sont considérés proches.

12. Par exemple, le système « Evalutron » de IMIRSEL permet de comparer les recommandations proposées par les différents algorithmes soumis à `MIREX`.

13. Le site `www.pandora.com` est considéré comme une référence en matière de recommandation dans la communauté MIR. Ces recommandations utilisent une description par 400 critères, annotés manuellement, de chaque morceau.

14. La « scalability » désigne ici le fait de pouvoir obtenir une estimation de la similarité en un temps réduit, soit par calcul exhaustif mais très rapide de ces distances, soit par déduction de ces distances à l'aide de données pivots.

Dans [35], nous proposons une modification de la SKLD de la forme $x \rightarrow \sqrt{log(x+1)}$ permettant de la rendre métrique et donc d'appliquer un algorithme de type M-tree. Sur une base d'un million d'objets, le coût de recherche est réduit à 90.8% de celui d'une recherche séquentielle ; ceci en gardant une précision de 100% (recherche exacte). Un algorithme est ensuite proposé afin de permettre le contrôle de la concavité de cette fonction. Cette concavité influe sur le respect de l'inégalité triangulaire (donc sur l'exactitude des résultats de la recherche) mais également sur sa dimension intrinsèque (donc son « indexabilité ») [15]. Cet algorithme permet de descendre à 8.2% du coût d'une recherche exhaustive pour une précision de 88% (recherche approximée).

Dans le projet Quaero, nous proposons une alternative à l'approche SKLD [36]. Celle-ci s'inspire des techniques utilisées en identification du locuteur [185] représentant les descripteurs au travers de super-vecteurs (voir partie 2.2.1.2). Les descripteurs modélisés sont les Mel Frequency Cesptral Coefficient [112] et les Spectral Flatness Measure [82]. Le principal avantage des super-vecteurs est de permettre l'utilisation d'une norme L_2. Le coût de calcul est ainsi réduit d'un facteur proche de 100 par rapport à l'approche SKLD. Deux nouvelles techniques de normalisation de la matrice de distance sont proposées pour éviter les orphelins et les attracteurs. Elles reposent sur la projection des super-vecteurs sur une sphère unitaire centrée soit sur l'UBM, soit sur le super-vecteur moyen d'une base d'apprentissage. Ces techniques sont comparées à l'état de l'art (Tim Pohle) et montrent une amélioration des performances (mesurées par « artist-filtered genre match ») de 48.3% à 52.6%. Cette méthode de calcul de similarité entre morceaux, complétée par une modélisation de type ARM (voir partie 2.2.1.1), a été soumise à l'évaluation MIREX 2011. Les tests d'évaluation perceptifs l'ont placée à la première place ex-aequo avec la soumission de Seyerlehner. Dans le projet Quaero, cet algorithme a été intégré dans les moteurs de recherche MSSE d'Orange et CMSE d'Exalead.

2.6 Identification audio par technique de signature

Les techniques d'identification audio visent à reconnaître un enregistrement particulier r d'un morceau w (voire un instant particulier t dans cet enregistrement) à partir de l'observation d'une diffusion de son signal audio. Ce signal peut être très dégradé (canal de diffusion de bande passante réduite, bruits additifs, dilatations temporelles, coupures ...). A la différence des techniques d'identification de morceaux w (cover-song identification, query-by-humming), l'objectif est ici de distinguer les différents enregistrements r d'un même morceau w : différents artistes, différentes versions studio, version concert ... Pour cette tâche, la technique du « tatouage audio » (**watermarking**) a d'abord été proposée. Celle-ci consiste à introduire un code identifiant i_r dans le signal audio. Ce code se doit d'être inaudible mais résistant aux dégradations. L'identification consiste donc à retrouver ce code dans un signal diffusé potentiellement dégradé. De ce fait, toutes les occurrences (CD, fichiers numériques ...) d'un même enregistrement r doivent avoir été préalablement tatouées. Depuis la fin des années 1990, la technique de l'« empreinte/signature » (**fingerprint**) audio a été proposée comme alternative au watermarking. Dans cette technique une signature f_r est extraite de chaque enregistrement r (de chaque instant t de chaque enregistrement r). Contrairement au watermarking, cette technique ne nécessite donc l'accès qu'à une seule occurrence de chaque enregistrement. L'ensemble des signatures est ensuite stocké dans une base. Lors de l'observation d'un signal inconnu, le même algorithme d'extraction de signature est utilisé et la signature extraite comparée à l'ensemble des signatures de la base. Si une correspondance est trouvée, celle-ci fournit l'identification de l'enregistrement correspondant au signal inconnu. Les deux verrous technologiques de cette technique concernent la conception de la signature et la recherche dans la base. La signature doit être à la fois discriminante (afin d'éviter la confusion entre enregistrements) et robuste (la même signature doit pouvoir être obtenue sur un signal dégradé), La recherche dans la base doit pouvoir être effectuée de manière très rapide. Ceci peut être obtenu par choix d'une signature de taille réduite (code compacte ou binaire) et/ou par utilisation d'algorithmes de recherche optimisés.

Etat de l'art. Même si l'identification audio par signature est sans doute la technologie MIR qui a connu le plus de succès auprès du grand public (à travers entre autre Shazam), le nombre de publications dans ce domaine n'est pas très important. De même, sans doute pour des raisons de compétitivité, les performances des techniques proposées n'ont pas été comparées. Parmi les techniques les plus connues, mentionnons celle de Philips [72] reposant sur une représentation compacte de la différence de l'énergie en sous-bandes et une recherche exacte par table de hachage ; celle de Shazam [216] reposant sur des clefs représentant des paires de pics spectraux et une recherche par accumulation temporelle de ces clefs (améliorée par Télécom ParisTech [51] par utilisation de la transformée à Q-constant) ; l'AudioID de Fraunhofer [3] reposant sur un schéma classique de plus proche voisin sur des descripteurs MPEG-7 et l'AudioDNA du MTG de Barcelone [28] reposant sur un codage de type HMM de la suite temporelle d'évènements.

15. Une dimension intrinsèque faible indique une grande disparité des distances entre objets, ils sont donc clusterisables et donc indexables, une dimension élevée indique que les objets sont équidistants et donc peu indexables.

Etat de l'art à l'IRCAM. Les contributions de l'IRCAM commencent avant l'état de l'art mentionné ci-dessus. En 1998, Laurent Worms et Xavier Rodet développent un premier système d'identification par technique de signature audio [219]. Le choix se porte sur une signature de taille réduite et permettant l'utilisation de la norme L_2 (ces deux points visent à accélérer la phase de recherche dans la base). Cette technique repose sur la modélisation spectrale de l'évolution temporelle du contenu énergétique à différentes fréquences (voir les « descripteurs dynamiques » dans la partie 2.2.1.1). Le signal audio d'un enregistrement s_r autour du temps o est codé par la matrice $Y(k, k', o)$. Le choix des bandes k et k' est effectué par utilisation d'un critère d'information mutuelle. La taille de la signature obtenue est de 13 nombres flottants pour 10 s. La recherche est ensuite effectuée par un algorithme de type « Branch-and-Bound ». Du fait d'une demande de brevet international [189], cette méthode ne sera pas publiée avant 2011.

Contribution. Ma contribution à cette recherche démarre en 2001 dans le cadre du projet `Cuidado`. L'objectif est la création d'un système de surveillance web de la musique (web-music-monitoring-system). Dans ce projet, un corpus de test de 1000 enregistrements est constitué (le système de [219] avait été développé et évalué sur un corpus de test de seulement 14 enregistrements). L'évaluation des performances du système [219] me conduit à apporter un premier ensemble de modifications dans le but d'augmenter la discrimination du code et de réduire l'influence du canal de diffusion et des bruits additifs. Faute d'accès à un corpus de test plus important cette recherche est arrêtée en 2004 et n'est reprise qu'en 2010 dans le cadre du projet `Quaero`. Dans ce projet la compagnie Yacast fournit des données à grande échelle permettant de redévelopper cette technique. L'objectif est cette fois la reconnaissance d'enregistrements dans un flux audio (flux radio) avec comme unique connaissance les signatures d'extraits audio de 30 s. Ceci présente un certain nombre de défis supplémentaires du fait des traitements audio utilisés en radio (dilatation temporelle, versions remontées, radio-edits) et du fait de la connaissance uniquement d'un extrait de 30 s (les fausses alarmes peuvent donc être nombreuses). La recherche est cette fois effectuée par Mathieu Ramona que j'encadre. Dans la suite, je résume thématiquement ces deux ensembles de contributions.

Amélioration de la signature. Alors que dans [219] l'échelle fréquentielle k (voir eq. 2.1) exprimait les fréquences de Fourier, j'ai proposé de l'exprimer en bandes perceptives (bandes de Bark [225]). Cette modification permet de représenter la totalité du contenu spectral tout en gardant une taille réduite. Dans le projet `Quaero` [177], nous avons remplacé ces filtres rectangulaires de Bark par des filtres cosinusoïdaux permettant une meilleure résistance aux transformations de type transposition. Alors que dans dans [219], la modélisation spectrale était directement appliquée au contenu énergétique ($d_m^k = |X(k, m|^2$ dans eq. 2.1), j'ai proposé de l'appliquer à la log-énergie ($d_m^k = \log(|X(k, m|^2))$ [16]. Le but est de s'affranchir du canal de transmission du signal. En effet, si les propriétés de ce canal de diffusion sont constantes au cours du temps, sa contribution est un terme additif, également constant au cours du temps, dans le spectre de log-énergie, donc une composante continue dans la seconde transformée de Fourier $Y(k, k', o)$.

Amélioration de la stratégie de recherche. Dans la méthode initiale chaque signature est calculée sur (et représente) un horizon de M=10 s. Pour l'identification, une seule recherche de signature dans la base est effectuée. La présence d'un bruit parasite sur la durée L_m modifie la totalité de la signature. La probabilité de la présence d'un tel bruit augmentant avec la durée L_m, j'ai donc décidé de réduire cette durée (à 2 s) et de pallier la diminution de discrimination résultante par une méthode de cumulation temporelle des recherches. Si nous notons (r_i, t_j) l'entrée dans la base correspondant à la signature de l'item r_i au temps t_j et $f(t)$ celle de la donnée inconnue mesurée au temps t, la méthode de cumulation cherche le minima de $d'(i, j) = \sum_{\delta=0}^{\Delta} d\left(f(t+\delta), (r_i, t_j + \delta)\right)$. Dans le projet `Quaero` [175], il apparaît que ces distances sont peu pertinentes mais que leur rang l'est. La méthode est donc modifiée vers une méthode de « rang ». Pour chaque temps t, les L plus proches entrées de la base (r_l, t_l) sont gardées. Sur une fenêtre glissante Δ, seules les entrées correspondant à des r_l revenant plusieurs fois sont considérées. Pour celles-ci, nous analysons leur alignement temporel avec celui du signal d'entrée. La décision de détection est prise en fonction de cette analyse. Ceci est illustré sur la Figure 2.8.

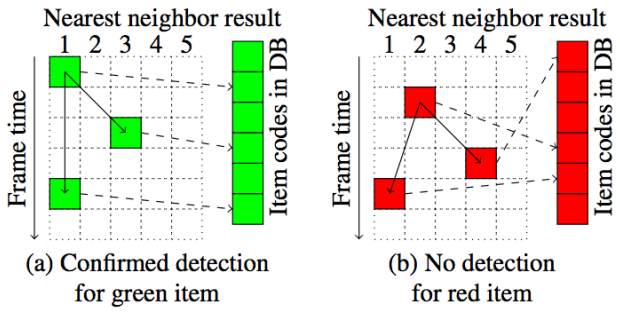

FIGURE 2.8 – Algorithme de recherche d'identification audio, d'après [177].

Synchronisation des codes. Dans [175], la robustesse de la signature IRCAM aux dégradations typiques du signal (encodage/décodage mp3 ou GSM, compression d'amplitude, égalisation, addition de bruit ...) est

16. En pratique, afin de mieux correspondre à la perception (et d'éviter les singularités en zéro de la fonction logarithmique) nous utilisons une échelle de Sone [50].

étudiée. Cette étude montre une bonne robustesse de la signature à la plupart de ces dégradations à l'exception du décalage temporel. La signature étant extraite toutes les 0.5 s, le décalage maximal entre signature extraite et signature de la base est de 0.25 s. Pour remédier à ce décalage, une méthode de calage de l'extraction des signatures sur les onsets du signal est proposée. Dans [175], celle-ci repose sur la détection d'onset par le logiciel `SuperVP` de l'IRCAM. Dans [177], une nouvelle méthode de synchronisation, reposant directement sur l'analyse des $|X(k,m|^2$ est proposée. Cette méthode beaucoup plus économique en temps de calcul est également plus robuste.

Evaluation. Les résultats de ce système, appelé `AudioPrint`, sont comparés à ceux de systèmes représentant Shazam [17] et Philips [18]. Sur un corpus de 240 heures de programmes radiophoniques codés en WMA 10kbps, les scores [19] sont de 98% pour `AudioPrint`, 89.5% pour Philips et 84.6% pour Shazam.

L'ensemble de cette recherche à nécessité la création de corpora proprement annotés (en particulier pour les études concernant le décalage temporel). Cela requiert l'annotation exacte de début et de fin de l'extrait de 30 s dans le flux radio. La difficulté de cette tâche provient, entre autre, du fait que la radio diffuse des items de durées pouvant être modifiées, la captation de ces flux pouvant également être tronquée. Pour de larges bases de données, cette annotation est une tâche lourde. Aussi, dans [176], une technique reposant sur notre `AudioPrint` a été développée dans le but de corriger automatiquement les annotations. Cette technique a permis la création d'un corpus `SyncOccur` mis à disposition de la communauté pour servir de référence aux travaux futurs. Cette technique a également servi au développement par la compagnie Vizion'R d'un prototype d'application iOS permettant la synchronisation des paroles à partir de l'identification d'un item capté à l'aide du micro d'un iPhone (voir Figure 2.9).

FIGURE 2.9 – Application iOS d'identification audio et de synchronisation temporelle de paroles développée par la société Vizion'R, reposant sur la technologie `AudioPrint` de l'IRCAM.

Dans le but de permettre à la communauté de recherche de comparer les performances de technologies d'identification audio, nous avons proposé dans [174], avec les partenaires du projet `Quaero`, un environnement d'évaluation publique. Celui-ci est constitué de l'implémentation d'un ensemble de métriques de références et d'un corpus radio annoté.

17. Shazam est représenté par l'implémentation de Dan Ellis.
18. Philips est représenté par une ré-implémentation de la méthode [72] faite par Mathieu Ramona.
19. Le score est ici défini comme la différence entre le nombre d'identifications correctes et le nombre de fausses alarmes.

CHAPITRE 3

Estimation de paramètres relatifs à la notation musicale

Dans cette partie je décris mes travaux portant sur l'estimation de paramètres relatifs à la notation musicale [1]. Ces paramètres concernent la description du rythme d'un morceau (tempo, métrique, position des battements et des premiers temps) et de son contenu harmonique (tonalité et suite d'accords). Notons que l'estimation des hauteurs de notes ainsi que leur assignation à des sources ne sont pas traitées ici mais font l'objet d'autres recherches dans l'équipe Analyse/Synthèse des sons de l'IRCAM [2]. A la différence des paramètres étudiés dans la partie 2, les paramètres « musicaux » ont la particularité d'avoir une définition sémantique (relativement) claire et peuvent dès lors être estimés par des algorithmes dont la connaissance est explicitement introduite (approche « **human-learning** »). Je présente cependant également des travaux incluant l'utilisation d'apprentissage machine pour leur estimation. Bon nombre de ces paramètres ont une définition **locale en temps** : soit que le concept varie dans le temps (le tempo, la tonalité, les accords), soit que le concept est le temps lui-même (la position des battements et des premiers temps). Ces paramètres « temporels » sont généralement mutuellement ou conditionnellement **dépendants** au travers d'une composition musicale (les accords varient généralement sur une grille de battements). Je représente ces dépendances à la Figure 3.1. Pour cette raison de temporalité et de dépendance temporelle, j'adjoins également dans cette partie mes travaux sur l'estimation d'une structure musicale, même s'il est vrai que cette description ne fait pas partie d'une notation musicale et n'a pas de sémantique claire.

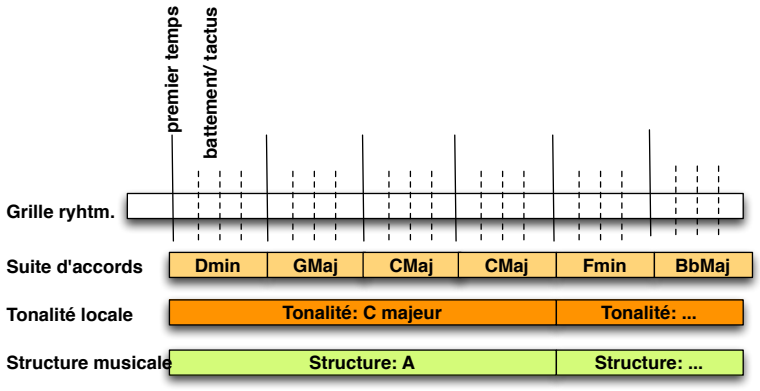

FIGURE 3.1 – Illustration des confusions possibles entre niveaux métriques en l'absence de partition et de connaissance du tempo.

3.1 Estimation de paramètres relatifs au rythme

Au cours de mes recherches sur l'estimation des paramètres « musicaux », l'estimation de la position des battements et des premiers temps m'est apparue rapidement comme centrale. En effet, ceux-ci permettent la définition d'une grille de lecture du temps et ainsi une interprétation sémantique et musicale des paramètres extraits au cours du temps [3]. De manière générale, de nombreux évènements musicaux se synchronisent sur

1. Je considère ici la notation occidentale de la musique.
2. Voir par exemple [222] pour l'estimation de hauteurs de notes et par exemple [187] et [21] pour la transcription en évènements percussifs.
3. On pourra ainsi passer d'une observation de type « les descripteurs changent aux instants 11.5 s et 12.5 s » à une interprétation de type « ces descripteurs changent sur les premiers temps ».

cette grille temporelle (accords, structure, hauteurs). Pour cette raison, depuis 2005, j'ai effectué ou encadré un nombre important de recherches sur ce sujet et publié 10 articles (dont 4 de journaux). Il est donc difficile de présenter toutes ces recherches ici. Aussi ai-je choisi de présenter deux de ces articles en annexes [4] et de décrire ici uniquement l'approche globale sous-jacente à mes recherches. Je joins en annexe une explication de la terminologie que j'utiliserai pour la description du rythme.

3.1.1 Etat de l'art

Le nombre d'algorithmes proposés dans le domaine de la description du rythme est très important (du moins pour l'estimation du tempo et de la position des battements, déjà moins pour la métrique et la classification, encore moins pour celle du premier temps ou du tempo perceptif). Ce domaine fait de plus l'objet d'un regain d'intérêt grâce, entre autres, aux recherches de l'INESC de Porto. Je me réfère donc aux états de l'art dressés dans mes publications [147] et [164], jointes en annexe de ce document, et à celui dressé par Gouyon et Dixon dans [66]. Je reprends uniquement ici un bref résumé. De manière globale, les méthodes proposées se divisent en deux catégories : celles basées sur une conversion audio-vers-symbolique et la détection d'onsets [42] et celles reposant sur des critères de variation énergétique (oscillateurs ou fonctions d'observation) [198]. Les périodicités de ces onsets, ou fonctions, sont ensuite utilisées pour inférer directement le tempo ou estimer simultanément l'ensemble de la structure métrique (tatum, tactus, mesure, facteur de swing), ceci à travers l'utilisation de modèles probabilistes [88] ou multi-agents [65]. D'autres types d'observations ont également été proposés dans le but de fournir un contexte informatif de plus haut niveau (comme la variation des Chromas [65] [88]). Récemment, [20] a proposé une approche exclusivement « machines » reposant sur l'utilisation de réseaux de neurones récursifs pour l'apprentissage des spécificités du signal autour des battements.

3.1.2 Problématiques

Notre objectif est l'estimation des paramètres relatifs au rythme d'un morceau : son tempo, sa métrique, la position de ses battements, premiers temps et les caractéristiques propres à son pattern rythmique. Malgré un état de l'art considérable, ce problème est encore aujourd'hui loin d'être résolu en dehors du cas de musiques de rythmes stéréotypés (pop, techno). Je liste ici les principaux problèmes rencontrés dans ce domaine.

Ambiguïté des niveaux métriques à estimer. Cette ambiguïté provient de l'aspect pyramidal des regroupements possibles d'un rythme (deux mesures en 2/4 peuvent souvent être regroupées en une mesure en 4/4). En l'absence de connaissance du tempo, plusieurs choix de niveau de regroupement peuvent également être possibles (une mesure en 6/8 à 60 bpm peut être considérée équivalente d'un point de vue acoustique à deux mesures de 3/4 à 180 bpm). Cette ambiguïté, et le fait que les algorithmes d'estimation produisent très souvent des erreurs « d'octave », ont conduit la communauté à utiliser deux mesures d'évaluation (Accuracy-1/Accuracy-2) considérant comme une erreur ou pas un changement de niveau hiérarchique par rapport au tempo de référence. Encore faut-il, quand la partition n'est pas connue, pouvoir définir ce tempo de référence. Dans ce cas, seule la perception peut déterminer le tempo. Moelants [116] montre à ce propos que pour un même morceau, les auditeurs peuvent percevoir différents tempi. Pour ces différentes raisons, dans nos algorithmes, l'estimation du tempo sera faite simultanément à celle d'un niveau métrique à travers l'utilisation de « gabarits-périodiques » (voir partie 3.1.3.3) : nous estimons simultanément le niveau correspondant à la mesure, au tactus et au tatum. De même, depuis la mise à disposition de données perceptives, celles-ci (lorsque des majorités claires se dégagent dans les jugements) seront considérées comme références (voir notre étude [158] décrite dans la partie 3.1.3.3).

Non-proéminence d'évènements. Une des approches possibles pour l'estimation de la position des battements, approche dite « temporelle », est de détecter la présence d'onsets dans le signal, et de déduire la période correspondant au tempo de l'analyse de l'inter-distance de ces onsets. Lors du stage de Jean-Baptiste Goyeau [68], que j'ai encadré, nous avons montré que cette approche est très sensible aux faux positifs et fausses réjections de la détection d'onsets. L'approche que j'ai développée est dès lors une approche dite « spectrale ». Dans cette approche, un ensemble de fonctions d'observation temporelles est extrait du signal. Ces fonctions ne représentent pas seulement la variation d'énergie mais différents points de vue sur la variation du contenu du signal susceptibles d'être corrélés au tempo (voir partie 3.1.3.1). Les périodicités de ces fonctions temporelles sont ensuite analysées (voir partie 3.1.3.2) afin d'estimer le tempo. Ensuite, à l'aide du tempo estimé et de ces fonctions d'observations, nous estimons le meilleur emplacement des battements et premiers temps [5].

4. Ceux décrivant les algorithmes d'estimation du tempo et de la métrique [147], de la position des battements et premiers temps [164] sous-jacents au logiciel `ircambeat`.

5. Cette approche spectrale (fonction-d'observation continue → mesure de périodicité → localisation temporelle) correspond également à celle que j'avais développée dans le cadre de ma thèse de doctorat [137] pour l'estimation des instants de fermeture de la glotte (IFG).

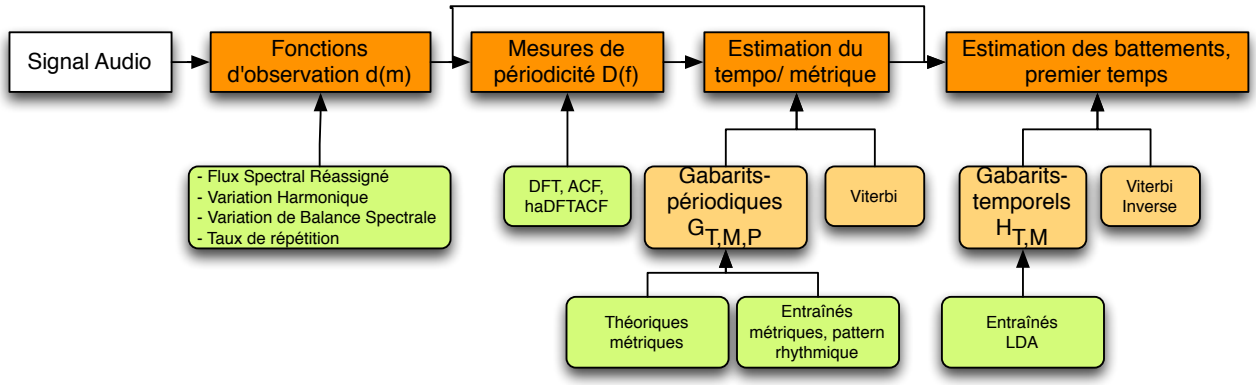

FIGURE 3.3 – Schéma général de l'algorithme d'estimation du tempo, de la métrique, des positions des battements et premiers temps. Les gabarits-périodiques utilisés sont issus des différentes méthodes : théoriques ou appris, relatifs ou absolus.

Lors de nos expériences, cette méthode s'est avérée plus robuste que la méthode « temporelle ». Ceci en particulier pour l'estimation du tempo sur des signaux pour lesquels il n'est pas possible de décider du commencement d'un évènement. Cette approche repose cependant sur une mesure de périodicité et implique donc que l'inter-distance entre battements soit (plus ou moins) périodique. Cette approche montre par exemple ses limites pour l'estimation des battements sur les Mazurka de Chopin (voir Figure 3.2). Cette figure illustre l'aspect très peu périodique de l'inter-distance de ces battements.

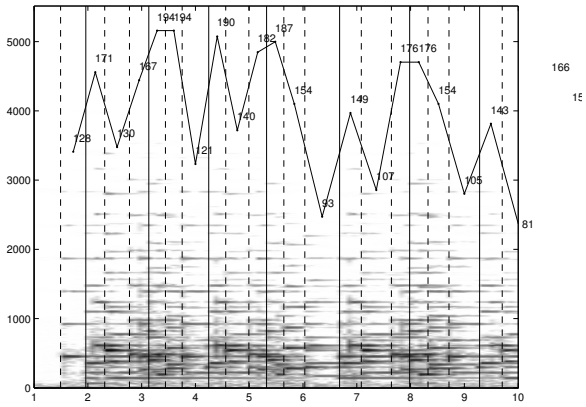

FIGURE 3.2 – Annotations des battements (lignes pointillées), premiers temps (lignes continues), et variations correspondantes de tempo, pour la Mazurka 06-4 de Chopin.

Variabilité du tempo et de la métrique. En dehors des Mazurka de Chopin, il existe bon nombre de musiques de tempo ou de métrique temporellement variables. Pour cela nous avons proposé de « décoder » temporellement le meilleur trajet de tempo et de métrique à l'aide d'un algorithme de Viterbi. Nous avons également proposé le décodage des temps de battements et premiers temps par un algorithme de Viterbi dit « inversé ». Un compromis doit néanmoins être trouvé entre une réactivité importante du système, afin de suivre des variations rapides du tempo (musique classique), et une certaine inertie, afin de permettre la continuation de l'estimation dans des régions sans évènements proéminents (par exemple des breaks en jazz).

Complexité rythmique. Les algorithmes d'estimation obtiennent de meilleures performances sur de la musique populaire ou de la techno. Ceci vient probablement du fait que, pour ces musiques, les battements correspondent aux maxima locaux d'énergie (grosse-caisse, caisse-claire). Lorsque les patterns rythmiques deviennent plus compliqués (jazz, musique cubaine ...) il devient nécessaire d'*apprendre* à l'algorithme leurs caractéristiques. Pour cela, nous avons proposé différentes techniques d'apprentissage des « gabarits-périodiques » (voir partie 3.1.3.3). Ces gabarits-périodiques permettent donc de représenter différentes métriques mais également différents patterns rythmiques.

3.1.3 Contributions

Je représente à la Figure 3.3 l'interconnexion des différents éléments étudiés pendant nos recherches et les détaille ci-dessous.

3.1.3.1 Fonctions d'observation $d^*(m)$

Au cours des années, nous avons étudié quatre fonctions d'observation. Chacune d'elle a pour but la mise en évidence de variations supposées relatives à la perception du tempo, de la métrique, des battements ou des premiers temps. Comme pour les descripteurs instantanés de la partie 2, je les note $d^k(m)$.

$d^e(m)$ **Flux-spectral-réassigné.** Cette fonction vise à mettre en évidence la présence d'onsets à travers les variations de contenu énergétique dans différentes bandes de fréquences. Elle est généralement désignée sous le nom de « flux spectral » [96]. Dans [147] j'ai proposé son calcul sur le spectrogramme « réassigné » [53]. Celui-ci permet d'améliorer la séparation entre bandes fréquentielles adjacentes (et donc de faciliter la mise en évidence des transitions entre notes proches) et la localisation temporelle.

$d^h(m)$ **Variation harmonique.** Cette fonction que nous avons proposée dans [164] et affinée dans [158] vise à mettre en évidence la variation de contenu harmonique du signal. Cette variation est souvent due aux changements d'accords ou de tonalités qui se produisent couramment sur les 1^{ers} temps (1^{ers} et/ou $3^{\text{èmes}}$ pour une métrique 4/4). Dans [164] , nous la calculons par mesure de variation de Chromas ; dans [158] par une mesure de « nouveauté » [56] obtenue par application de noyaux de segmentation sur une matrice d'auto-similarité de Chromas. Nous l'utilisons dans [164] pour la localisation des 1^{ers} temps et utilisons sa périodicité dans [158] comme attribut du tempo perceptif.

$d^b(m)$ **Variation de la balance spectrale.** Cette fonction que nous avons proposée dans [164] vise à mettre en évidence la variation de balance spectrale (rapport entre les contenus énergétiques haute et basse fréquence). Cette variation est souvent due au phénomène « poum-tchak » (alternance grosse-caisse et caisse-claire ou encore alternance de la ligne de basse) présent en musique populaire. Nous utilisons cette fonction dans [164] pour la localisation des 1^{ers} et $3^{\text{ème}}$ temps (correspondant au « poum ») et utilisons sa périodicité dans [158] comme attribut du tempo perceptif.

$d^r(\tau)$ **Taux de répétition.** Cette fonction que nous avons proposée dans [158] vise à mettre en évidence les répétitions à court terme de contenu timbral et harmonique (une séquence d'évènements ou de notes répétées périodiquement). Son calcul repose sur la conversion d'une matrice d'auto-similarité représentant le contenu timbral et harmonique [146] en matrice de décalage. Notons que cette fonction s'exprime directement dans le domaine des décalages τ et non du temps m. Sa périodicité est utilisée dans [158] comme attribut du tempo perceptif.

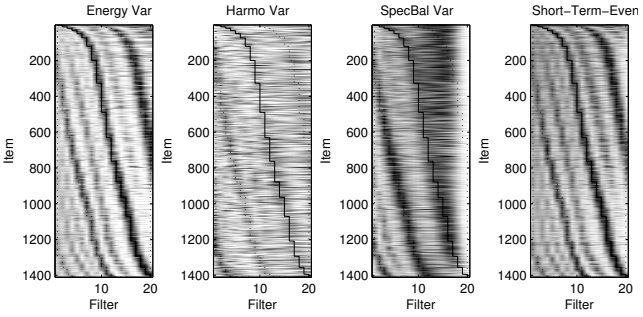

FIGURE 3.4 – Chaque matrice représente une fonction d'observation (de gauche à droite : $d^e(m), d^h(m), d^b(m), d^r(\tau)$). Chaque ligne d'une matrice représente la périodicité (ACF) de la fonction d'observation pour un item donné. Les items sont classés par ordre croissant de tempo.

Pour une tâche d'estimation du premier temps, nous montrons dans [164] que l'utilisation des fonctions $d^h(m)$ et $d^b(m)$ en complément de $d^e(m)$ apporte une amélioration significative des résultats. Pour une tâche d'estimation du tempo perceptif, nous montrons dans [158] que la combinaison de $d^e(m)$, $d^b(mt$ et $d^r(\tau)$ fournit les meilleures prédictions. J'illustre le comportement de ces fonctions à la Figure 3.4 sur le corpus de test `Last-FM` utilisé dans [158].

3.1.3.2 Mesures de périodicité $D^*(f)$

La périodicité de ces fonctions d'observation $d^*(m)$ (voir partie 3.1.3.1) est ensuite calculée. Pour cela, j'ai étudié trois mesures de périodicité. Les deux premières sont la Transformée de Fourier Discrète (**DFT**) et la Fonction d'Auto-Corrélation (**ACF**).

La troisième, proposée dans notre article de journal [151], est appelée « Product Hybrid Axis DFT/ACF » (**haDFTACF**). L'utilisation de cette fonction est motivée par les très bons résultats que nous avons obtenus avec la fonction produit DFT/FM-ACF dans le cadre de l'estimation de hauteurs [143] [6] ou dans le cadre de l'estimation du tempo [147]. Cette fonction résulte de la combinaison par multiplication d'une DFT et d'une ACF exprimée sur un axe hybride. Cette méthode tire parti du fait que la DFT d'un train périodique d'impulsions de période T_0 est une série d'harmoniques aux fréquences $f_h = h/T_0,\ h \in \mathbb{N}^*$ alors que son ACF est une série de périodes aux décalages $\tau_h = hT_0,\ h \in \mathbb{N}^+$. Leurs séries harmoniques se « déroulent » donc dans des directions opposées. En exprimant le domaine des décalages τ dans celui des fréquences $f = 1/\tau$, il est donc possible (théoriquement), par simple multiplication des deux fonctions, de réduire la série à une composante unique $f = 1/T_0$ (dans le cas d'un signal à périodicité unique)

6. Dans l'article [147], je compare la combinaison de représentations fréquentielles —DFT, auto-corrélation de la DFT (ACFofDFT) et auto-corrélation du spectre réassigné (ACFofRES)— et temporelles —auto-corrélation (ACF) et cepstre (CEP)— pour une tâche d'estimation de hauteurs de notes. Sur un corpus de test de 5371 sons d'instruments, les meilleurs résultats sont obtenus avec les combinaisons ACFofDFT/CEP et ACFofREAS/CEP : 97%. Ces résultats constituent une amélioration par rapport à l'algorithme Yin (94.9%).

ou à un ensemble réduit de composantes représentatives des différentes périodicités présentes (dans le cas d'un signal aux périodicités multiples comme le rythme). Dans la pratique, du fait que la discrétisation de l'axe des décalages (déterminé par le taux d'échantillonnage) n'est pas compatible avec celle de l'axe des fréquences (déterminé par la taille de la DFT), des techniques d'interpolation doivent être utilisées. Dans la méthode **DFT/FM-ACF** (Frequency-Mapped ACF), les décalages de l'ACF sont exprimés dans le domaine fréquentiel et pour cela ses valeurs sont interpolées. Dans la méthode **TM-DFT/ACF** (Temporally-Mapped DFT), les fréquences de la DFT sont exprimées dans le domaine temporel. Ces interpolations résultent cependant en une perte d'information (une partie des informations en basses fréquences de l'ACF sont perdues dans la FM-ACF, en bas décalages de la DFT dans la TM-DFT). Nous proposons donc d'effectuer cette interpolation sur un axe hybride (hybride entre discrétisation temporelle et fréquentielle). Cette méthode est appelée « Product Hybrid Axis DFT/ACF » (**haDFTACF**).

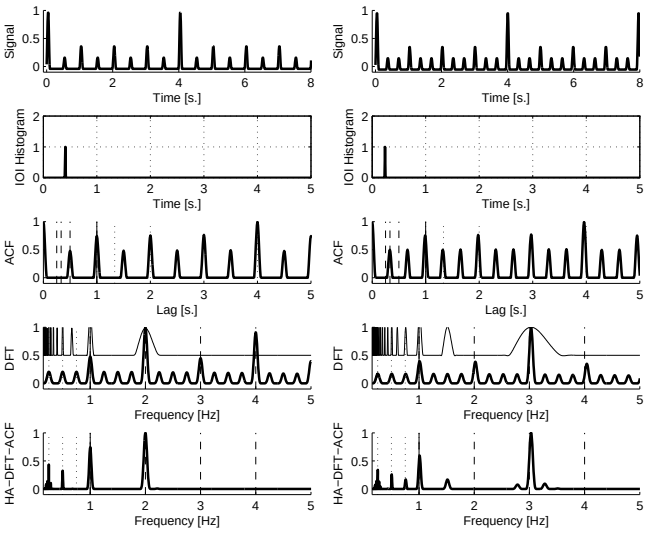

FIGURE 3.5 – Chaque panneau représente de haut en bas : le signal, l'Inter-Onset-Histogram, l'ACF, la DFT (nous lui superposons la FM-ACF) et l'haDFTACF. [Panneau de gauche] métrique quadruple/simple, [Panneau de droitel] métrique quadruple/composée.

Dans [151], nous comparons l'ACF, la DFT et l'haDFTACF pour la représentation du contenu rythmique d'un signal. Nous les comparons au travers de leur sensibilité aux positions relatives des évènements (et donc leur capacité à représenter ceux-ci) ; de leur insensibilité aux décalages intentionnels (tel le swing ou plus généralement le groove) et de la possibilité de les rendre indépendantes du tempo. La Figure 3.5 illustre leur comportement pour des signaux de métriques différentes. Sur cet exemple, l'haDFTACF se montre plus discriminante et compacte que la DFT et l'ACF. Nous montrons cependant que globalement, sans doute du fait de la sensibilité de l'ACF aux décalages intentionnels (qui est donc propagée à l'haDFTACF), la meilleure représentation est la DFT.

3.1.3.3 Utilisation de gabarits-périodiques

Comme indiqué, nos algorithmes d'estimations considèrent simultanément le tempo (T) dans le contexte d'une métrique (M) et/ou d'un pattern rythmique (P). Notre hypothèse est que la périodicité $D^*(f)$ des fonctions d'observation $d^*(m)$ dépend de T, M et P. Dans nos algorithmes cette dépendance est représentée par des gabarits-périodiques. Ils sont notés $G_{T,M,P}$ quand ils représentent les trois concepts T, M et P. L'objectif de ces gabarits est de permettre l'estimation des meilleurs T, M et P expliquant les périodicités de $D(f)$: $(D(f) \leftrightarrow G_{T,M,P}) \rightarrow (T, M, P)$.

Représentations relatives et absolues. La création de gabarits-périodiques $G_{T,M,P}$ est une tâche compliquée manuellement (nombre important de degrés de liberté) et automatiquement (nombre réduit de données pour l'apprentissage de chaque catégorie). Aussi dans beaucoup de nos recherches avons-nous utilisé des gabarits-périodiques indépendants du tempo $G_{M,P}$. Lorsque le tempo n'est pas inclus dans une représentation, nous parlons de **représentation relative**. Elle est « relative » à un tempo. A l'inverse, nous désignons par **représentations absolues** celles l'intégrant. $D(f)$ est une représentation absolue puisqu'elle contient l'information de tempo. Le passage d'une représentation relative à une représentation absolue est fait par l'ajout d'une information de tempo. Ainsi, si le gabarit-périodique $G_{M,P}$ est relatif, une information de tempo devra être ajoutée pour pouvoir le comparer à $D(f)$: $(G_{M,P} + T) \rightarrow G_{T,M,P} \leftrightarrow D(f)$. Une autre possibilité est de créer une représentation de $D(f)$ relative au tempo T, notée $D_T(k)$: $G_{M,P} \leftrightarrow D_T(k)$.

La représentation relative choisie consiste à représenter les périodicités attendues aux subdivisions et multiples du tempo. Plus précisément, nous utilisons le fait que le spectre d'un signal périodique prend la forme d'un peigne harmonique dont la fréquence fondamentale f_0 correspond à la périodicité la plus basse (souvent la périodicité de la mesure dans le cas du rythme) et dont les amplitudes a_h aux différentes harmoniques hf_0 dépendent de la structure rythmique M et P. De même, nous faisons l'hypothèse que seules ces fréquences harmoniques contiennent une information relative au rythme. Cette approche est donc similaire à celle utilisée dans le « cepstre discret » [60] qui considère également une discrétisation du spectre aux fréquences harmoniques. La représentation relative est un vecteur formé par l'échantillonnage de $D(f)$ aux fréquences correspondant à l'union des harmoniques d'une métrique 4/4 et 3/3 : $f = T \cdot [\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, ...4]$, dans lequel T est le tempo. Nous

montrons dans [151] que ces fréquences correspondent à la position des pics de la DFT, même en présence de swing. Pour un échantillonnage reposant sur T, cette représentation relative est notée $D_T(k)$ ou $G_{M,P}$.

Gabarits-périodiques théoriques et entraînés. Ces gabarits-périodiques peuvent être créés **théoriquement** par des Humains G^H, ou être Appris G^A à l'aide de techniques d'apprentissage machine.

Nos différentes études sur le rythme (dans le cadre de l'estimation du tempo ou de la classification de rythme) se distinguent par ce que représentent ces gabarits-périodiques (en terme d'observation – DFT, ACF ou haDFTACF –, en terme de dépendance vis-à-vis de T, M ou P) et dans la manière dont ces gabarits-périodiques sont créés (théoriquement ou par apprentissage). Notons que dans ces études ces « gabarits-périodiques » peuvent revêtir la forme d'un vecteur de valeurs ou de paramètres d'un algorithme de classification (modèle Gaussien, GMM ou SVM).

Gabarits-périodiques pour l'estimation du tempo. Dans notre première recherche [147], l'estimation du tempo et de la métrique repose sur l'utilisation de gabarits-périodiques relatifs de type G^H_M (représentant uniquement les métriques). Ceux-ci ont été créés théoriquement à partir de l'observation du comportement de la fonction DFT/FM-ACF sur des signaux de métriques 2/2, 3/4, 6/8 et 9/8 (voir Figure 3.6). Ces gabarits-périodiques sont ensuite comparés aux vecteurs $D_T(k)$ pour différentes hypothèses de tempo T. En pratique, afin de permettre l'estimation de tempi et de métriques variables au cours du temps, un modèle de Markov caché est utilisé. Dans celui-ci, les état cachés correspondent aux couples (T, M) et les observations aux $D(f)$ au cours du temps. Le décodage temporel est ensuite effectué par un algorithme de type **Viterbi**. Les évaluations de cette méthode, proposée dans [147], montrent des performances supérieures à l'état de l'art pour les corpora de test `Ballroom` (en Accuracy1 et Accuracy2) et `Loops` (en Accuracy2). Cette méthode, évaluée lors de la campagne `MIREX` 2005 (nous n'avons pas participé les années suivantes), a obtenu la première place dans la catégorie « At Least One Correct Tempo » [7].

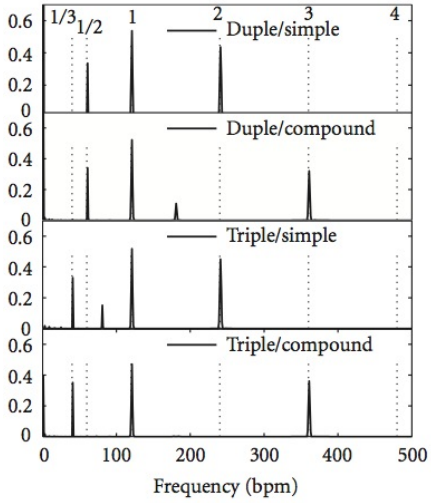

FIGURE 3.6 – Gabarits-périodiques théoriques correspondant à différentes métriques dans le cas de la TFD/FM-ACF.

Gabarits-périodiques pour la classification. Mes recherches sur la classification automatique des rythmes, publiées dans l'article de journal [151], reposent sur l'entraînement de gabarits-périodiques relatifs de type $G^A_{M,P}$ (représentant simultanément la métrique et le pattern rythmique). Dans cet article je montre qu'une représentation relative permet une bonne représentation des patterns rythmiques. A titre d'exemple je représente à la Figure 3.7 les vecteurs de périodicités $D(f/T)$ (partie de gauche) et $D_T(k)$ (partie de droite) correspondant à tous les morceaux de chacune des 8 catégories de rythme du corpus de test `Ballroom`. Chacune des catégories est une combinaison de M et P. La figure montre bien les patterns distinctifs de $D_T(k)$ en fonction de M et P. L'abscisse représente les fréquences normalisées par rapport au tempo (1), l'ordonnée les différents items appartenant à une classe. Je compare dans cet article l'utilisation de différentes mesures de périodicité (ACF, DFT, haDFTACF) pour le calcul de $D_T(k)$ à travers une tâche de reconnaissance de classes de rythmes. A cet effet, des modèles $G^A_{M,P}$ sont appris à l'aide de différents algorithmes de classification (J48, PART, ClassViaReg, AdaBoost, Random Forest, SVM). Dans le cas où le tempo T est connu (utilisé pour la création de $D_T(k)$), la TFD permet une classification correcte à 93.4% ; dans le cas où T est estimé elle est de 80%. Je propose également dans [151], un ensemble de descripteurs audio propres au rythme : percussivité (mesure du taux de décroissance moyen des évènements audio), périodicité (mesure du taux d'énergie du signal expliqué par des composantes périodiques), vitesse (mesure du centre de gravité des fréquences). A l'aide de ces descripteurs, les performances de reconnaissance montent jusqu'à 95.6% (T connu), 86.5% (T estimé).

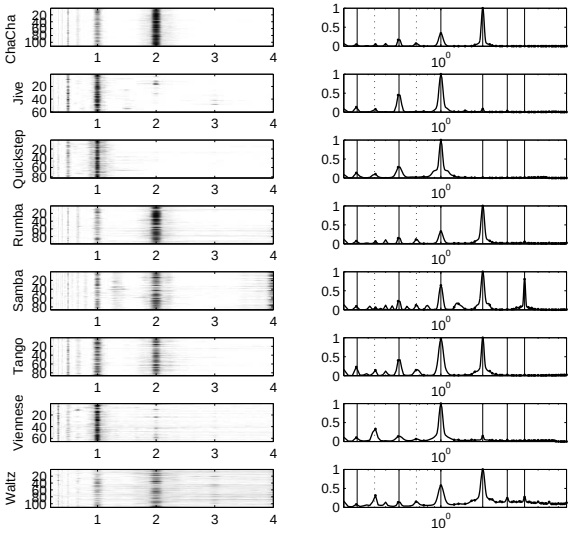

FIGURE 3.7 – Voir texte

7. Les meilleurs résultats globaux ont été obtenus par Alonso [5].

Gabarits-périodiques pour l'estimation de tempo. Dans [148], je réétudie l'apprentissage de gabarits-périodiques relatifs de type $G^A_{M,P}$ (représentant simultanément la métrique et le pattern rythmique) mais pour l'estimation du tempo cette fois. Alors que dans [151], un apprentissage supervisé était utilisé (nécessitant la définition de classes de rythme et leur annotation), je cherche ici à valider l'utilisation d'apprentissages non-supervisés pour les $G^A_{M,P}$. L'algorithme utilisé est de type « K-means ». Pour une tâche de reconnaissance du tempo exact (Accuracy-1), je compare les résultats obtenus avec notre technique initiale (les G^H_M de [147] sont théoriques et représentent uniquement la métrique) à ceux obtenus avec les $G^A_{M,P}$ appris de manière non-supervisée et supervisée (par modèle Gaussien). Les résultats sont respectivement 44%, 72.9% et 75.2%. Les meilleurs résultats sont donc obtenus en apprentissage supervisé. Ceux obtenus en apprentissage non-supervisé sont malgré tout très bons. Cette étude montre que l'utilisation de gabarits-périodiques représentant simultanément la métrique et le pattern rythmique permet une meilleure estimation du tempo. Lors de cette étude je montre également l'influence néfaste que peut avoir l'utilisation d'une probabilité a priori fixe sur l'estimation du tempo, et dès lors je mets indirectement en évidence le fait que les patterns rythmiques sont préférentiellement joués à certains tempi.

Gabarits-périodiques pour l'estimation du tempo perceptif. Ce dernier résultat est mis à profit dans nos dernières études par l'apprentissage de gabarits-périodiques absolus $G^A_{T,M,P}$ (donc modélisant la relation existante entre le tempo, la métrique et le pattern rythmique). Dans le stage de Master de Joachim Flocon-Cholet [54] et dans l'article [158] nous étudions l'estimation du tempo perceptif. La publication des résultats de l'expérience faite à `Last-FM` en 2011 [98], relative à la perception du tempo, nous a rendu possible cette étude. Dans cette expérience, 4000 morceaux ont été annotés en tempo par un large panel d'utilisateurs. Pour 1500 de ces 4000 morceaux, la majorité des utilisateurs s'est mise d'accord sur la perception d'un tempo unique. Nous étudions donc la perception de ces tempi. A la différence de nos autres études sur les gabarits-périodiques (utilisant uniquement la fonction d'observation $d^e(m)$), nous utilisons ici les quatre fonctions d'observation. Nous considérons que la perception du tempo est potentiellement liée à ces quatre indices. Comme l'apprentissage des gabarits-périodiques $G^A_{T,M,P}$ nécessiterait un nombre considérable de données (du fait du nombre important de degrés de liberté), nous optons pour l'utilisation d'une technique de régression sur des modèles de mélange Gaussien (modèle GMR de [46]) appris sur les représentations « absolues » de périodicité $D(f)$). Afin de réduire la dépendance du modèle envers le tempo, $D(f)$ est convolué par un banc de filtres. Le modèle GMM utilisé pour la régression représente l'ensemble de la structure rythmique (tempo, métrique, pattern rythmique) et est donc du type $G^A_{M,P,T}$. Les résultats obtenus à l'aide de cette méthode montrent une large augmentation de la qualité de la prédiction du tempo obtenue : 67.3% pour [147] contre 72.9% pour la méthode GMR (sur le corpus de test `Last-FM`), 66.1% pour [147] contre 87% pour la méthode GMR (sur le corpus de test `Ballroom`).

Gabarits-périodiques complexes pour la méthode « copy and scale ». Dans les méthodes précédentes, nous faisions l'hypothèse que $D(f)$ dépend de T, M et P. Connaissant cette dépendance (modélisée par des gabarits-périodiques), nous pouvions ainsi déduire le tempo de l'observation de $D(f)$. Dans la méthode « copy and scale », je fais l'hypothèse que la position (la localisation temporelle) des battements dépend de T, M et P. Les gabarits-périodiques $G^A_{M,P}$ permettant de représenter ces patterns rythmiques, je cherche à les étendre de manière à pouvoir également représenter la localisation temporelle des évènements. L'objectif est d'attacher à ces nouveaux gabarits-périodiques l'information de positions des battements propre à chaque pattern rythmique ; et de rechercher ensuite, pour un signal inconnu de périodicité $D(f)$, le gabarit-périodique le plus proche et de recopier ses marqueurs attachés.

Je propose pour cela, dans l'article de journal [149], l'utilisation de gabarits-périodiques dans le domaine complexe : $G^{*A}_{M,P} \in \mathbb{C}$. Ces gabarits complexes $G^{*A}_{M,P}$ permettent de mieux représenter la localisation temporelle des événements que $G^A_{M,P}$, comme l'attestent leurs erreurs respectives de reconstruction de $d(m)$ [8]. La méthode proposée repose ensuite sur un simple 1-PPV (plus proche voisin). Il n'y a donc pas d'entraînement ni de modèle ; la connaissance du placement des battements spécifique à chaque T, M et P est introduite simplement sous forme d'exemples. Pour un ensemble de morceaux annotés en position des battements, leurs gabarits-périodiques $G^{*A}_{M,P}$ sont calculés et stockés dans une base de données. Pour un signal inconnu, le gabarit-périodique le plus proche est recherché (utilisation de distance dans le domaine complexe) et les positions de battements attachées à ce gabarit-périodique sont recopiées (partie **copy**). Telle quelle la méthode nécessiterait cependant la création d'une base de données très grande, aussi la représentation est-elle rendue indépendante du tempo (les $G^{*A}_{M,P}$ sont indépendants du tempo, et peuvent donc représenter des items de tempi différents), et indépendante de la position relative des battements (relative par rapport au début du signal observé). Ces deux ajouts nécessitent la partie **scale** de l'algorithme décrite dans l'article. Dans celui-ci, j'étudie également certaines optimisations permettant de rendre la recherche par 1-PPV plus rapide. L'estimation de la position des battements fournie par la méthode « copy and scale » est comparée à celle d'`ircambeat` (combinaison de l'algorithme d'estimation

8. L'erreur de reconstruction de $o(m)$ est définie par $\epsilon = \sqrt{\sum_m (o(m) - \hat{o}(m))^2}$ dans laquelle $\hat{o}(m)$ est obtenue par synthèse sinusoïdale à partir d'un gabarit. Les erreurs de reconstruction obtenues par les gabarits-périodiques $G^A_{M,P}$ et $G^{*A}_{M,P}$ sont comparées sur une base de 600 morceaux. L'erreur moyenne de $G^A_{M,P}$ est de 1.62, celle de $G^{*A}_{M,P}$ de 0.41.

du tempo [147] et de localisation des battements [164]) sur le corpus `Ballroom`. Pour l'ensemble du corpus, cette méthode conduit à une F-measure moyenne de 68.2% (78.4% pour `ircambeat`), Pour le sous-ensemble d'items dont la classe de rythme a été correctement identifiée, elle conduit à 81.3% (80.4% pour `ircambeat`). Ces résultats montrent donc le potentiel d'une telle méthode qui ne repose pourtant sur aucune notion de tempo ou de battement.

3.1.3.4 Estimation de la position des battements et premiers temps

La méthode que nous avons proposée dans l'article de journal [164] pour l'estimation de la position des battements et des premiers temps repose sur l'utilisation d'un algorithme de **Viterbi dit « inversé »**.

Pour cela, nous définissons comme dans [132] les « beat-position-inside-a-bar » (bpib) comme les positions relatives des battements par rapport à leur premier temps.Nous cherchons à estimer simultanément la position des battements et de leur indice bpib. Sans connaissance a priori, n'importe quel instant d'un morceau peut être un battement. Nous définissons de ce fait les *états cachés* d'un modèle de Markov comme un temps t spécifique dans un bpib spécifique. Nous cherchons ensuite à décoder le chemin à travers les états cachés qui expliquent le mieux les observations $d^*(m)$ du signal. Comme le temps est la variable cachée, nous parlons de Viterbi « inversé » (l'algorithme de Viterbi étant généralement utilisé pour décoder les états à travers le temps et non pour décoder le temps lui-même). Le décodage fournit ensuite les états c'est-à-dire les temps de battements et leur bpib associé, c'est-à-dire les premiers temps.

Les *probabilités de transition* entre états de ce HMM tiennent compte du fait que les états doivent avoir une inter-distance inverse du tempo locale estimé, et que les index bpib de ces états doivent suivre la permutation circulaire propre à la métrique estimée.

Les *probabilités d'émission* sont obtenues en utilisant plusieurs critères. Le premier est relatif à la corrélation entre un « gabarit-temporel » $H_{T,M}$ et la fonction de flux-spectral-réassigné $d^e(m)$. Le second est un critère de positionnement dans un pattern harmonique dérivé de $d^h(m)$ et dans un pattern rythmique dérivé de $d^b(m)$. Nous proposons également une technique d'apprentissage des « gabarits-temporels » $H_{T,M}$ permettant d'accroitre la discrimination entre les valeurs de la corrélation à la position des battements et en dehors. Pour cela, chaque échantillon de $d^e(m)$ à l'intérieur d'une mesure est considéré comme une dimension d'un espace de description. Sur un corpus d'entraînement, nous assignons aux différents m de $d^e(m)$ les labels « battement » ou « non-battement ». L'Analyse Linéaire Discriminante (LDA) permet ensuite de trouver le meilleur hyper-plan de séparation entre ces deux classes. La matrice de transformation de cette analyse sert ensuite à la création des gabarits-temporels $H_{T,M}$. Dans [164], nous montrons que ces gabarits-temporels permettent une meilleure localisation des battements que les gabarits-temporels habituels (comme ceux de [96]).

Dans [164], nous comparons les résultats obtenus avec notre algorithme à ceux de l'état de l'art. Pour une tâche de détection de battements, il n'y a pas d'amélioration par rapport à l'état de l'art. A l'"inverse, pour une tâche d'estimation des premiers temps, il y a une nette amélioration : 46% pour l'algorithme de Klapuri [88] contre 61% pour le notre sur le corpus de test `Klapuri` pour le critère CMLc [9]. Notre algorithme d'estimation des battements a également fait l'objet de nombreuses évaluations lors des campagnes `MIREX`. Sur le corpus de test `McKinney`, il a obtenu la première place en 2009, la deuxième en 2010 et 2011. `MIREX` n'a cependant pas de tâche concernant l'estimation des premiers temps.

3.1.4 Applications

Les algorithmes d'estimation du tempo [147] et de la position des battements et premiers temps [164] ont fait l'objet du développement du logiciel et de la librairie C++ `ircambeat` développés par Patrice Tisserand et depuis plusieurs années par Frédéric Cornu. L'optimisation réalisée a permis d'atteindre des performances de l'ordre de 4 s pour le traitement de 4 min d'audio. Cette librairie permet les estimations « beat-sychrones » des logiciels `ircamchord` et `ircamsummary`. Elle est également maintenant intégrée dans le logiciel `Audiosculpt 3.0` de l'IRCAM grâce au travail de Charles Picasso. J'illustre à la Figure 3.8, son interface graphique dans cette application.

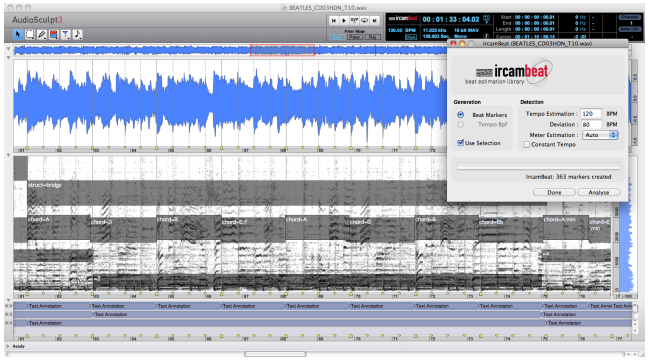

FIGURE 3.8 – `ircambeat` dans l'application `Audiosculpt 3.0`.

9. Continuity required at the correct Metrical level [71]

3.2 Estimation de paramètres relatifs au contenu harmonique

Dans cette partie je décris les recherches que j'ai effectuées ou encadrées concernant la description du contenu harmonique d'un morceau de musique. Ces recherches visent à obtenir une description de haut niveau de ce contenu ; aussi n'étudions-nous pas l'estimation des notes [10] mais d'une abstraction de ces notes : la tonalité (globale ou locale en temps) et la suite d'accords. Pour ces recherches, nous nous sommes appuyées sur des représentations de type Chroma et non sur l'estimation des hauteurs de notes [11]. Ces recherches se sont effectuées de 2006 à 2010 dans le cadre du projet `SemanticHIFI` et `Ecoute`, du stage de Master et de la thèse d'Hélène Papadopoulos. Ces recherches sont poursuivies aujourd'hui dans le cadre du projet `Quaero` par Johan Pauwels sous ma direction, dans l'objectif d'établir une estimation jointe [accords, structure]. Elles ont donné lieu à la publication de 7 articles (dont 2 de journaux). J'ai choisi de présenter en annexe notre article [132] qui correspond aux algorithmes d'estimation sous-jacents au logiciel `ircamchord`.

3.2.1 État de l'art

Je ne reprends ici qu'un bref état de l'art et renvoie le lecteur à notre publication [132] pour un état de l'art complet. Les estimations de tonalités et de suite d'accords sont deux sujets étroitement liés. Dans ce domaine, les algorithmes reposent pour la plupart sur l'extraction de Chroma ou d'Harmonic Pitch Class Profiles [62], incluant éventuellement une séparation harmonique (« pitch-enhancement ») ou un blanchissement du spectre [120]. Un modèle (issu d'expériences perceptives, entraîné sur un corpus ou reposant sur la théorie musicale) est ensuite utilisé pour établir un lien entre observations et labels de tonalité ou d'accords. Dans ce domaine, la modélisation des dépendances entre les différents paramètres musicaux (à l'aide de modèles de Markov ou de réseaux Bayésiens) est une pratique courante : dépendance entre accords et tonalité [135] [188], entre accords successifs, entre accords, position métrique [132] et notes de la basse [107].

3.2.2 Contributions

3.2.2.1 Estimation de tonalités globales

Cette recherche vise à estimer la tonalité globale d'un morceau de musique (parmi l'ensemble des 24 tonalités majeures et mineures [12]) à partir de l'analyse de son signal audio. Dans les articles [144] et [142] nous comparons plusieurs méthodes pour cela reposant cependant toutes sur l'utilisation d'une représentation de type Chroma. Cette représentation est ensuite comparée à des modèles G_{key} représentant les 24 tonalités.

Chromas. Les Pitch-Class-Profiles [59] (également appelés Chromas [215] dans le cas d'une utilisation de transformée à Q-constant – CQT – [24]) visent à résumer le contenu harmonique local d'un signal par sommation des énergies présentes dans un ensemble de filtres centrés sur des hauteurs de notes de « chroma » équivalent. Il s'agit d'une représentation vectorielle très compacte du contenu harmonique. Cette représentation est extraite au cours du temps par analyse à fenêtre « glissante », il s'agit donc d'un descripteur audio instantané (voir partie 2.2). Du fait de la définition des filtres (devant séparer au minimum les notes adjacentes), l'extraction des Chromas nécessite de considérer les propriétés de résolution fréquentielle relatives aux choix des fenêtres d'analyse. Pour cette raison la transformée à Q-constant est souvent utilisée. A la différence des hauteurs de notes multiples qui sont issues d'algorithmes d'estimation, les Chromas ne sont pas issus d'un estimateur mais d'une cascade d'opérateurs mathématiques. Il n'y a donc pas d'appariement effectué entre l'énergie présente à une fréquence fondamentale f_0 et celles présentes à ses harmoniques hf_0. Ceci induit la présence de composantes « parasites » dans le vecteur de Chroma pouvant amener à des confusions de hauteur. Face à ces composantes « parasites » deux stratégies sont utilisées :

Méthode 1. Réduire l'importance des harmoniques supérieures dans la représentation spectrale (et donc réduire l'importance des composantes « parasites »),

Méthode 2. Prendre en compte la présence de ces composantes « parasites » dans les modèles G_{key} représentant les tonalités.

10. Comme indiqué, l'estimation des hauteurs de notes (EHN) fait l'objet d'autres études dans l'équipe Analyse/Synthèse des sons de l'IRCAM (voir par exemple [222]).

11. Dans nos recherches, nous avons essayé de nous appuyer sur le résultat de ces études sur l'EHN. Cependant, du fait de l'état d'avancement des études sur l'EHN au démarrage de nos recherches, leur utilisation ne s'est pas avérée convaincante pour l'estimation des attributs d'accords et tonalités (voir à ce propos l'étude faite dans [130]). Ceci provient sans doute, comme pour l'estimation de tempo reposant sur une détection d'onsets, de l'impact des faux positifs et fausses réjections de l'EHN sur un système aval.

12. Il s'agit donc des tonalités et non des modes que nous avons étudiés plus tard.

De même, les Chromas étant extraits directement d'une transformée spectrale, ils représentent l'intégralité du spectre, donc également les transitoires et le bruit.

Contributions concernant l'extraction des Chromas. Afin de limiter l'influence des transitoires et du bruit dans le calcul des Chromas, nous proposons dans [144] et [142] une méthode de calcul reposant sur un filtrage médian temporel du contenu des filtres (méthode proche conceptuellement de celle de [52]). Afin de limiter l'influence de la présence des harmoniques supérieures dans ce calcul, nous proposons dans [142] la technique appelée « Harmonic Peak Subtraction » (HPS). Celle-ci reprend le principe sous-jacent aux méthodes DFT/FM-ACF ou haDFTACF (voir partie 3.1.3.2) pour l'étendre aux signaux de périodicités multiples. Le principe sous-jacent à la méthode DFT/FM-ACF est la combinaison de la DFT (notée $A(f_k)$) et de l'ACF exprimée dans le domaine fréquentiel (notée $g_\tau(f_k) = g_\tau^+(f_k) - g_\tau^-(f_k)$). Dans la méthode HPS, nous testons l'hypothèse que chaque fréquence f_k est une hauteur (H1) contre l'hypothèse que f_k est une harmonique supérieure (H2) ou une sous-harmonique d'une autre hauteur (H3). Les vraisemblances de H1, H2, H3 sont ensuite données par la combinaison des projections de $A(f_k)$ sur $g_\tau^+(f_k)$ et $g_\tau^-(f_k)$. La représentation résultante permet de mieux discriminer les fréquences des hauteurs de celles correspondant aux harmoniques supérieures ou inférieures. La Figure 3.9 illustre l'application de cet algorithme. Les Chromas peuvent donc être extraits d'une représentation de type DFT ou du résultat de l'HPS. Dans [142] nous étudions également l'influence de l'échelle des valeurs de ces représentations : échelle d'amplitude, d'énergie et de Sone (inspirée des propositions de [127] faites dans le cas de la similarité musicale). Dans [142], les meilleures estimations de tonalité sont obtenues à l'aide du HPS et de cette échelle de Sone, les différences ne sont cependant pas significatives. Dans la suite, nous notons $\mathbf{d}_m$ ces vecteurs de Chroma au cours des trames m.

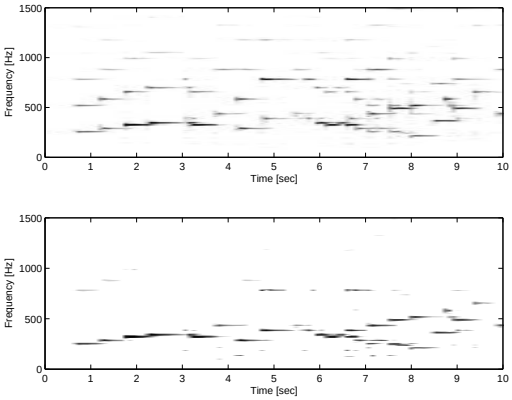

FIGURE 3.9 – [Haut] Spectrogramme [Bas] Algorithme HPS appliqué au spectogramme sur les 10 premières secondes du morceau « Le clavier bien tempéré, 02 Fugue en do majeur » de J.S. Bach.

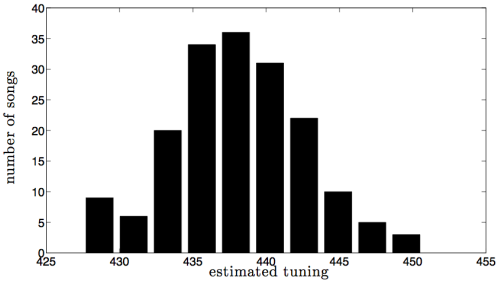

FIGURE 3.10 – Histogramme de l'accordage estimé sur tous les morceaux du corpus de test « Beatles », d'après [130]

Estimation de l'accordage. Les vecteurs de Chroma $\mathbf{d}_m$ sont utilisés par comparaison à des modèles de tonalités G_{key}. Ces derniers correspondent à un accordage de 440 Hz. Cet accordage peut cependant être très différent pour les morceaux observés comme l'illustre la Figure 3.10. Pour cette raison, nous proposons dans [144] un algorithme d'estimation de l'accordage d'un morceau. Celui-ci repose sur la minimisation d'une erreur quadratique moyenne de modélisation de l'énergie du spectre obtenue par un ensemble de peignes harmoniques correspondants à différents candidats de fréquence d'accordage. Cet algorithme a fait l'objet du logiciel `ircamtuning` licencié auprès de la compagnie MakeMusic. Cette estimation sert ensuite à « accorder » les vecteurs $\mathbf{d}_m$.

Estimation de tonalités globales. Dans le cas de l'utilisation de la Méthode 2, la présence des composantes « parasites » dans le vecteur $\mathbf{d}_m$ (dues à la présence d'harmoniques supérieures) est prise en compte au niveau des modèles G_{key}.

Lorsque les modèles sont établis **manuellement** (sur base de la théorie musicale ou sur base de résultats d'expériences perceptives), cette prise en compte doit être effectuée de manière explicite. Les gabarits G_{key} sont des vecteurs représentant l'importance des différentes hauteurs dans une tonalité. La contribution des « valeurs » attendues aux différentes harmoniques de ces hauteurs est donc introduites dans G_{key}. Ces « valeurs » peuvent être apprises sur une base d'entraînement. Izmirli [79] propose cela dans le cas du piano. Lors de nos expériences, nous ne sommes pas parvenu à des résultats concluants du fait de la trop grande variabilité de ces valeurs (intra-instrument et à travers le temps pour un même son, voir à ce propos [27]). Nous avons donc choisi l'utilisation du modèle théorique de contribution des harmoniques proposé par Gomez [62]. Lorsque les modèles sont appris **automatiquement** (à partir de l'observation d'exemples de Chromas appartenant à une tonalité) la contribution des harmoniques est prise en compte implicitement.

Dans [144], nous comparons la création manuelle et automatique des modèles G_{key}. La création manuelle repose sur l'utilisation des résultats d'expériences perceptives de Krumhansl & Schmukler [92]. Celles-ci déterminent les hauteurs devant être représentées dans les modèles G_{key} ainsi que leur poids. Ces modèles sont ensuite étendus afin de prendre en compte la présence des harmoniques supérieures de chaque hauteur. Nous avons proposé une nouvelle méthode pour l'apprentissage automatique des modèles G_{key}. Celle-ci s'inspire de techniques

utilisées en reconnaissance de la parole, consistant à entraîner un modèle de Markov caché pour chaque tonalité. Nous obtenons ainsi 24 modèles. Les états internes de chaque modèle représentent les variations locales pouvant exister dans cette tonalité. Nous avons en particulier proposé une méthode permettant de pallier le manque de données d'entraînement pour certaines tonalités. Celle-ci repose sur un recentrage des Chromas sur deux tonalités de référence, un apprentissage de deux modèles (majeur et mineur), et un décentrage afin d'obtenir les 24 modèles. Un résultat remarquable de cette étude réside dans le fait que certains des vecteurs moyenne du HMM obtenus sont très corrélés aux profils cognitifs de Krumhansl & Schmukler. Ceci est illustré à la Figure 3.11. Dans [144], nous montrons cependant que, pour une tâche de prédiction de tonalité globale, l'utilisation de modèles G_{key} manuels, issus d'expériences perceptives, surpasse celle des HMMs appris.

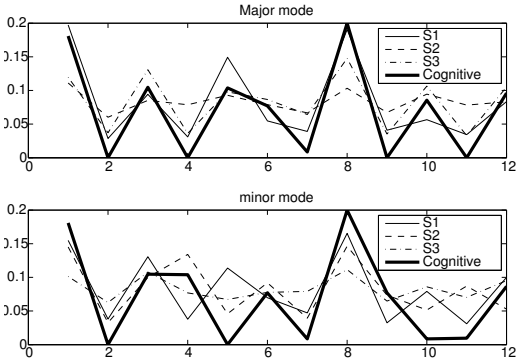

FIGURE 3.11 – Comparaison entre les profils cognitifs de Krumhansl & Schmukler et ceux obtenus par apprentissage HMMs (S1, S2, S3) pour les tonalités do majeur [Haut] et do mineur [Bas].

L'algorithme d'estimation de la tonalité globale, reposant sur l'utilisation directe des Chromas en échelle de Sone et l'utilisation des profils cognitifs, a fait l'objet du développement du logiciel C++ `ircamkeymode` par Patrice Tisserand dans le cadre du projet `SemanticHIFI`. Ce logiciel a obtenu de très bonnes performances aux évaluations `MIREX`.

3.2.2.2 Estimation jointe [tonalités, accords, premiers temps]

La recherche sur l'estimation de tonalités globales est poursuivie par une recherche plus générale sur l'estimation de suites d'accords et de tonalités variables au cours du temps, à travers le stage de Master [129] et la thèse de doctorat [130] d'Hélène Papadopoulos que j'ai encadrés. Ces travaux ont en commun l'utilisation des modèles de Markov cachés. Afin de mieux pallier la décroissance exponentielle des durées d'auto-transition des états, les observations sont extraites de manière synchrone aux battements (analyse dite « beat-synchrone »). Dans ces travaux, une étude assez exhaustive de l'utilisation des modèles de Markov cachés pour l'estimation d'accords est effectuée. Ces recherches démarrent de manière concomitante à la diffusion du corpus de test « Beatles » de [74]. Notre article [131] constitue sans doute la première étude de reconnaissance d'accords sur un corpus important. Je résume ici les principaux apports de cette recherche et renvoie le lecteur à notre publication [132] jointe en annexe pour plus de détails.

Chroma. Une étude détaillée des différents modes d'extraction des Chromas est proposée (à partir de la DFT, de la CQT, à partir des résultats d'un algorithme d'estimation de hauteurs de notes multiples [222]. Cette étude conclut en une meilleure représentation du contenu harmonique (en particulier des basses) et une meilleure robustesse de la CQT.

Analyse « beat-sychrone ». Une étude détaillée des analyses « beat-sychrones » (BS) est proposée, distinguant l'extraction des Chromas de manière BS de leur intégration temporelle de manière BS (voir partie 2.2.1.3). Dans les deux cas, l'analyse peut correspondre à un positionnement BS ou une durée BS. La synchronisation peut s'effectuer au niveau du tactus mais également au niveau du tatum (voir annexe). Cette étude met en évidence la sensibilité de l'analyse « beat-synchrone » à la qualité de l'estimation des battements, et la dépendance des paramètres d'une analyse BS au style musical considéré.

Probabilités d'émission du HMM. Plusieurs méthodes sont étudiées pour estimer les probabilités d'émission du HMM : approche par corrélation entre gabarits d'accords G_{chord} et vecteurs de Chroma et approche par modélisation Gaussienne des Chromas de chaque accord (apprise sur un corpus d'entraînement ou dérivée de la théorie musicale avec prise en compte des harmoniques supérieures).

Probabilités de transition du HMM. Plusieurs méthodes sont également étudiées pour calculer la matrice de transition du HMM : utilisation des proximités/transitions entre accords dans le « cercle des quintes », utilisation des corrélations entre les profils cognitifs de Krumhansl et utilisation de techniques d'entraînement (sur des signaux audio, sur des données transcrites).

Estimation jointe [accords, premiers temps] et modèle HMM à états doubles. Une des contributions importantes de cette recherche est la proposition d'estimation jointe [accords, premiers temps]. Cette estimation repose sur l'hypothèse que les accords changent préférentiellement sur les 1^{er} temps (1^{er} et/ou $3^{ème}$ pour une métrique 4/4) et que cette information peut donc être introduite dans les probabilités de transition. Pour cela un modèle de Markov à états doubles (représentant simultanément l'accord et sa position dans la mesure) est proposé. La matrice de transition entre états doubles représente simultanément les spécificités des transitions entre chaque accord et entre leur position dans la mesure. J'illustre à la Figure 3.12 cette matrice

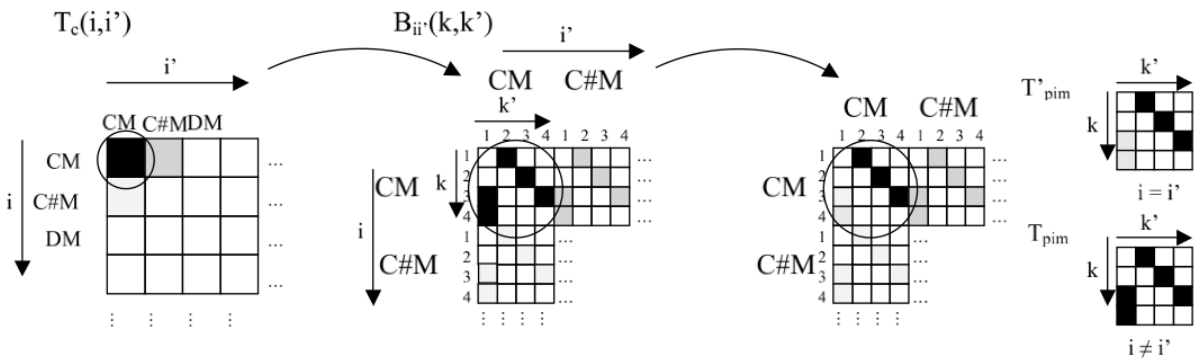

FIGURE 3.12 – Matrice de transition du modèle de Markov à états doubles (accords, positions dans la mesure) proposé par [132].

de transition. Ses différents sous-éléments représentent le même accord dans différentes positions à l'intérieur de la mesure. Ce modèle est étendu pour inclure les changements temporels de métrique (les métriques 3/4 et 4/4 sont considérées dans l'étude).

Algorithmes en cascade et prise en compte de propagation d'erreurs. Le modèle HMM est modifié afin de tenir compte d'erreurs potentielles dans l'estimation des battements utilisés pour l'analyse « beat-synchrone ». La matrice de transition est modifiée de manière à ne pas propager ces erreurs lors du décodage.

Estimation jointe [tonalités, accords, premier temps]. A la différence des méthodes étudiées dans la partie 3.2.2.1 (reposant sur analyse directe des Chromas), nous utilisons ici l'interdépendance existant entre accords et tonalités. L'estimation jointe [tonalités, accords, premier temps] est obtenue par l'utilisation de matrices de transition spécifiques à chaque tonalité. La matrice de transition entre états doubles présentée ci-dessus est démultipliée en 24 matrices représentant les transitions d'accords spécifiques à chaque tonalité. Ceci conduit à 24 HMMs. Le HMM permettant le meilleur décodage du signal fournit la tonalité du morceau, le décodage de ses états fournit les accords et les premiers temps.

Estimation de tonalité locale. Finalement dans notre article de journal [133], nous étudions l'estimation de la tonalité variable en temps. Pour cela, une représentation appelée « chordgram » est proposée. Celle-ci représente, sur des fenêtres longues, le contenu en « accords » du morceau. Une étude des tailles optimales de ces fenêtres longues est proposée. Deux méthodes sont ensuite comparées : soit le chordgram représente les probabilités de chaque accord, soit il représente le résultat du décodage Viterbi de la succession d'accords. Le chordgram sert ensuite d'observation pour un HMM de niveau supérieur. Dans celui-ci, chaque état caché représente une tonalité observée au travers du chordgram. La matrice de transition de ce HMM repose sur des règles de modulations, dérivées des expériences de Krumhansl. Le décodage de ce HMM fournit la suite de tonalités au cours du temps expliquant le mieux le chordgram observé. A la différence du modèle précédent, ce modèle n'effectue pas l'estimation jointe [tonalités, accords, premiers temps] mais permet une estimation de tonalité temporellement variable. Pour cette étude un nouveau corpus de `musique classique` annoté en tonalité locale est également proposé.

L'algorithme d'estimation jointe [accords, premiers-temps] a fait l'objet du développement du logiciel C++ `ircamchord` par Frédéric Cornu. Ce logiciel est par exemple utilisé pour la navigation intra-document par accords dans l'interface Flash `MCIpa` (Music Content Interface Player and Annotation). Cette interface, développée par David Fenech que j'ai encadré dans le projet `Music Discover`, permet la visualisation et l'annotation de l'ensemble des données de contenu (structure, battements, premiers temps, accords, hauteurs multiples et reconnaissance des évènements de batterie). Il permet également une navigation inter-document au travers d'une base de morceaux par critère de genre et de similarité musicale. Il repose sur les technologies d'estimation des caractéristiques de contenu issues des travaux de l'IRCAM, de Télécom ParisTech et de l'Ecole Centrale de Lyon. Je décris ce travail dans l'article [157] et l'illustre à la Figure 3.13.

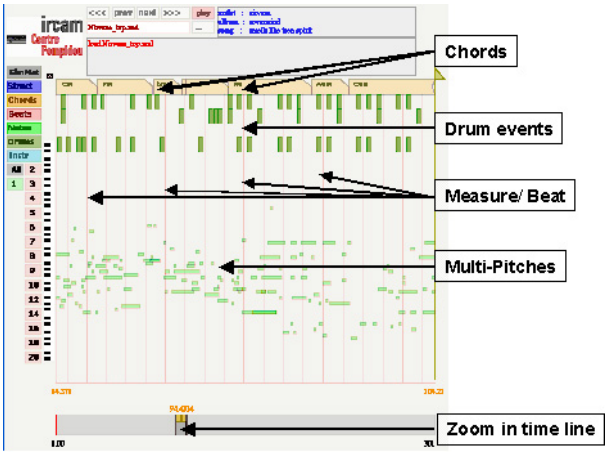

FIGURE 3.13 – Interface `MCIpa`, projet `Music Discover`.

3.3 Estimation d'une structure musicale et d'un résumé audio

Dans cette partie je décris les recherches que j'ai effectuées ou encadrées se rapportant à l'estimation d'une structure musicale et à la génération d'un résumé audio. J'utilise intentionnellement le terme « une » structure musicale et non « la » structure musicale puisqu'il faut bien reconnaître que dans ce domaine l'« élément significatif » à estimer n'a pas de définition unique. Même si de nombreuses techniques ont été proposées pour l'estimation de « la » structure musicale, il apparaît de manière de plus en plus claire qu'il existe une multitude de définitions possibles de cette structure (selon le point de vue adopté).

Pluralité de structures. J'illustre ceci sur la Figure 3.14 où les matrices dites « d'auto-similarité » [55] (encore appelées « recurrence plots » [44]) de deux morceaux sont représentées. Ces matrices $\mathbf{S}_v(t,t')$ représentent la similarité de contenu (mis en évidence selon un point de vue v) entre toutes les paires de temps (t,t') d'un morceau. La figure du haut représente les 100 premières secondes du morceau « Natural Blues » de Moby. Les deux carrés représentent des segments de contenu instrumental homogène. La rupture à la 32$^{\text{ème}}$ seconde correspond à l'entrée de la batterie. La figure du bas représente le même morceau mais selon un point de vue v différent. Les sous-diagonales représentent les répétitions successives des sous-thèmes mélodiques. Malgré le changement d'instrumentation, seule la structure due aux mélodies apparaît. Même si ces deux figures représentent le même morceau, les deux structures visibles sont différentes. Les deux sont pourtant correctes mais illustrent deux points de vue v différents sur le contenu du morceau et donc sur son organisation temporelle. Il existe donc plusieurs « structures » d'un même morceau. Cette multiplicité n'est pas un problème en soi. Cependant, du fait que de nombreux algorithmes ont été proposés pour l'estimation de « la » structure musicale, il est maintenant nécessaire de les comparer. Le problème provient du fait que chaque algorithme a_v utilise un point de vue ou une définition de structure rarement explicitée. De même un corpus annoté $c_{v'}$, utilisé pour comparer ces algorithmes, repose sur une définition d'annotation souvent non explicitée [13]. Il y a donc une forte probabilité pour que $v \neq v'$. Je ne discute pas davantage ici la notion de structure et renvoie le lecteur à la partie 4.1 où je décris nos réflexions et propositions sur la définition et l'annotation en structure.

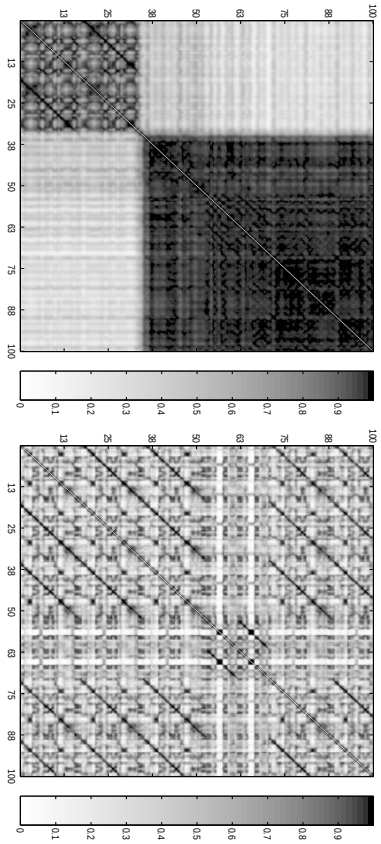

FIGURE 3.14 – Matrices d'auto-similarité correspondants à [Haut] « Natural Blues » de Moby, [Bas] « Natural Blues » de Moby.

Mes recherches sur l'estimation d'une structure ont démarré en 2002 dans le cadre du projet `Cuidado` dans l'objectif d'utiliser cette structure pour la génération d'un résumé audio. Ces recherches se poursuivent actuellement par l'encadrement de Florian Kaiser. Elles ont donné lieu à la publication de 8 articles de conférence (celui de l'ICMC 2003 a reçu le prix du « best paper »), un chapitre d'ouvrage et un brevet international. Je joins en annexe de ce document ce chapitre d'ouvrage.

3.3.1 Etat de l'art

La recherche sur l'estimation de structures musicales vise à déterminer la manière dont les éléments constitutifs d'un morceau de musique sont organisés à un niveau macro-temporel. A l'inverse, l'estimation des battements, premiers temps ou accords concerne la structure micro-temporelle. Les deux sont liés par le fait que la macro-structure se calque généralement sur la micro-structure. Dans nos méthodes d'estimation de structures, cette liaison est réalisée par l'utilisation d'analyses « beat-synchrones ». L'estimation de cette structure peut servir à la génération d'un résumé audio. Elle peut également fournir une information utile à l'utilisateur lui permettant de développer une écoute interactive : comparaison de parties jugées équivalentes par l'algorithme, compréhension visuelle de l'organisation temporelle d'un morceau, navigation par parties ou accès directe aux parties visuellement les plus répétées.

Pour présenter les travaux existants, je reprends la distinction en approches algorithmiques par « états » et par « séquences » que j'ai proposée initialement dans [140] et affinée dans [150], puisque cette distinction est maintenant largement utilisée par la communauté. Pour l'estimation de la structure musicale, deux **hypothèses** sont généralement utilisées. La première considère le morceau comme étant formé d'une succession de

13. Ceci à l'exception des travaux de formalisation de Bimbot [18] ou de Smith [206]. Ce genre de formalisation doit être encouragé.

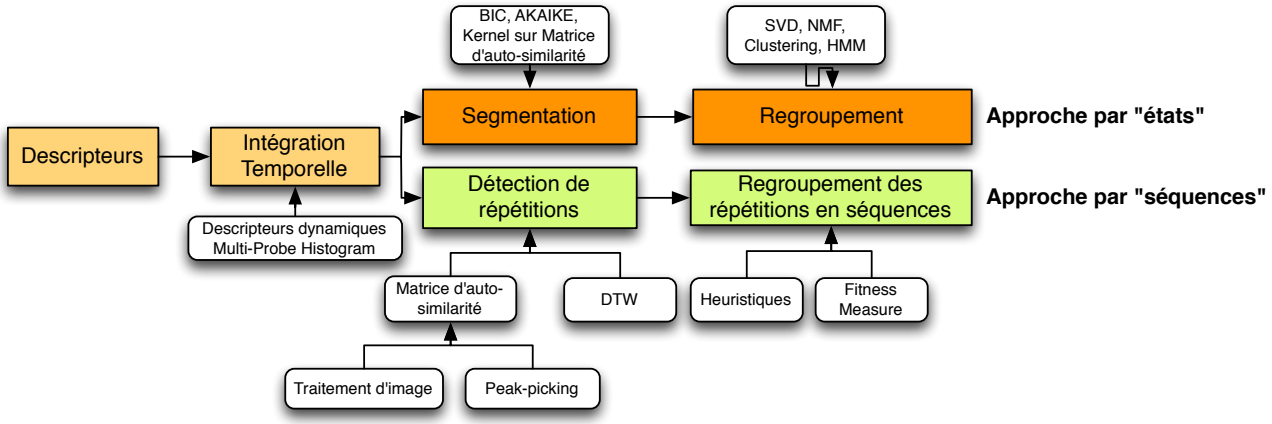

FIGURE 3.16 – Schéma général d'un algorithme de structure musicale.

segments temporels **homogènes** (i.e. contenant une information similaire au sens d'un critère d'observation) et de segments non homogènes. Ces segments homogènes sont illustrés par les lettres « A » et « B » sur la Figure 3.15. La deuxième hypothèse considère que le morceau renferme des **répétitions** temporelles. Elles peuvent correspondre à des répétitions de segments homogènes (comme le segment « B » dans la figure) ou non homogènes (la succession des notes d'une mélodie ne forme pas nécessairement un segment homogène, les notes pouvant être différentes entre elles, mais la mélodie peut être répétée, comme la séquence « abc » dans la Figure 3.15). Ces deux hypothèses conduisent aux deux approches algorithmiques :
— l'approche par **états**, qui considère le morceau comme une succession de segments homogènes répétés ou non,
— l'approche par **séquence**, qui considère uniquement les segments non homogènes répétés.

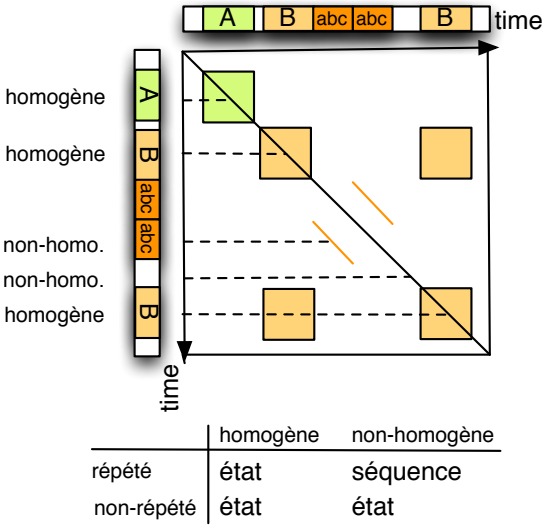

FIGURE 3.15 – Illustration des hypothèses d'homogénéité et de répétition sur une matrice de similarité.

Beaucoup plus de systèmes ont été proposés pour l'approche par **états**. Ceci s'explique par le fait que les systèmes peuvent se rapprocher de ceux utilisés en « speaker diarization » (segmentation couplée à un système de regroupement des segments). Pour la segmentation, les algorithmes reposent donc sur des critères de type BIC, Akaike ou GLR appliqués aux observations temporelles [7] [170] ou encore « Novelty Measure » de [56]. Le regroupement s'effectue par des algorithmes de type K-means [161], clustering hiérarchique par agglomération [102], modèles de Markov cachés [11], clustering spectral [40] voire même NMF [86]. Dans ces approches, il n'y a pas besoin de distinguer les répétitions des non-répétitions puisque toutes formeront des états. Lors de l'utilisation d'algorithmes de clustering, les segments non homogènes non-répétés formeront des états « poubelles ». Cette approche ne peut cependant pas détecter les répétitions quand les segments ne sont pas homogènes. Ceci est l'objectif de l'approche suivante.

Dans l'approche par **séquences**, les critères de segmentation usuels ne peuvent plus être utilisés. Ces approches nécessitent en premier lieu d'estimer les temps de contenu répété puisque seules les répétitions sont utilisées pour caractériser la structure. Pour cela, les algorithmes reposent souvent sur l'analyse d'une matrice dite d'auto-similarité (détection des répétitions représentées par les sous-diagonales [15] [63]), sur l'alignement dynamique du temps [34] [119] ou encore sur les facteurs d'oracle [93] [43]). Les répétions détectées sont ensuite regroupées en séquences. Ceci peut se faire par utilisation d'heuristiques de regroupement [63] ou par des approches de type maximum de vraisemblance [146] [121].

3.3.2 Contributions

Nos recherches sur l'estimation de structures musicales s'articulent autour de la conceptualisation et de la création de **corpora annotés** en structure (voir partie 4.1), du développement d'algorithmes pour l'estimation de structure par approche **états** (partie 3.3.2.1), par approche **séquences** (voir partie 3.3.2.2) et du dévelop-

pement d'algorithmes pour la génération de résumés audio (voir partie 3.3.2.4). Je présente à la Figure 3.16 l'interconnexion des différents éléments étudiés pendant nos recherches et je les résume ci-dessous.

3.3.2.1 Approches par « états »

Modèle HMM et descripteurs dynamiques. Notre première étude, publiée en 2002 dans [161] repose sur un algorithme en deux étapes : une segmentation temporelle suivie d'un regroupement des segments (approche devenue très classique aujourd'hui). Une des spécificités de cette approche est l'utilisation d'une intégration temporelle à fenêtre longue de type « descripteurs dynamiques » (voir partie 2.2.1.1). Cette intégration vise à résumer les caractéristiques locales du signal, communes à un ensemble successifs de trames, afin de permettre (dans une certaine mesure) le traitement de séquences (segments non homogènes répétés) par une approche par états. L'estimation de la structure s'effectue ensuite en deux étapes. La première segmente le flux temporel d'observations selon un critère de variation trame-à-trame. La seconde regroupe ces segments en états. Celle-ci procède elle-même en trois étapes. Les segments sont d'abord regroupés selon un critère de similarité de leur contenu moyen puis le résultat sert à l'initialisation d'un algorithme de type « Fuzzy-K-Means » dont le résultat est lui-même utilisé pour l'initialisation d'un modèle de Markov caché. L'utilisation du HMM vise à introduire une contrainte temporelle entre les états. La Figure 3.17 illustre, de bas en haut, le flux temporel de vecteurs d'observation, l'assignation obtenue par « Fuzzy-K-Means », et le résultat du décodage HMM. Les performances de cette estimation ne sont pas mesurées dans [161] puisqu'il n'existait, à l'époque, ni corpus annoté [14] ni métrique d'évaluation. Cette estimation de structure est utilisée pour la navigation intra-document par parties (j'illustre l'interface présentée lors de la conférence ISMIR-2002 à la Figure 3.18) ainsi que pour la génération d'un résumé audio (voir partie 3.3.2.4).

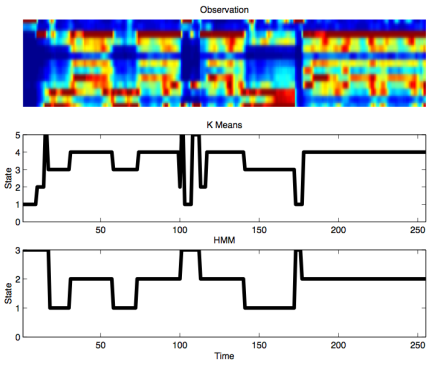

FIGURE 3.17 – Illustration de l'estimation de structure pour le morceau « Head Over Heels » de Alanis Morisette.

FIGURE 3.18 – Interface de navigation intra-document par parties présentée à ISMIR-2002.

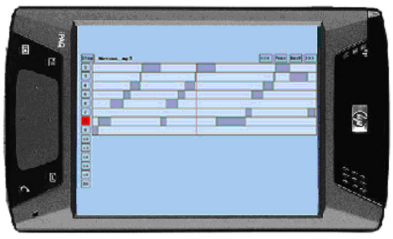

FIGURE 3.19 – Interface Flash de navigation intra-document par représentation hiérarchique portée sur PDA HP, projet `SemanticHIFI`.

Late-fusion et clustering hiérarchique par agglomération. L'étude suivante, ayant conduit au développement du logiciel C++ `ircamsummary`, n'a malheureusement jamais été publiée en dehors des « Extended Abstracts » de `MIREX`. Dans cette étude le nombre d'observations représentant le contenu du signal est élargi. Comme dans nos approches par « séquences » [146] nous considérons maintenant trois points de vue sur le signal : les contenus timbral (représenté par des MFCCs), harmonique (par des Chromas) et bruité (par les coefficients Spectral Valley et Crest [80]). Le calcul de ces observations est effectué de manière « beat-synchrone ». Les trois matrices de similarité correspondant à ces trois points de vue sont ensuite combinées (late-fusion). Le critère de segmentation « trame-à-trame » est remplacé par un critère reposant sur le calcul de la « Novelty Measure » proposé par Foote [56]. Ceci permet de prendre simultanément en compte la variation inter-segments et l'homogénéité intra-segment. L'algorithme de regroupement est également modifié du fait de la difficulté à estimer correctement le nombre d'états cachés du HMM. Notons que ce nombre est en partie subjectif (comme l'est le fait de regrouper deux segments homogènes [A,A] en un segment unique [A]). C'est pourquoi nous voulons laisser ce choix à l'utilisateur au travers d'une interface. Pour cela, les segments doivent cependant posséder une organisation hiérarchique. J'illustre à la Figure 3.19 l'interface développée par Samuel Goldszmidt et Ludovic Gaillard pour PDA HP dans le projet `SemanticHIFI` permettant la navigation hiérarchique. Cette interface a fait l'objet d'un ensemble de tests utilisateurs que nous décrivons dans l'article [22]. Pour garantir cette hiérarchie, nous optons pour un algorithme de clustering hiérarchique par agglomération des segments détectés. Celui-ci opère le regroupement sur base de deux distances : une distance entre segments considérés comme « états » (donc

14. Suite à cet article, j'ai créé le premier corpus en « états » et « séquences » dans le cadre de `MPEG-7` Audio. Ce corpus figure aujourd'hui parmi ceux distribués par l'université Queen-Mary de Londres.

homogènes), une distance entre segments considérés comme des instances d'une « séquence » (donc considérés comme non homogènes mais répétés [15]). Le regroupement de deux segments est ensuite contraint par des critères visant à favoriser la jonction de segments adjacents et à homogènéiser leurs durées. Ces contraintes peuvent être rapprochées de celles utilisées par Paulus [134] ou Sargent [195].

Recherches actuelles. Les études actuelles sur l'approche par « états » sont effectuées par Florian Kaiser que j'encadre dans le projet `Quaero`. Ces études visent à étendre ses travaux précédents sur l'utilisation de Factorisations Matricielles Non-Négatives (NMF) appliquées aux matrices d'auto-similarité [86] et d'intégrations temporelles des Chromas par « multi-probe histogram » [83] (voir partie 2.2.1.2). Pour ce dernier point, nous avons récemment proposé dans [84] une technique reposant sur une mesure de nouveauté permettant de déterminer automatiquement la position et la taille de fenêtre optimale pour effectuer cette intégration sans effet de « lissage ». L'article que nous avons récemment soumis [85] propose quant à lui l'utilisation de noyaux de segmentation [56] multi-échelles (permettant de résoudre le problème de la détermination de la taille du noyau que comporte la méthode de Foote [56]). Nous proposons également de nouveaux types de noyaux pour la mise en évidence des transitions entre segments homogènes et non homogènes.

Evaluations. Depuis 2009, nous évaluons ces algorithmes lors des campagnes `MIREX`. Pour les raisons évoquées dans l'introduction de cette partie, et du fait de la multiplicité des métriques utilisées dans `MIREX` (14 métriques), il est cependant difficile de tirer des conclusions de ces évaluations. Les performances de notre algorithme varient assez fortement selon les critères mais obtiennent généralement de bonnes performances pour les métriques « Random Clustering Index », « Normalized Conditional Entropy Under-Segmentation Score » et « Frame Pairwise Clustering Under-Segmentation », indiquant donc une tendance à la sur-segmentation.

3.3.2.2 Approches par « séquences »

La spécificité des approches par « séquences » se trouve dans la nécessité de détecter en premier lieu les segments répétés $s_1, s_2 \dots s_N$. La répétition d'un segment est visible sous la forme d'une sous-diagonale dans une matrice d'auto-similarité. Cette diagonale apparie donc deux segments (par exemple s_2 et s_5). Dans un second temps, les segments détectés sont regroupés en différentes séquences $S_A, S_B \dots$ sur la base de leur similarité. La notation $\{s_2, s_5, s_8, s_9\} \in S_A$ indique que ces segments sont des répétitions d'une même « séquence » S_A (par exemple les différentes répétitions d'une même mélodie). On dit que la séquence S_A est « instanciée » en différents segments temporels s_2, s_5, s_8 et s_9. Dans nos approches, les segments appartenant à une même séquence ne se recouvrent pas temporellement. Nos différentes propositions se distinguent par les algorithmes utilisés pour ces deux étapes : la détection de segments répétés ainsi que leur regroupement en séquences. L'utilisation de l'oracle des facteurs, que je présente en premier, est cependant une exception à cette décomposition en deux étapes.

Oracle des facteurs. En 2002, dans le stage de Master de Amaury Laburthe [93] que j'ai encadré, nous avons étudié la représentation de la structure par oracle des facteurs. Il s'agit donc implicitement d'une approche par séquences. Dans cette approche, l'application de l'oracle des facteurs [4] sur des données de valeurs continues de type descripteurs audio est rendue possible par une quantification de leurs valeurs ainsi considérées comme symboles. La méthode « Audio Oracle » de Dubnov [43] fournira une autre solution par seuillage de distances entre descripteurs audio (ce seuillage est équivalent à la comparaison de symboles).

Détection des segments répétés. Alors que dans l'oracle des facteurs la détection de segments répétés et leur regroupement en séquences sont faits de manière concomitante, dans nos autres algorithmes ces étapes sont successives. Dans [166] et [146] les segments répétés sont obtenus par détection des sous-diagonales dans une matrice d'auto-similarité. Pour cela, je propose dans [166] l'utilisation d'un algorithme de filtrage structurant en 2-dimensions (traitement de l'image) appliqué à la matrice d'auto-similarité. Le filtrage structurant présente l'avantage par rapport aux filtrages classiques de ne pas lisser temporellement l'information de départ et de fin des segments. Cet algorithme se montre cependant très sensible au paramétrage et coûteux en temps de calcul. Nous étudions donc dans [146] la méthode de détection proposée par Goto [63] (détection des pics de la somme à travers les temps d'une matrice d'auto-similarité exprimée en décalages). Dans [146] nous proposons d'appliquer cette méthode à la matrice fusionnée des matrices de contenu timbral, harmonique et bruité. Nous proposons également les matrices d'auto-similarité d'ordres supérieurs. Ces matrices d'ordres supérieurs visent à pallier le phénomène de transitivité attendu par les algorithmes de détection de structure (si $A \simeq B$ et que $A \simeq C$, nous devrions observer $B \simeq C$) mais souvent occulté par des variations entre segments. Les matrices de similarité d'ordres supérieurs intègrent pour cela la similarité à travers des temps « pivots » u : $\mathbf{S}_2(t,t') = \int_u \mathbf{S}(t,u)\mathbf{S}(u,t')\partial u$. Cette nouvelle méthode permet une estimation plus robuste des segments répétés mais ne permet pas la prise en compte de variations temporelles entre répétitions (ralentis, accélérations). Pour cette raison, nous avons étudié dans le stage d'Alexandre Wronecki [220] que j'ai encadré,

15. Pour cela, parmi l'ensemble des diagonales permettant la connexion des deux segments, nous utilisons celle de moindre coût.

l'utilisation d'algorithmes de type Alignement Dynamique du Temps (DTW). Dans ce cas, toutes les possibilités d'alignements entre segments possibles (choix du démarrage et de la durée) sont testées. Les distances utilisées dans le DTW correspondent à l'inverse des valeurs de la matrice d'auto-similarité. Un alignement de coût global faible met en évidence la présence de deux segments s_n et $s_{n'}$ répétés. Cet algorithme se montre cependant extrêmement coûteux.

Regroupement en séquences. Dans un second temps, les segments répétés détectés sont regroupés en séquences. Dans [166] nous proposons pour cela une méthode heuristique permettant leur appariement progressif (voir chapitre joint en annexe). Dans [146], nous proposons une approche globale visant à déterminer l'ensemble minimal de séquences expliquant le mieux les segments détectés. Chaque séquence possède une séquence « mère » qui est le segment prototype duquel les autres segments de la séquence sont dérivés (avec ou sans modifications). Nous testons l'ensemble des intervalles possibles comme candidats pour cette séquence « mère ». Un score, inspiré du « summary score » de Cooper [39], permet de mesurer ensuite l'explication des segments observés fournie par cette séquence « mère ». La séquence « mère » la plus explicative fournit simultanément l'ensemble des segments qu'elle instancie. Le regroupement est donc simultané, et non progressif. La méthode est ensuite itérée sur les segments non encore expliqués. Pour l'évaluation de cette méthode, nous proposons de nouvelles métriques permettant différentes prises en compte de la sous-segmentation ou sur-segmentation ainsi qu'un corpus public [16]. Cet algorithme s'est montré très efficace pour le regroupement des segments en séquences (voir Figure 3.20 pour un exemple d'estimation) mais nécessite cependant une détection préalable de ces segments. A ce titre, l'algorithme récemment proposé par [122], proche conceptuellement de nos travaux mais ne nécessitant pas de détection préalable des segments répétés, m'apparaît intéressant.

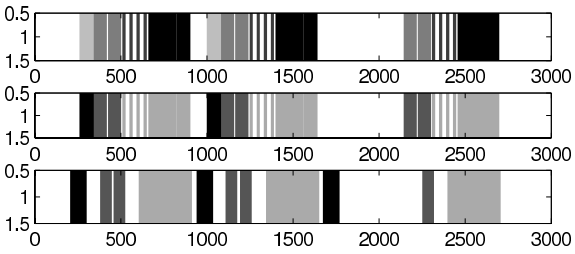

FIGURE 3.20 – Exemple d'estimation de structure : [Haut] annotation de la structure, [Milieu] alignement de l'annotation avec la structure estimée, [Bas] structure estimée sur le morceau « Smells like teen spirit » de Nirvana.

3.3.2.3 Choix entre une approche par « états » et par « séquences »

Les approches par « états » et par « séquences » sont des approches algorithmiques. Le choix de l'une ou l'autre pour représenter la structure d'un morceau dépend non seulement de son contenu (les premières compositions des Beatles tendent à se baser sur des répétitions de mélodies, donc de « séquences » ; à l'inverse la musique « grunge » fonctionne souvent par blocs d'instrumentation très différents donc par « états ») mais dépend également, comme l'illustre l'exemple de la Figure 3.14, de l'observation que nous avons de son contenu (le point de vue v). Cette observation est déterminée par le choix des descripteurs et des post-traitements appliqués sur ceux-ci. Pour cela, je propose dans [150] une mesure permettant d'estimer l'approche la plus appropriée pour un morceau en fonction des observations utilisées.

3.3.2.4 Génération de résumés audio

Un « résumé audio » se doit de résumer le contenu d'un morceau de musique. Alors que ces résumés (encore appelés « thumbnails », « snippets » ou « previews ») sont généralement créés en sélectionnant un extrait de 30 s au début ou à partir de la 45$^{\text{ème}}$ seconde du morceau (Amazon, 7-Digital, iTunes), nous proposons de les générer de manière à ce qu'ils renferment les éléments **représentatifs** du morceaux. A la différence de la technique proposée par Cooper [39] (détectant l'instant clef du morceau comme le plus répété dans une matrice d'auto-similarité), la technique que j'ai publiée dans [161] et brevetée dans [139] vise à créer une bande-annonce obtenue par concaténation de segments sélectionnés dans la structure musicale estimée. Pour cela, j'étudie dans [161] différents critères de sélection des segments et différentes techniques de montage.

Sélection des segments. Le choix des segments dépend de la signification donnée à la notion « représentatifs du morceau ». Le résumé peut reproduire l'ordonnancement temporel des segments (A, B, A, B, C, A, B), fournir un aperçu unique de chaque segment (A, B, C), fournir un aperçu des transitions entre segments $(A \rightarrow B, B \rightarrow A, B \rightarrow C)$ ou fournir un aperçu audio de la partie la plus importante. En pratique, la configuration A, B, C est jugée la plus pertinente (en terme de quantité d'informations apportée par le résumé, rapportée à sa durée). Lors du stage de François Mislin [115] que j'ai co-encadré, cette technique a servi à la génération des résumés audio de la base musicale de la médiathèque de l'IRCAM.

16. Il s'agit du corpus MPEG-7 Audio Structure Sequence.

Evaluation. Dans le cadre du projet `Quaero`, les résumés audio ont été intégrés dans l'application MSSE développée par Orange. Dans l'article [152], nous décrivons les études complémentaires (qualitatives et quantitatives) ayant été effectuées à cette occasion. Nous comparons quatre méthodes de résumé : un extrait de 30 s au début, un extrait de 30 s aléatoire, l'extrait le plus représentatif de 30 s (dénommé 1x30) et la concaténation beat-synchrone des trois extraits les plus représentatifs (3x10). Ces méthodes ont été comparées par des tests utilisateurs sur des musiques connues ou inconnues. Les questions suivantes leur étaient posées : « quelle technique résume le mieux le morceau ? » (musiques connues), « quelle technique fournit le plus d'information ? » (musiques inconnues). Dans les deux cas, le résumé 3x10 a été jugé meilleur. Une étude quantitative a également été effectuée pour comparer les résumés 1x30 et 3x10. A défaut de pouvoir utiliser la « présence du chorus dans le résumé » comme critère (puisque la définition du chorus s'avère problématique, voir partie 4.1), nous utilisons la « présence du titre du morceau dans les paroles chantées ». Sur un corpus de test de 160 morceaux, nous avons mesuré le nombre de résumés contenant le titre du morceau. Le résumé 3x10 a atteint 95% de bonne détection contre 90% pour le 1x30.

Technique de montage. La technique de montage proposée dans [161] est une concaténation beat-synchrone des segments sélectionnés (Beat-Synchronous Overlap-Add). Cette technique est illustrée à la Figure 3.21. Je l'ai étendue plus tard à la concaténation downbeat-synchrone. L'article [152] traite également des tests perceptifs effectués pour cette technique de montage.

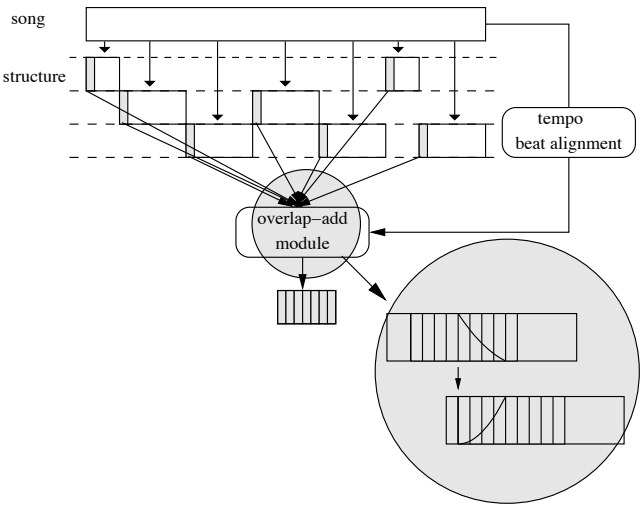

FIGURE 3.21 – Génération de résumés audio par concaténation beat-synchrone de segments

CHAPITRE 4

Création de corpora annotés et campagnes d'évaluation

Dans cette partie je décris nos travaux concernant la création de corpora annotés et les campagnes d'évaluations. Les deux étant étroitement liés j'ai choisi de les présenter dans une même partie.

4.1 Création de corpora annotés

Le développement des technologies d'indexation automatique de contenus audio nécessite l'accès à des données annotées. L'ensemble de ces données annotées constitue un « corpus annoté ». Cette accès est nécessaire pour permettre l'observation de signaux correspondant aux concepts à estimer, l'apprentissage automatique de ces concepts et l'évaluation des performances des technologies développées.

Dès mes premières recherches (reconnaissance d'instruments), l'accès à des corpora (correctement) annotés m'est apparue indispensable. Au cours des années, j'ai dû bien souvent « mettre la main à la pâte » pour la création de ces corpora. Je ne peux d'ailleurs que recommander à tout chercheur de faire de même ; cette tâche souvent déconsidérée est finalement pleine d'enseignements sur ce que l'on cherche réellement à estimer. Mais c'est en particulier dans le cadre du projet `Quaero` que mon activité dans ce domaine s'est développée. Une partie de ce projet étant dédiée à la constitution de corpora, il m'a paru important de profiter de cette occasion pour démarrer une réelle réflexion sur l'annotation de corpora en MIR. Pour cela, j'ai mis en place à l'IRCAM une équipe dédiée à l'annotation, constituée de spécialistes de différents domaines : musicologie, ingénierie du son, oreille d'or, programmation radio (Emmanuel Deruty, Maxence Riffault, Jean-François Rousse, Nicolas Baubillier). Je décris nos travaux au regard de la terminologie que j'ai récemment proposée avec Karën Fort de l'INIST dans l'article [159].

Proposition de description des corpora MIR annotés. Dans [159] nous proposons une méthodologie de description des corpora MIR annotés. Cette proposition vise à permettre un meilleur partage et une meilleure utilisation de ces corpora. En effet, comme il n'existe pas d'institution dédiée à la constitution de corpora annotés en MIR, ceux-ci sont directement produits par les différents laboratoires. De ce fait les méthodologies sont souvent très différentes. Malheureusement, ces corpora sont souvent distribués sans aucune description des méthodologies utilisées. Ceci conduit parfois à leur utilisation à contresens ou à des conclusions erronées dérivées de leur utilisation. Notre proposition vise à aider les créateurs de corpora dans leur démarche de description. Pour la description de ces corpora, nous distinguons deux grands axes. Le premier décrit le **choix des items constituant le corpus**. Celui-ci distingue les items synthétiques (résultant d'une synthèse audio [221]), les items enregistrés spécialement pour la création d'un corpus (par exemple les corpora `Studio-OnLine` [14], `RWC` [64] ou `ENST-Drum` [61]) et les items issus du monde réel (`Isophonic` [106] ou `Million Song Dataset` [16]). Dans ce dernier cas se pose la question du critère de sélection ayant été utilisé pour les choisir.

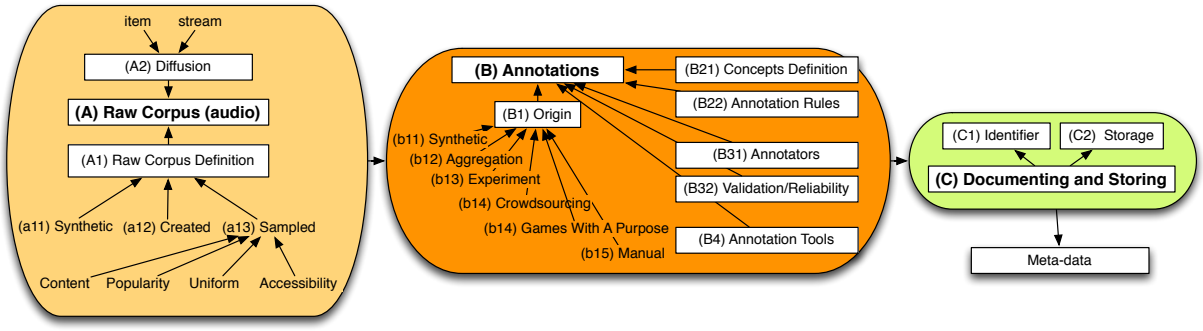

FIGURE 4.1 – Principaux axes de description d'un corpus MIR annoté selon notre proposition [159].

Le deuxième axe distingue la **provenance des annotations**. Celles-ci peuvent être obtenues :

automatiquement : soit à partir des paramètres de synthèse [221], soit par agrégation de différents contenus présents sur le web [16],

manuellement : soit par des expériences (comme celles sur le timbre), soit par des techniques de foule de type « crowdsourcing » ou « games with a purpose » (comme celles sur le tempo perceptif [98]), soit par l'annotation manuelle traditionnelle effectuée par des experts.

Les autres axes visent à définir les concepts annotés, les règles d'annotation en ces concepts, les annotateurs et la validité des annotations. Ces axes de description sont résumés à la Figure 4.1.

Je décris le travail que nous avons accompli dans le projet `Quaero` au regard de cette proposition de description. Les références (A) (a13) (B21) données dans le texte correspondent à celles de la Figure 4.1.

(A1) Choix des items. Les items de notre corpus `Quaero` correspondent à des items réels (a13), plus spécifiquement à des morceaux de musique commercialisés. Différents critères de sélection ont été utilisés. Un premier ensemble d'items a été sélectionné sur la base de leur popularité (statistiques de vente du Billboard magazine), un autre sur la base de leur contenu (mise en évidence de musique possédant certains rythmes, certaines suites d'accords), le plus grand ensemble d'items a été sélectionné selon la méthode de l'échantillonnage uniforme. Cette technique consiste à échantillonner un espace de description de manière uniforme. Les axes de cet espace correspondent aux différentes catégories de description fournies par la base de métadonnées `All-Music-Guide` : genre, sous-genre, humeur et année. Finalement, un critère d'accessibilité des items a été ajouté puisque seuls les morceaux du catalogue EMI nous étaient accessibles. Un corpus de 50.000 items a ainsi été créé.

(B1) Origine des annotations. Les annotations ont été faites manuellement (b15) par les quatre experts (B31) mentionnés dans l'introduction.

(B21) Définition des concepts à annoter. Afin de sélectionner les concepts d'annotation les plus pertinents pour l'annotation MIR, un critère de validation de concepts a été développé. Celui-ci, appelé « Perceptual Recognition Rate » (PRR), mesure l'applicabilité d'un concept à un morceau. Nous avons ainsi montré que le concept de « chorus » était difficilement applicable puisque pour un nombre important de morceaux il n'est pas possible de dire si oui ou non il y a un « chorus ». En utilisant ce PRR, les critères d'annotation usuels en MIR ont été testés (battements, structure, accords, voix, humeur ou genre musical . . .). De nouveaux concepts ainsi que des redéfinitions de concepts existants ont ainsi été proposés pour la description du *genre* musical (sur une base purement acoustique), des *accords* (sur base du rôle harmonique de l'accord dans la tonalité locale) et de la *structure* musicale. Je décris à titre d'exemple ce dernier travail que nous avons publié dans [155] et [154].

Recherche sur l'annotation en structure musicale. Pour expliquer le concept de structure musicale, les notions de « couplet » et « refrain » sont généralement utilisées et semblent être comprises de tous. Il existe de fait une catégorie de musiques populaires stéréotypées pour laquelle ces notions sont applicables. En dehors de celle-ci ces notions deviennent très floues. Dans [155] et [154], nous étudions, avec Emmanuel Deruty, l'annotation en structure musicale. Nous proposons d'abord une formalisation des systèmes d'annotation existants. Pour cela, nous les distinguons selon qu'ils reposent sur le *rôle musical* du segment (utilisation des notions d'intro, couplet, refrain, pont), sur la *similarité acoustique* entre segments, sur l'*application* finale [1], ou sur les résultats de *tests perceptifs* [25]. Signalons l'existence d'une autre approche, proposée par l'INRIA/Metiss [17], très intéressante mais non publiée à l'époque de notre étude, définissant la structure en terme de paradigme et de syntagme dans l'optique de la rapprocher

1. Une application visant la génération automatique de résumés se concentrera par exemple sur le segment le plus représentatif quel que soit son rôle.

des modèles de langage. Dans [155] et [154], nous proposons ensuite un nouveau système d'annotation multi-dimensionnel permettant de rendre compte des différents points de vue possibles sur la structure. L'élément de base, appelé Constitutive Solid Loops (CSL), représente une phrase musicale ou une exposition musicale (succession d'accords). Ce CSL peut être répété ou non. Les *variations* de ces CSL (variation d'instrumentation ou d'intensité) sont décrites comme des caractéristiques additionnelles. Les instruments présents ne sont pas décrits par leur nom mais par leur *fonction* dans le morceau : un « primary lead » indique qu'il s'agit de l'instrument principal qu'il s'agisse d'une voix ou d'un saxophone. Le *rôle* musical d'un segment est uniquement indiqué lorsque celui-ci est évident : « obvious chorus ». La Figure 4.2 illustre un exemple d'application de ce système d'annotation où chaque ligne représente un point de vue descriptif.

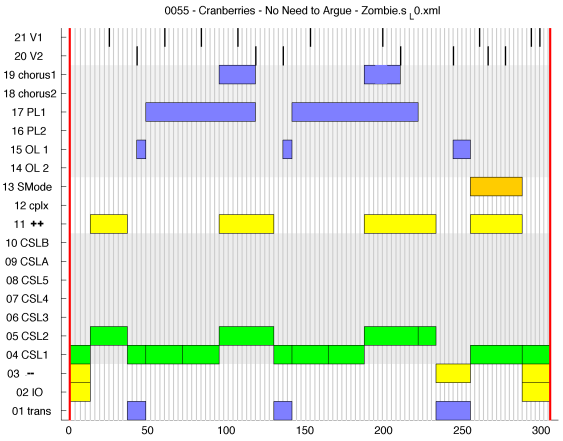

FIGURE 4.2 – Illustrations d'une annotation multi-dimensionnelle sur le morceau « Zombie » de The Cranberries

(B31) Annotation. J'ai mis en place deux campagnes d'annotation : la première s'est déroulée de juillet 2008 à septembre 2009, la deuxième de juin 2010 à mai 2011. Suivant les recommandations de [57], chaque campagne a été précédée d'une précampagne sur un nombre réduit de morceaux. Le but de ces précampagnes était de vérifier l'*adéquation* des concepts à annoter aux morceaux, d'améliorer l'*utilisabilité* des interfaces d'annotations et de vérifier l'*accord entre les annotateurs* pour chaque concept. Au cours des deux campagnes, environ 500 morceaux ont été annotés en critères locaux (battements, premiers temps, accords, structures, segments chantés) et 8.000 en critères globaux (de type genre, humeur, caractéristiques vocales).

(B32) Validité des annotations. Les annotations globales de la première campagne ont été effectuées indépendamment par trois annotateurs, la deuxième campagne par deux. Les corpora issus de ces campagnes d'« annotation » servent pour les campagnes d'« évaluation » du projet `Quaero`, nommées `Quaero-Eval` (voir partie 4.2). Pour celles-ci, les données test sont sélectionnées sur la base de l'accord inter-annotateur. Si nous notons c_n^a le concept c appliqué par un annotateur a à la donnée n, seuls les critères c ayant un accord inter-annotateur élevé ont été gardés ; pour ce sous-ensemble de critères, seules les données n amenant à un accord élevé sont gardées. Pour des raisons de coût, les annotations locales ont été effectuées par un seul annotateur et vérifiées par un second. Lors des campagnes d'évaluations, ces données font donc l'objet d'une phase d'adjudication.

(B4) Outils d'annotation. Vu l'importance du travail d'annotation à accomplir, j'ai développé deux plates-formes d'annotation ainsi qu'un format de stockage de ces annotations. La première plate-forme, nommée QIMAG (voir Figure 4.3), est dédiée à l'annotation des concepts globaux (genre, humeur, tags). Elle est développée sous la forme d'une interface web connectée à une base de donnée (Flash, Php, Mysql, Apache) afin de permettre la gestion centralisée d'un très grand nombre de tâches d'annotation. Son développement reprend les principes proposés par Herrera pour la plate-forme MUCOSA [75]. En fin de campagne d'annotation, la base contenait plus d'un million de quintuplets [ID-morceaux, ID-annotateur, ID-questionnaire, ID-question, réponse)]. La seconde plate-forme, nommée QIMAL (voir Figure 4.4), est dédiée à l'annotation des concepts locaux en temps (battements, premiers temps, accords, structures, segments chantés). Elle est développée sous la forme d'une interface `matlab` compilée. Les spécificités de cette application sont l'utilisation de calques superposés (calques de représentations – signal, spectrogramme, chromagramme, matrice de similarité – et calques d'annotations) et l'utilisation de marqueurs/segments typés (chacun appartient à un concept dont les valeurs possibles, parmi son dictionnaire, sont indiquées par un « piano-roll » vertical). Ces spécificités ont pour but d'accroître la vitesse d'annotation (par rapport à l'utilisation d'outils non dédiés) et de les rendre conformes à la définition des concepts (dictionnaire de valeurs).

(C) Stockage et distribution. Le format de stockage des descriptions que nous avons défini et qui a été utilisé dans la partie « musique » du projet `Quaero` est un format XML, nommé `musicdescription`. Il peut être considéré comme une version simplifiée (beaucoup plus légère) du format `MPEG-7`, tout en gardant les possibilités de typage de données, de définition de dictionnaires de valeurs et en y rajoutant la possibilité de description de signaux multi-canaux. Ce format est maintenant intégré dans le logiciel `Audiosculpt 3.0` ainsi que dans l'ensemble des logiciels `ircambeat`, `ircamchord` et `ircamsummary`.

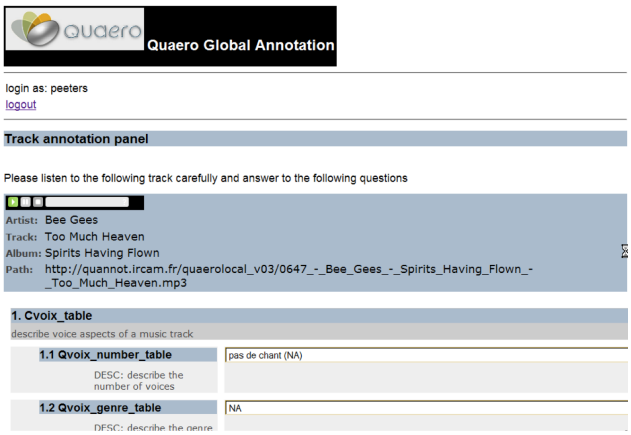

FIGURE 4.3 – Interface de QIMA-G (Quaero Interface Music Annotation Global), IRCAM, projet `Quaero`.

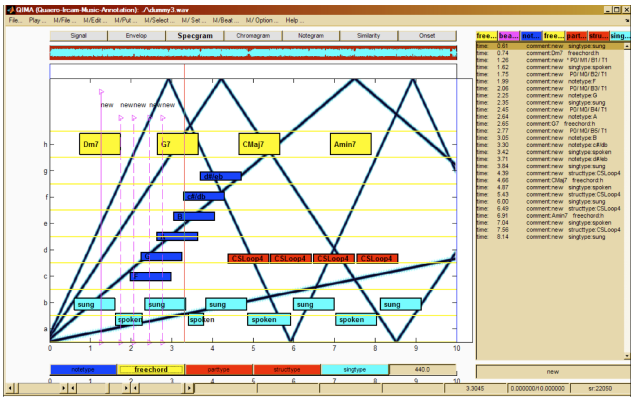

FIGURE 4.4 – Interface de QIMA-L (Quaero Interface Music Annotation Local), IRCAM, projet `Quaero`.

4.2 Campagnes d'évaluation

Comme dans d'autres domaines relatifs à l'indexation, l'évaluation des technologies développées revêt une place importante en indexation musicale. Outre le fait de permettre la comparaison de technologies, l'établissement d'une évaluation implique de réfléchir au domaine de leur application. En effet, l'évaluation nécessite de définir conjointement le concept qui doit être estimé par les algorithmes, le type de données sur lequel ces algorithmes doivent pouvoir être appliqués, et la manière dont la qualité des estimations doit être mesurée (celle-ci vise à mettre en avant ce qui est important dans l'estimation du concept). Dans le domaine de l'indexation musicale (MIR), ces évaluations ont vu progressivement le jour depuis 2004. Elles ont permis de changer profondément le domaine MIR : dans un premier temps en ancrant dans les plannings de recherche la nécessité d'évaluer ce qui est proposé, et actuellement, en incitant les chercheurs à réfléchir sur la bonne manière d'évaluer les technologies (concepts et données). Dans cette partie je décris mes contributions concernant les campagnes d'évaluation. En effet, si nous participons depuis 2005 aux évaluations `MIREX` (participation à 12 tâches en 2011), nous n'en avons pas moins un regard averti sur la manière dont ces évaluations sont organisées. Ces considérations m'ont conduit aux spécifications des campagnes `Quaero-Eval`, à ma participation à `Media-Eval` et à l'organisation d'une session dédiée à l'évaluation à ISMIR-2012.

Considérations sur l'état de l'art. La systématisation des campagnes en MIR commence en 2004 avec l'organisation par le MTG de l'« Audio Descriptor Contest ». En 2005, le laboratoire IMIRSEL de l'Université d'Illinois propose les évaluations `MIREX` (Music Information Retrieval Evaluation eXchange) qui sont depuis lors devenues le standard d'évaluation en MIR. `MIREX` évalue chaque année environ 20 tâches (de la détection d'onsets à l'estimation de tags utilisateurs). La particularité de `MIREX` réside dans le modèle **algorithm-to-data** utilisé : chaque participant soumet son algorithme au laboratoire IMIRSEL, celui-ci l'applique aux données, les résultats sont ensuite retournés aux participants qui n'accèdent donc jamais à ces données. Cette non-distribution s'explique du fait des droits associés à ces données audio, et par le coût que constituerait la création de nouveaux corpora chaque année. Cette non-distribution permet de réutiliser les données d'année en année. L'inconvénient de ce modèle est que les participants n'ont que la seule information des scores obtenus. Il est difficile de tirer de ceux-ci les enseignements nécessaires à l'amélioration des algorithmes et de la recherche en général. Ces évaluations ont malgré tout un grand succès (337 systèmes ont ainsi été évalués en 2010). La participation est entièrement gratuite, il suffit de soumettre un algorithme. Le coût de fonctionnement de l'ensemble de l'évaluation est pris en charge par IMIRSEL. La forme de facto **centralisée** de ces évaluations renferme cependant un danger. Comme il suffit de soumettre son algorithme, peu de gens se soucient de la pertinence des protocoles expérimentaux (certaines tâches résultent du don fait par un étudiant en fin de thèse d'un corpus de test et de ses métriques d'évaluation), protocoles reconduits d'année en année (les nouveaux arrivants considérant les protocoles comme mûrement réfléchis par leur prédécesseurs). Compte tenu de l'impact que peuvent avoir les résultats obtenus à `MIREX`, ce dernier point est particulièrement problématique.

A ce propos, l'article que nous avons récemment soumis avec Johan Pauwels [136] illustre clairement (dans le cadre d'une tâche d'estimation d'accords `MIREX`) que le choix d'une métrique d'évaluation particulière impacte directement les classements des algorithmes participants.

Quaero-Eval. Quaero-Eval est la plate-forme d'évaluation des technologies d'indexation musicale que nous avons créée avec les partenaires du projet (l'IRIT, Télécom ParisTech, l'INRIA/Metiss et le LIMSI). Quaero-Eval vise à pallier les défauts de MIREX tout en gardant ses qualités. A cette fin, un modèle intermédiaire entre MIREX et ESTER [2] est proposé. Comme dans MIREX, les algorithmes sont évalués par un partenaire indépendant, l'IRIT, qui est chargé de fournir la puissance de calcul nécessaire pour faire tourner tous les algorithmes et de livrer les résultats à date fixée. Comme dans ESTER, à la suite de chaque campagne, les données de test sont distribuées aux participants (de nouvelles données de test sont créées chaque année). Cette distribution permet une phase d'adjudication (correction d'éventuelles erreurs dans les données test et donc correction des résultats) et une analyse détaillée des résultats. Les données distribuées servent de données de développement pour l'année suivante. Chaque participant engrange donc de nouvelles données chaque année pour faire avancer sa recherche. Une autre différence majeure avec MIREX est l'élaboration collaborative de l'évaluation. Les tâches à évaluer et le protocole expérimental sont définis d'un commun accord par les partenaires du projet. Les environnements d'évaluation (implémentations exactes des mesures de performances) sont créés de manière collaborative (gestion sous forme de SVN). Etant donnée l'accessibilité de ces environnements et des données tests, l'ensemble de l'expérience peut être reproduite. Suite à chaque campagne d'évaluation, un rapport est écrit et une réunion est généralement tenue afin de tirer les conclusions de la campagne. Depuis 2009, Quaero-Eval évalue chaque année environ 7 tâches différentes dont certaines, comme la séparation de sources, ne font pas partie des tâches MIREX.

Media-Eval. Récemment, je me suis également rapproché de l'initiative d'évaluation multimédia Media-Eval [3]. Celle-ci propose une organisation décentralisée dans laquelle chaque tâche est élaborée et prise en charge par un sous-groupe de personnes directement intéressées par l'accomplissement de cette tâche (sous-groupe nommé « task leader »). Aucune tâche ne peut être évaluée sans task-leader. Ce sous-groupe est formé de personnes motivées, connaissant bien le domaine de la tâche et ses problématiques. Les tâches sont donc généralement bien définies et correctement évaluées. Media-Eval fournit seulement la structure nécessaire à la communication et à la synchronisation des tâches, et organise, après chaque campagne, des réunions permettant aux participants de discuter les modalités d'évaluation et les résultats obtenus. En 2012, avec Nicola Orio, Markus Schedl et Cynthia Liem nous avons organisé la tâche MusiClef [123] [100]. L'objectif de cette tâche est l'estimation de labels de genre et d'humeur (assignés à des morceaux par des professionnels) à l'aide d'informations multimodales (contenu audio, textes dérivés de pages web, tags collaboratifs utilisateurs). De manière similaire à Quaero-Eval, un environnement d'évaluation est créé et distribué aux participants. Une implémentation « de référence », illustrant une extraction multimodale type, est également fournie à des fins pédagogiques. Le modèle suivi par Media-Eval est également proche de celui d'ESTER, c.-à-d. que les participants soumettent à l'évaluateur leurs estimations sur les données tests ; ce dernier leur retourne les résultats obtenus. Les données sont distribuées aux participants après la campagne.

Session Evaluation à ISMIR-2012. Finalement, lors de la conférence ISMIR-2012, j'ai co-organisé avec Fabien Gouyon et Julian Urbano une session spéciale dédiée à l'évaluation en MIR. L'objectif de cette session était de comparer les pratiques d'évaluation actuellement utilisées en MIR à celles utilisées dans d'autres domaines. Pour cela, j'ai invité les représentants des principales campagnes d'évaluation en MIR (MIREX, MillionSongContest et MusiClef) ainsi que Gareth Jones, co-fondateur de Media-Eval. Cette session, ainsi que les discussions très intéressantes qui ont suivi, ont permis de mettre en évidence la demande d'une plus grande ouverture de MIREX, la nécessité d'une meilleure définition des tâches à accomplir, de la manière de les évaluer et finalement la nécessité de pouvoir tirer plus d'enseignements de ces évaluations. Inspiré par Media-Eval, j'ai proposé la décentralisation des évaluations MIREX à travers un modèle de « task leader », et la tenue de réunions post-évaluations permettant aux participants d'échanger. Avec Julian Urbano et Gareth Jones nous avons résumé ces discussions et contributions dans l'article [167].

2. Campagne d'évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques, voir http://www.afcp-parole.org/camp_eval_systemes_transcription/.
3. MediaEval Benchmarking Initiative for Multimedia Evaluation, voir http://www.multimediaeval.org/.

CHAPITRE 5

Conclusion et perspectives

Dans ce document, j'ai résumé les recherches que j'ai effectuées ou encadrées depuis ma soutenance de thèse en 2001. Ces recherches concernent l'indexation automatique de contenus audio et ne sont donc pas dans la continuité de mes recherches de thèse [1]. Je ne peux dès lors m'empêcher de comparer ces deux domaines dans le but de faire émerger les spécificités de l'indexation audio. Leur principale différence est sans doute l'objectif de la description recherchée. En effet, si dans le cas de la transformation sonore, l'objectif est l'obtention de paramètres permettant une modification ou une re-synthèse de haute qualité, dans le cas de l'indexation la description est l'objectif en soi. Ces domaines diffèrent également par les échelles considérées. En effet, si dans le cas de la transformation sonore, les analyses et l'évaluation de leur qualité s'effectuent à un niveau microscopique, pour l'indexation elles s'effectuent à un niveau macroscopique. Les verrous de ces deux domaines sont également différents. En effet, si dans le cas de la transformation sonore, ces verrous concernent essentiellement les algorithmes de traitement du signal, dans l'indexation audio ils incluent également la nécessaire accessibilité à des données de masses issues du monde réel et annotées : « de masse » afin d'obtenir des mesures statistiquement significatives ; « issues du monde réel » afin qu'elles soient représentatives de la complexité du monde musical. Ces données sont annotées en concepts (i.e. en « éléments significatifs »). Il reste encore un travail important à effectuer sur ces éléments significatifs pour définir précisément ce qu'ils sont pour le domaine musical. Comment définit-on la similarité, le genre, l'humeur ou encore la structure musicale ? Bien sûr, les algorithmes estimant le genre musical sans l'avoir définit au préalable sont néanmoins utiles pour le développement d'applications. Mais sans un travail sur ces définitions, il existe un danger que le domaine ne se développe que par avancées technologiques (meilleurs algorithmes de classification). La richesse de ce domaine se situe précisément dans ce croisement entre technologies et sémantique apportée aux données. Enfin une dernière spécificité concerne cette fois les algorithmes et leur application à des « masses » de données. En effet, il devient nécessaire de prendre en compte le « passage à l'échelle » des technologies (incluant des algorithmes d'estimation rapides [2] et des techniques de comparaison rapides), leur robustesse (la technologie doit pouvoir être appliquée quelles que soient les propriétés du contenu audio) ainsi que la non-possibilité de paramétrer les algorithmes pour chaque item devant être traité.

Je propose ci-dessous un ensemble de réflexions et de directions pour mes futurs travaux de recherche.

Améliorations des systèmes de classification. Faut-il rechercher de meilleurs descripteurs audio ? Des descripteurs spécifiques ? Plus sémantiques ? Faut-il orienter les recherches vers la génération automatique de descripteurs (comme le système EDS de Sony CSL [126] ou les Deep Believe Networks) ? Faut-il chercher de meilleurs algorithmes de classification ? Force est de constater que je n'ai pas de réponses à ces questions. Même s'il est intéressant de développer des descripteurs plus sémantiques (puisqu'ils amènent à une meilleure compréhension des résultats en termes acoustiques, permettent un meilleur contrôle des modèles appris, et peuvent parfois conduire à de meilleurs résultats), leur développement reste une tâche lourde et n'est pas possible sans la compréhension de la caractéristique à mettre en évidence (quelle caractéristique acoustique fait qu'un morceau est « pop » ?). Le choix de l'intégration temporelle appliquée à ces descripteurs semble en tout

1. Celles-ci concernaient le traitement du signal pour les transformations sonores de hautes qualités.

2. La version « recherche » d'estimation du tempo prend 1 min pour traiter un morceau de 4 min. Son application sur une base de 50.000 morceaux prend donc 34 jours. La version très optimisée, `ircambeat`, prend pour cela 4 s, et permet donc le traitement de ces 50.000 morceaux en 55 heures.

cas déterminant, comme l'a montré l'utilisation des intégrations UBM et ARM dans nos systèmes. Ceux-ci ont amené une amélioration significative des résultats ; au prix cependant d'une augmentation très importante de la dimensionnalité, rendant toute interprétation sémantique impossible.

Après la communauté parole, le débat sur la génération automatique de descripteurs a actuellement lieu dans la communauté MIR à travers l'utilisation des « Deep Believe Network ». Les résultats présentés en 2011 à MIREX (qui se sont cependant révélés erronés) ont contribué à lancer ce débat. Théoriquement, cette approche pourrait surpasser celle par descripteurs « manuels » ; cependant, il n'existe pas encore d'éléments quantitatifs appuyant cette hypothèse. De plus, qu'en est-il du coût de calcul et du passage à l'échelle de cette approche ?

Reste la conception de meilleurs algorithmes de classification. La frontière est cependant floue entre l'algorithme de classification et l'ensemble des outils utilisés pour la conception de meilleurs descripteurs. Dans le domaine des algorithmes de classification, je me positionne en tant qu'utilisateur. Faute de pouvoir déterminer un meilleur algorithme pour toutes les tâches, une approche de plus en plus courante est de les tester tous pour chaque tâche [110]. Mon expérience m'a cependant montré que les performances d'un même algorithme peuvent varier énormément selon son paramétrage (Ramona [173] montre bien cela dans le cas des SVMs). Il est donc important de bien connaître les propriétés des algorithmes de classification utilisés et la bonne manière de les paramétrer. Finalement, l'amélioration de ces systèmes de classification se fait très souvent par des tests sur des corpora annotés. Je ne peux que recommander d'analyser en détail les résultats de ces tests et de ne pas se fier uniquement aux scores globaux (F-measure). En effet, ces corpora, en tout cas en MIR, contiennent souvent bon nombre d'erreurs d'annotation, erreurs non apparentes dans les scores globaux (voir à ce propos l'étude récente de Sturm [207] sur les corpora de genre).

Prise en compte de la perception et de « vérités terrain » multiples. Les techniques d'indexation reposent très souvent sur la reproduction d'une « vérité terrain » unique (un genre ou un tempo est considéré comme « vérité terrain »). Celle-ci résulte de l'annotation manuelle souvent effectuée par une seule personne. L'expérience de création de corpora annotés dans le projet Quaero m'a montré que plusieurs vérités peuvent coexister, chacune avec plus ou moins de pertinence. Ces « vérités terrain » peuvent donc être multiples pour un même item. Elles peuvent également provenir d'expériences perceptives (comme les différents jugements de dissimilarité entre paires de sons des expériences de timbre) ou de techniques de foule (comme le corpus de tempo perceptif de Last-FM [98]). Dans le premier cas, les descripteurs de timbre sont obtenus à partir d'une analyse de la MDS représentant ces jugements ; ces descripteurs visent donc à reproduire un jugement *moyen* de dissimilarité. Dans le deuxième cas, seul le sous-ensemble d'items pour lesquels la perception du tempo était *partagée* a été étudiée. Mon objectif est maintenant de considérer ces « vérités terrain » multiples tant dans la conception des algorithmes d'estimation que dans leur évaluation. Cette multiplicité peut être introduite au niveau des descripteurs mais également des algorithmes d'apprentissage.

Modèle sémantique de description de contenu. Estimer les descriptions de haut niveau (telles qu'étudiées dans la partie 2 : genre musical) à partir de descriptions musicales (telles qu'étudiées dans la partie 3 : tempo, accords) est un rêve depuis le début des années 2000. C'était l'objectif d'une des publications les plus citées du domaine MIR : Tzanetakis [213]. A cette époque cependant, la qualité des descriptions musicales obtenues était telle que leur utilisation pour estimer des descriptions de haut niveau était difficile et que les résultats étaient inférieurs à ceux obtenus en utilisant directement des descriptions de bas niveau (MFCCs, Chromas). Aujourd'hui, les résultats obtenus pour ces descriptions musicales se sont grandement améliorés et leur utilisation comme étage sémantique pour l'estimation du haut niveau est envisageable. Ceci a par exemple été montré par [2]. Mon objectif sera donc l'estimation de ce haut niveau à partir de nos descriptions musicales.

J'ai montré que les technologies d'apprentissage machine (partie 2) peuvent également être utilisées pour l'estimation des descriptions musicales (partie 3). A l'inverse des technologies d'apprentissage, les descriptions de haut niveau (genre musical) ne sont pas aujourd'hui utilisées pour l'estimation de ces descriptions musicales. Mes travaux se focaliseront donc sur la liaison inverse : utiliser les informations de hauts niveaux comme contexte (probabilité a priori) pour l'estimation des descriptions musicales (les musiques de genre « pop » et « jazz » ont « a priori » des complexités d'accords très différentes).

Systèmes d'estimation en cascade et propagation d'erreurs. Aujourd'hui chaque système de description part d'une analyse du signal audio sans utilisation d'une information sémantiquement plus élevée préalablement calculée. Ainsi l'estimation d'accords s'effectue par l'extraction de Chromas (et non par l'utilisation d'une estimation de hauteurs de notes multiples) et l'estimation des battements par l'extraction d'une fonction d'énergie (et non par utilisation d'une estimation d'onsets). Ceci s'explique par le fait qu'au démarrage de ces recherches les technologies d'estimation de hauteurs de notes multiples ou de séparation de sources permettaient difficilement (du fait de leurs performances et de leur coût de calcul) leur utilisation comme base pour l'estimation de paramètres de niveau supérieur. Il en va autrement aujourd'hui. L'empilement de systèmes d'estimation nécessite cependant la prise en compte des performances de ces différents systèmes, de leurs incertitudes et de la manière dont celles-ci se propagent à travers les systèmes. Dans [130], nous avons étudié

l'influence des erreurs d'un système d'estimation de battements sur un système aval « beat-synchrone » d'estimation d'accords. Ozerov propose dans [124] une étude sur la propagation d'erreurs d'un système de séparation de sources à un système aval d'extraction de MFCCs. Mon objectif est plus globalement d'étudier la définition des fiabilités de chaque étage d'un système et la manière dont celles-ci se propagent.

Estimations jointes des descriptions de contenu. Aujourd'hui, les descriptions du contenu musical sont pour l'essentiel estimées de manières « séparées ». Dans mes recherches, j'ai étudié l'estimation « jointe » de concepts interdépendants [accords, premiers temps] [132], [battements, premiers temps] [164] ou [tonalités, accords] [133]. Dans le projet `Quaero`, j'ai récemment démarré une recherche avec Johan Pauwels et Florian Kaiser sur l'estimation jointe [structure musicale, suite d'accords]. Mon objectif est la généralisation de ces estimations « jointes ». Je propose pour cela de mettre en évidence l'ensemble des relations d'interdépendances entre les descriptions de contenus, et d'étudier la meilleure manière de formaliser ces dépendances (au niveau de leur représentation et de leur interaction). J'envisage pour cela un travail conjoint avec l'équipe Représentation Musicale de l'IRCAM spécialiste des formalismes de représentation.

ANNEXE

A. Terminologie de description des caractéristiques du rythme

Je définis ici l'ensemble des termes relatifs à la description des caractéristiques d'un rythme utilisés dans ce document. Mon but est de clarifier l'utilisation de ces termes parfois utilisés à des fins différentes.

Métrique ou chiffrage de mesure (« time signature, meter signature ») : Notation indiquée sur une partition définissant le nombre d'unités ($\#u$) dans une mesure et la durée symbolique de ces unités d_u (par exemple $\frac{\#u}{d_u} = \frac{2}{4}, \frac{3}{4}, \frac{6}{8}, \frac{9}{8}$).

Mesure : (« measure, bar ») : Niveau rythmique correspondant à la distance entre deux « barres » sur une partition. Sa durée d_m correspond à la multiplication du numérateur et du dénominateur de la métrique : $d_m = \#u \cdot d_u$.

métrique simple, composée : Dans une métrique **simple**, chaque battement b se divise en 2. Par convention, le battement est défini égal à l'unité : $b = u$. Le numérateur désigne donc le nombre de battements dans une mesure et le dénominateur la durée d'un battement. Des exemples sont 2/2, 4/4, 4/4. Dans une métrique **composée** chaque battement b se divise en 3. Par convention, le battement est défini égal à trois unités : $b = 3u$. De ce fait, le numérateur ne désigne pas le nombre de battements dans une mesure (qui est égale à $\#u/3$) et le dénominateur ne désigne pas la durée d'un battement (qui est égale à $3d_u$). Des exemples sont 6/8, 9/8, 12/8.

métrique double, triple et quadruple : spécifie le nombre de battements $\#b$ dans une mesure ($\#b$=2, 3 ou 4). Dans le cas d'une métrique simple (puisque $b = u$), les exemples sont 2/4, 3/4 et 4/4. Dans le cas d'une métrique composée (puisque $b = 3u$), les exemples sont 6/8, 9/8 et 12/8.

Pattern rythmique : Ensemble des rythmes possibles pour un tempo et une métrique donnée. Le « ChaCha-Cha », la « Samba » sont deux exemples de patterns rythmiques possibles dans une métrique 4/4.

Structure rythmique (« metrical structure, metric structure, musical meter ») : Combinaison d'un tempo, d'une métrique et d'un pattern rythmique particulier (notons cependant qu'un pattern rythmique implique généralement une métrique et se joue à des tempi préférentielles).

Tactus, battement, tempo (parfois également appelé pulsation) : Dans le cas de la musique notée il est égale à la durée du dénominateur (pour des métriques simples) ou à trois fois celle-ci (pour des métriques composées). Sa définition perceptive est donnée par le niveau rythmique correspondant à la fréquence à laquelle la majorité des être humains taperait du pied en écoutant un morceau.

Tatum (« temporal atom ») : Il correspond au niveau rythmique des évènements les plus rapides arrivant autrement qu'accidentellement (souvent la croche ou le triolet).

Mesure (« measure, bar ») : Outre la définition donnée ci-dessus, nous utilisons également (par extrapolation) ce terme pour désigner le niveau de regroupement des battements en dehors de la connaissance d'une partition. Dans ce cas ce niveau est ambigu (voir partie 3.1.2).

Structure métrique Nous désignons par structure métrique, l'ensemble des niveaux de pulsations compatibles avec le tempo et la métrique ; donc indépendamment du pattern rythmique

Nous illustrons ces terminologies sur la Figure 5.1.

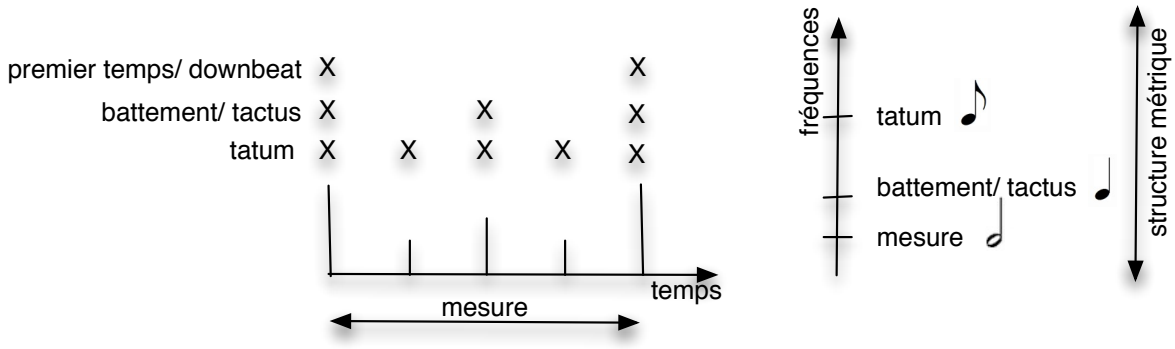

FIGURE 5.1 – Terminologie de description des caractéristiques du rythme.

Bibliographie

[1] S. Abdallah, K. Nolan, M. Sandler, M. Casey, and C. Rhodes. Theory and evaluation of a bayesian music structure extractor. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 420–425, London, UK, 2005.

[2] J. Abeßer, H. Lukashevich, C. Dittmar, and G. Schuller. Genre classification using bass-related high-level features and playing styles. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Kobe, Japan, 2009.

[3] E. Allamanche, J. Herre, O. Hellmuth, B. Froba, and M. Cremer. Audioid : Towards content-based identification of audio material. In *Proc. of AES 110th Convention*, Amsterdam, The Nederlands, 2001.

[4] C. Allauzen, M. Crochemore, and M. Raffinot. Factor oracle : A new structure for pattern matching. In *SOFSEM99 : Theory and Practice of Informatics*, pages 758–758. Springer, 1999.

[5] M. Alonso. *Extraction d'information rythmique a partir d'enregistrements musicaux.* PhD thesis, Telecom Paris, ENST, 2006.

[6] J. Andén and S. Mallat. Multiscale scattering for audio classification. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 657–662, Miami, Florida, USA, 2011.

[7] R. Andre-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1) :29–40, 1988.

[8] G. Assayag and G. Bloch. Navigating the oracle : A heuristic approach. In *Proc. of ICMC (International Computer Music Conference)*, volume 7, pages 405–412, Copenhagen, Denmark, 2007.

[9] L. Atlas and S. A. Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Appiled Signal Processing*, 7 :668–675, 2003.

[10] J. Aucouturier and F. Pachet. Finding songs that sound the same. In *Proc. of IEEE Benelux Workshop on Model based Processing and Coding of Audio*, pages 1–8, 2002.

[11] J.-J. Aucouturier and M. Sandler. Segmentation of musical signals using hidden markov models. In *Proc. of AES 110th Convention*, Amsterdam, The Netherlands, 2001.

[12] R. Badeau, B. David, and G. Richard. High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials. *IEEE Transactions on Signal Processing*, 54(4) :1341–1350, 2006.

[13] P. Balabko. Speech and music discrimination based on signal modulation spectrum. Technical report, 1999.

[14] G. Ballet, R. Borghesi, P. Hoffman, and F. Lévy. Studio online 3.0 : An internet 'killer application' for remote access to ircam sounds and processing tools. In *Proc. of JIM (Journées d'Informatique Musicale)*, Issy-Les-Moulineaux, France, 1999.

[15] M. Bartsch and G. Wakefield. To catch a chorus : Using chroma-based representations for audio thumbnailing. In *Proc. of IEEE WASPAA (Workshop on Applications of Signal Processing to Audio and Acoustics)*, pages 15–18, New Paltz, NY, USA, 2001.

[16] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[17] F. Bimbot, E. Deruty, S. Gabriel, and E. Vincent. Methodology and resources for the structural segmentation of music pieces into autonomous and comparable blocks. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[18] F. Bimbot, E. Deruty, S. Gabriel, and E. Vincent. Methodology and conventions for the latent semiotic annotation of music structure. Technical report, IRISA, 2012.

[19] F. Bimbot, L. Mathan, A. De Lima, and G. Chollet. Standard and target driven ar-vector models for speech analysis and speaker recognition. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, volume 2, pages 5–8, San Francisco, California, USA, 1992.

[20] S. Böck and M. Schedl. Enhanced beat tracking with context-aware neural networks. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Paris, France, 2011.

[21] A. Bonnefoy. Transcription automatic de la partie percussive d'un morceau de musique. Master's thesis, Université Paris VI, Master ATIAM (encadrement : A. Roebel, M. Lagrange, G. Peeters), September 2012.

[22] G. Boutard, S. Goldszmidt, and G. Peeters. Browsing inside a music track, the experimentation case study. In *Proc. of LSAS (International Workshop on Learning the Semantics of Audio Signals)*, Athens, Greece, 2006.

[23] G. Box and D. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, pages 211–252, 1964.

[24] J. Brown. Calculation of a constant q spectral transform. *JASA (Journal of the Acoustical Society of America)*, 89(1) :425–434, 1991.

[25] M. Bruderer, M. McKinney, and A. Kohlrausch. Structural boundary perception in popular music. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Victoria, Canada, 2006.

[26] J.-J. Burred and G. Peeters. An adaptive system for music classification and tagging. In *Proc. of LSAS (International Workshop on Learning the Semantics of Audio Signals)*, Graz, Austria, 2009.

[27] J. J. Burred, A. Roebel, and T. Sikora. Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. *Audio, Speech and Language Processing, IEEE Transactions on*, 18(3) :663–674, 2010.

[28] P. Cano, E. Batlle, H. Mayer, and H. Neuschmied. Robust sound modeling for song detection in broadcast audio. *Proc. of AES 112th Convention*, pages 1–7, 2002.

[29] C. Cao and M. Li. Thinkit's submissions for mirex2009 audio music classification and similarity tasks. In *MIREX (Extended Abstract)*, Kobe, Japan, 2009.

[30] M. Carey, E. Paris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Phoenix, Arizona, USA, 1999.

[31] G. Carpentier. *Computational Approach of Musical Orchestration : Constrained Multiobjective Optimization in Large Sound Databases*. PhD thesis, Université Paris VI, 2008.

[32] G. Carpentier, D. Tardieu, and A. Gérard. An evolutionary approach to computer-aided orchestration. In Springer-Verlag, editor, *Proc. of EvoMUSART*, pages 488–498, 2007.

[33] M. Casey and M. Slaney. Song intersection by approximate nearest neighbor search. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, volume 6, pages 144–149, Victoria, Canada, 2006.

[34] W. Chai. Semantic segmentation and summarization of music : Methods based on tonality and recurrent structure. *IEEE Signal Processing Magazine*, 23(2) :124–132, 2006.

[35] C. Charbuillet, G. Peeters, S. Barton, and V. Gouet-Brunet. A fast algorithm for music search by similarity in large databases based on modified symetrized kullback-leibler divergence. In *Proc. of IEEE CBMI (International Workshop on Content-Based Multimedia Indexing)*, pages 1–6, Grenoble, France, June 2010.

[36] C. Charbuillet, D. Tardieu, and G. Peeters. Gmm supervector for content based music similarity. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, pages 425–428, Paris, France, September 2011.

[37] W. Chou and L. Gu. Robust singing detection in speech/music discriminator design. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Salt Lake City, Utah, USA, 2001.

[38] A. Cont. Antescofo : Anticipatory synchronization and control of interactive parameters in computer music. international computer music conference. In *Proc. of ICMC (International Computer Music Conference)*, Belfast, Ireland, 2008.

[39] M. Cooper and J. Foote. Automatic music summarization via similarity analysis. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 81–85, Paris, France, 2002.

[40] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In *Proc. of IEEE WASPAA (Workshop on Applications of Signal Processing to Audio and Acoustics)*, New Paltz, NY, USA, 2003.

[41] A. Derenzweig and D. Ellis. Locating singing voice segments within music signals. In *Proc. of IEEE WASPAA (Workshop on Applications of Signal Processing to Audio and Acoustics)*, New Paltz, NY, USA, 2001.

[42] S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1) :39–58, 2001.

[43] S. Dubnov, G. Assayag, and A. Cont. Audio oracle : A new algorithm for fast learning of audio structures. In *Proc. of ICMC (International Computer Music Conference)*, Copenhagen, Denmark, 2007.

[44] J. Eckmann, S. Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *EPL (Europhysics Letters)*, 4 :973, 1987.

[45] D. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in music artist similarity. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 170–177, Paris, France, 2002.

[46] T. En-Najjary, O. Rosec, and T. Chonavel. A new method for pitch prediction from spectral envelope and its application in voice conversion. In *Proc. of Eurospeech*, Geneva, 2003.

[47] P. Esling. *Analyse multi-objective des séries temporelles"*. PhD thesis, Université Paris VI, 2012.

[48] P. Esling, C. Grégoire, and A. Carlos. Dynamic musical orchestration using genetic algorithms and a spectro-temporal description of musical instruments. In *Lecture Notes in Computer Science*, volume 6025, EvoApplications Part II. Springer-Verlag, 2010.

[49] S. Essid. *Classification automatique des signaux audio-frequences : reconnaissance des instruments de musique*. Phd thesis, Télécom Paris-Tech, Paris, France, 2005.

[50] H. Fastl and E. Zwicker. *Psychoacoustics : Facts And Models*, page 207. Springer, 3rd ed. edition, 2007.

[51] S. Fenet, G. Richard, Y. Grenier, et al. A scalable audio fingerprint method with robustness to pitch-shifting. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[52] D. Fitzgerald. Harmonic/percussive separation using median filtering. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Graz, Austria, 2010.

[53] P. Flandrin. *Time-Frequency/Time-Scale Analysis*. Academic Press, San Diego, California, 1999.

[54] J. Flocon-Cholet. Estimation du tempo percetif et diminution des erreurs d'octave. Master's thesis, Université Paris VI, Master ATIAM (encadrement : G. Peeters), September 2012.

[55] J. Foote. Visualizing music and audio using self-similarity. In *Proc. of ACM Multimedia*, pages 77–80, Orlando, Florida, USA, 1999.

[56] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proc. of IEEE ICME (International Conference on Multimedia and Expo)*, pages 452–455, New York City, NY, USA, 2000.

[57] K. Fort. *Vers une methodologie de l'annotation manuelle de corpus*. PhD thesis, Université Paris 13, Villetaneuse, France, 2012.

[58] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 329–336, London, UK, 2005. Citeseer.

[59] T. Fujishima. Realtime chord recognition of musical sound : a system using common lisp music. In *Proc. of ICMC (International Computer Music Conference)*, pages 464–467, Bejing, China, 1999.

[60] T. Galas and X. Rodet. Generalized functional approximation for source-filter system modeling. In *Proc. of Eurospeech*, pages 1085–1088, Genova, Italy, 1991.

[61] O. Gillet and G. Richard. Enst-drums : an extensive audio-visual database for drum signals processing. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 156–159, Victoria, BC, Canada, 2006.

[62] E. Gomez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing, Special Cluster on Computation in Music*, 18(3), 2006.

[63] M. Goto. A chorus-section detecting method for musical audio signals. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 437–440, Hong Kong, China, 2003.

[64] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database : Popular, classical, and jazz music databases. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages pp. 287–288, Paris, France, 2002.

[65] M. Goto and Y. Muraoka. Real-time beat tracking for drumless audio signals : Chord change detection for musical decisions. *Speech Communication*, 27 :311–335, 1999.

[66] F. Gouyon and S. Dixon. A review of rhythm description systems. *Computer Music Journal*, 29(1) :34–54, 2005.

[67] F. Gouyon, L. Fabig, and J. Bonada. Rhythmic expressiveness transformations of audio recordings : Swing modifications. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, London, UK, 2003.

[68] J.-B. Goyeau. *Descripteurs et algorithmes de caracterisation de l'aspect rythmiques du son et de la musique*. Master thesis, Master ATIAM, Université Paris VI, 2004.

[69] J. M. Grey and J. W. Gordon. Perceptual effects of spectral modifications on musical timbres. *JASA (Journal of the Acoustical Society of America)*, 63(5) :1493–1500, 1978.

[70] E. Guaus. *Audio content processing for automatic music genre classification : descriptors, databases, and classifiers*. Phd thesis, Univsersitat Pompeu Fabra, 2009.

[71] S. Hainsworth. *Techniques for the automated analysis of musical audio*. Phd thesis, Cambridge University, 2004.

[72] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Paris, France, 2002.

[73] M. Hall. Feature selection for discrete and numeric class machine learning. Working paper 99/04, Hamilton, New Zealand : University of Waikato, Department of Computer Science., 1999.

[74] C. Harte, M. Sandler, S. Abdallah, and E. Gomez. Symbolic representation of musical chords : A proposed syntax for text annotations. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 66–71, London, UK, 2005.

[75] P. Herrera, O. Celma, J. Massaguer, P. Cano, E. Gomez, F. Gouyon, M. Koppenberger, D. Garcia, J.-P. Garcia, and N. Wack. Mucosa : A music content semantic annotator. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, London, UK, 2005.

[76] P. Herrera, A. Klapuri, and M. Davy. Automatic classification of pitched musical instrument sounds. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*. Springer-Verlag, New York, 2006.

[77] P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1) :3–21, 2003.

[78] I. International Music Information Retrieval Systems Evaluation Laboratory. Introducing m2k and d2k. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Barcelona (Spain), 2004.

[79] O. Izmirli. Template based key finding from audio. In *Proc. of ICMC (International Computer Music Conference)*, pages 211–214, Barcelona, Spain, 2005.

[80] D. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai. Music type classification by spectral contrast. In *Proc. of IEEE ICME (International Conference on Multimedia and Expo)*, Lausanne Switzerland, 2002.

[81] C. Joder, S. Essid, and G. Richard. Temporal integration for audio classification with application to musical instrument classification. *Audio, Speech and Language Processing, IEEE Transactions on*, 17(1) :174–186, 2009.

[82] J. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2) :314–323, 1988.

[83] F. Kaiser. Multi-probe histograms : A mid-level harmonic feature for music structure segmentation. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Paris, France, 2011.

[84] F. Kaiser and G. Peeters. Adaptive temporal modeling of audio features in the context of music structure segmentation. In *Proc. of AMR (Adaptive Multimedia Retrieval)*, Copenhagen, Denmark, October 2012.

[85] F. Kaiser and G. Peeters. Multiple hypotheses at multiple scales for audio novelty computation within music. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Vancouver, British Columbia, Canada, May 2013.

[86] F. Kaiser and T. Sikora. Music structure discovery in popular music using non-negative matrix. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Utrecht, The Netherlands, 2010.

[87] A. Klapuri. Automatic transcription of music. Master thesis, Tampere University of Technology, Tampere, Finland, 1997.

[88] A. Klapuri, A. Eronen, and J. Astola. Analysis of the meter of acoustic musical signals. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(1) :342–355, 2006.

[89] I. Kononenko. Estimating attributes : Analysis and extensions of relief. In *Proc. of ECML (European Conf. on Machine Learning)*, pages 171–182, Catania, 1994.

[90] J. Krimphoff, S. McAdams, and S. Windsberg. Caractérisation du timbre des sons complexes. ii : Analyses acoustiques et quantification psychophysique. *Journal de physique*, 4 :625–628, 1994.

[91] C. Krumhansl. Why is musical timbre so hard to understand. In S. Nielzen and O. Olsson, editors, *Structure and perception of electroacoustic sound and music*, volume 9, pages 43–53. Elsevier, Amsterdam (Expcerpta Medica 846) edition., 1989.

[92] C.-L. Krumhansl. *Cognitive foundations of musical pitch.* Oxford University Press, New-York, 1999.

[93] A. Laburthe. *Résumé sonore.* Master thesis, INPG, Université Joseph Fourier, Grenoble, France, 2002.

[94] H. Lachambre. *Caractérisation de l'environnement musical dans les documents audiovisuels.* Phd thesis, Université de Toulouse, December 2009.

[95] S. Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics*, 2000.

[96] J. Laroche. Efficient tempo and beat tracking in audio recordings. *JAES (Journal of the Audio Engineering Society)*, 51(4) :226–233, 2003.

[97] A. Lenoir, R. Landais, G. Peeters, L. Oudre, and T. Fillon. Muma : A music search engine based on content analysis. In *Proc. of IEEE ICME (International Conference on Multimedia and Expo)*, Barcelona, Spain, July 2011.

[98] M. Levy. Improving perceptual tempo estimation with crowd-sourced annotations. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[99] Y. Li and D. Wang. Separation of singing voice from music accompaniment for monaural recordings. *Audio, Speech and Language Processing, IEEE Transactions on*, 15(4) :1475–1487, 2007.

[100] C. C. S. Liem, N. Orio, G. Peeters, and M. Schedl. Brave new task : Musiclef multimodal music tagging. In *Proc. of MediaEval (Multimedia Benchmark Workshop)*, Pisa, Italy, October 2012.

[101] A. Livshin and X. Rodet. The importance of cross database evaluation in musical instrument sound classification. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 241–242, Baltimore, Maryland, USA, 2003.

[102] B. Logan and S. Chu. Music summarization using key phrases. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, volume II, pages 749–752, Istanbul, Turkey, 2000.

[103] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proc. of IEEE ICME (International Conference on Multimedia and Expo)*, 2001.

[104] N. Maddage, C. Xu, and Y. Wang. Singer identification based on vocal and instrumental models. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 375–378. IEEE, 2004.

[105] K. Martin. *Sound source recognition : a theory and computational model.* Phd thesis, MIT (Massachusetts Institute of Technology), Cambridge, Massachusetts, United States, June 1999.

[106] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Klozali, D. Tidhar, and M. Sandler. Omras2 metadata project 2009. In *Proc. of ISMIR (Late-Breaking News)*, Kobe, Japan, 2009.

[107] M. Mauch and S. Dixon. Simultaneous estimation of chords and musical context from audio. *Audio, Speech and Language Processing, IEEE Transactions on*, 18(6) :1280 – 1289, 2010.

[108] S. McAdams, S. Windsberg, S. Donnadieu, G. DeSoete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres : common dimensions specificities and latent subject classes. *Psychological Research*, 58 :177–192, 1995.

[109] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle. jaudio : A feature extraction library. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, London, UK, 2005.

[110] C. McKay, R. Fiebrink, D. McEnnis, B. Li, and I. Fujinaga. Ace : A framework for optimizing music classification. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, London, UK, 2005.

[111] M. McKinney and J. Breebaart. Features for audio and music classification. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Baltimore, Maryland, USA, 2003.

[112] P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116 :91–103, 1976.

[113] A. Mesaros, T. Virtanen, and A. Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 375–378, Vienna, Austria, 2007.

[114] N. Misdariis, B. Smith, D. Pressnitzer, P. Susini, and S. McAdams. Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres. In *Proc. of 135th Meet. Ac. Soc. of America / 16th Int. Cong. on Acoustics*, Seattle, Washington, USA, 1998.

[115] F. Mislin, M. Fingerhut, and G. Peeters. Automatisation de la production et de la mise en ligne de resumes sonores. Master's thesis, ISTY, 2005.

[116] D. Moelants and M. F. McKinney. Tempo perception and musical content : What makes a piece slow, fast, or temporally ambiguous ? In *Proc. of ICMPC (International Conference of Music Perception and Cognition)*. Evanston, IL, 2004.

[117] L. Molina, L. Belanche, and A. Nebot. Feature selection algorithms : A survey and experimental evaluation. In *Proc. of ICDM (IEEE Int. Conf. on Data Mining)*, Maebashi City, Japan, Dec. 2002.

[118] MPEG-7. Information technology - multimedia content description interface - part 4 : Audio (iso/iec 15938-4), 2002.

[119] M. Müller and S. Ewert. Joint structure analysis with applications to music annotation and synchronization. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Philadelphia, PA, USA, 2008.

[120] M. Müller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *Audio, Speech and Language Processing, IEEE Transactions on*, 18(3) :649–662, 2010.

[121] M. Müller, P. Grosche, and N. Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[122] M. Müller, N. Jiang, and P. Grosche. A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *Audio, Speech and Language Processing, IEEE Transactions on*, To Appear in 2012.

[123] N. Orio, C. C. S. Liem, G. Peeters, and M. Schedl. Musiclef : Multimodal music tagging task. In *Proc. of CLEF (Conference on Multilingual and Multimodal Information Access Evaluation)*, Roma, Italy, September 2012.

[124] A. Ozerov, M. Lagrange, and E. Vincent. Gmm-based classification from noisy features. In *1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011.

[125] F. Pachet and P. Roy. Markov constraints : Steerable generation of markov sequences. *Constraints*, 16(2) :148–172, 2011.

[126] F. Pachet and A. Zils. Automatic extraction of music descriptors from acoustic signals. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Barcelona (Spain), 2004.

[127] E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 201–208, Baltimore, Maryland, USA, 2003.

[128] E. Pampalk, A. Flexer, G. Widmer, et al. Improvements of audio-based music similarity and genre classification. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, volume 5, London, UK, 2005.

[129] H. Papadopoulos. *Extraction automatique d'une suite d'accords a partir de l'analyse d'un signal audio musical*. Master sic (system intelligents et communicants) / diplome d'ingenieur, Université de Cergy-Pontoise / ENSEA, 2006.

[130] H. Papadopoulos. *Joint Estimation of Musical Content Information*. Phd thesis, University Paris VI, 2010.

[131] H. Papadopoulos and G. Peeters. Large-scale study of chord estimation algorithms based on chroma representation. In *Proc. of IEEE CBMI (International Workshop on Content-Based Multimedia Indexing)*, Bordeaux, France, 2007.

[132] H. Papadopoulos and G. Peeters. Joint estimation of chords and downbeats from an audio signal. *Audio, Speech and Language Processing, IEEE Transactions on*, 19(1) :138 – 152, January 2010.

[133] H. Papadopoulos and G. Peeters. Local key estimation from an audio signal relying on harmonic and metrical structures. *Audio, Speech and Language Processing, IEEE Transactions on*, 20(4) :1297 – 1312, May 2012.

[134] J. Paulus and A. Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *Audio, Speech and Language Processing, IEEE Transactions on*, 17(6) :1159–1170, 2009.

[135] J. Pauwels and J.-P. Martens. Integrating musicological knowledge into a probabilistic system for chord and key extraction. In *Proc. AES 128th Conv*, London, UK, 2010.

[136] J. Pauwels and G. Peeters. Evaluating automatically estimated chord sequences. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Vancouver, British Columbia, Canada, May 2013.

[137] G. Peeters. *Modèles et modèlisation du signal sonore adaptés à ses caractéristiques locales.* Phd thesis, Université Paris VI, July 2001.

[138] G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Proc. of AES 115th Convention*, New York, NY, USA, 2003.

[139] G. Peeters. Method for processing an audio sequence for example a piece of music. Patent Audio summary (FR04/01493, 2004/06/16), Europe 04767355.3, Japan 516296/2006, US 2006/0288849 A1, 2003.

[140] G. Peeters. *Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation : Sequence and State Approach*, pages 142–165. Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg 2004, 2004.

[141] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado project report, Ircam, 2004.

[142] G. Peeters. Chroma-based estimation of musical key from audio-signal analysis. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 115–120, Victoria, BC, Canada, 2006.

[143] G. Peeters. Music pitch representation by periodicity measures based on combined temporal and spectral representations. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, volume V, pages 53–56, Toulouse, France, 2006.

[144] G. Peeters. Musical key estimation of audio signal based on hmm modeling of chroma vectors. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Montreal, Canada, 2006.

[145] G. Peeters. A generic system for audio indexing : application to speech/ music segmentation and music genre. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Bordeaux, France, 2007.

[146] G. Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Vienna, Austria, 2007.

[147] G. Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Applied Signal Processing*, 2007(1) :158–158, 2007. doi :10.1155/2007/67215.

[148] G. Peeters. Template-based estimation of tempo : using unsupervised or supervised learning to create better spectral templates. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, pages 209–212, Graz, Austria, September 2010.

[149] G. Peeters. Copy and scale method for doing time-localized m.i.r. estimation : Application to beat-tracking. *Journal of New Music Research*, 40 (Special issue on Music and Machine Learning)(2) :153–164, June 2011.

[150] G. Peeters. Music structure discovery : measuring the 'state-ness' of times. In *Proc. of ISMIR (Late-Breaking News)*, Miami, Florida, USA, October 2011.

[151] G. Peeters. Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal. *Audio, Speech and Language Processing, IEEE Transactions on*, 19(5) :1242–1252, July 2011.

[152] G. Peeters, F. Cornu, D. Tardieu, C. Charbuillet, J. J. Burred, M. Ramona, M. Vian, V. Botherel, J.-B. Rault, and J.-P. Cabanal. A multimedia search and navigation prototype, including music and video-clips. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Porto, Portugal, October 2012.

[153] G. Peeters and E. Deruty. Automatic morphological description of sounds. In *Proc. of Acoustics08*, Paris, France, 2008.

[154] G. Peeters and E. Deruty. Is music structure annotation multi-dimensional ? a proposal for robust local music annotation. In *Proc. of LSAS (International Workshop on Learning the Semantics of Audio Signals)*, Graz, Austria, 2009.

[155] G. Peeters and E. Deruty. Toward music structure annotation. In *Proc. of ISMIR (Late-Breaking News)*, Kobe, Japan, 2009.

[156] G. Peeters and E. Deruty. Sound indexing using morphological description. *Audio, Speech and Language Processing, IEEE Transactions on*, 18(3 (Special issue on Signal Models and Representations of Musical and Environmental Sound)) :675–687, March 2010.

[157] G. Peeters, D. Fenech, and X. Rodet. Mcipa : A music content information player and annotator for discovering music. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Philadelphia, PA, USA, 2008.

[158] G. Peeters and J. Flocon-Cholet. Perceptual tempo estimation using gmm regression. In *Proc. of ACM Multimedia/ MIRUM (Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies)*, Nara, Japan, November 2012.

[159] G. Peeters and K. Fort. Towards a (better) definition of the description of annotated m.i.r. corpora. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Porto, Portugal, October 2012.

[160] G. Peeters, B. Giordano, P. Susini, N. Misdariis, and S. McAdams. The timbre toolbox : Extracting audio descriptors from musical signals. *JASA (Journal of the Acoustical Society of America)*, 130(5), November 2011.

[161] G. Peeters, A. Laburthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 94–100, Paris, France, 2002.

[162] G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of mpeg-7. In *Proc. of ICMC (International Computer Music Conference)*, pages 166–169, Berlin, Germany, 2000.

[163] G. Peeters, S. McAdams, J. Krimphoff, P. Susini, N. Misdariis, and B. Smith. Method for characterizing the timbre of a sound signal in accordance with at least a descriptor. Brevet FR2830118 (26.09.01) / Patent 20040220799 / U.S. Patent Number : 7,406,356, 2001.

[164] G. Peeters and H. Papadopoulos. Simultaneous beat and downbeat-tracking using a probabilistic framework : theory and large-scale evaluation. *Audio, Speech and Language Processing, IEEE Transactions on*, 19(6) :1754–1769, August 2011.

[165] G. Peeters and X. Rodet. Automatically selecting signal descriptors for sound classification. In *Proc. of ICMC (International Computer Music Conference)*, pages 455–458, Göteborg, Sweden, 2002.

[166] G. Peeters and X. Rodet. Signal-based music structure discovery for music audio summary generation. In *Proc. of ICMC (International Computer Music Conference)*, Singapore, Singapore, 2003. Peeters03e.

[167] G. Peeters, J. Urbano, and G. Jones. Notes from the ismir 2012 late-breaking session on evaluation in music information retrieval. In *Proc. of ISMIR (Late-Breaking News)*, Porto, Portugal, October 2012.

[168] F. Pellegrino and R. Andre-Obrecht. Automatic language identification : an alternative approach to phonetic modelling. *Signal Processing*, 80(7) :1231–1244, 2000.

[169] J. Pinquier and R. André-Obrecht. Audio indexing : primary components retrieval. *Multimedia Tools and Applications*, 30(3) :313–330, 2006.

[170] J. Pinquier and R. André-Obrecht. Audio indexing : Primary components retrieval - robust classification in audio documents. *Multimedia Tools and Applications*, 30(3) :313–330, 2006.

[171] T. Pohle and D. Schnitzer. Striving for an improved audio similarity measure. *4th annual music information retrieval evaluation exchange*, 2007.

[172] D. Psenicka. Sporch : An algorithm for orchestration based on spectral analyses of recorded sounds. In *Proc. of ICMC (International Computer Music Conference)*, Singapore, Singapore, 2003.

[173] M. Ramona. *Classification automatique de flux radiophoniques par Machines a Vecteurs de Support*. Phd thesis, Télécom Paris-Tech, Paris, France, 2010.

[174] M. Ramona, S. Fenet, R. Blouet, H. Bredin, T. Fillon, and G. Peeters. A public audio identification evaluation framework for broadcast monitoring. *Journal of Experimental and Theoretical Artificial Intelligence (Special Issue on Event Recognition)*, 26(1-2) :119–136, February 2012.

[175] M. Ramona and G. Peeters. Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 477 – 480, Prague, Czech Republic, May 2011.

[176] M. Ramona and G. Peeters. Automatic verification and high-precision alignment of audio fingerprinting annotations. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, pages 429–436, Paris, France, September 2011.

[177] M. Ramona and G. Peeters. Audioprint : An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Vancouver, British Columbia, Canada, May 2013.

[178] M. Ramona, G. Richard, and B. David. Vocal detection in music with support vector machines. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 1885–1888, Las Vegas, Nevada, USA, 2008.

[179] L. Régnier. Détection de la voix chantée dans un morceau de musique. Master thesis, Université Paris VI, 2008.

[180] L. Régnier. *Localization, Characterization and Recognition of Singing Voices*. PhD thesis, Université Paris VI, Paris, France, March 2012.

[181] L. Régnier and G. Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 1685 – 1688, Taipei, Taiwan, April 2009.

[182] L. Régnier and G. Peeters. Partial clustering using a time-varying frequency model for singing voice detection. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 441–444, Dallas, Texas, USA, March 2010.

[183] L. Régnier and G. Peeters. Combining classifications based on local and global features : application to singer identification. In G. Peeters, editor, *Proc. of DAFx (International Conference on Digital Audio Effects)*, pages 127–134, Paris, France, September, 19-23 2011. IRCAM - Centre Pompidou.

[184] L. Régnier and G. Peeters. Singer verification : singer model .vs. song model. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Kyoto, Japan, March 2012.

[185] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3) :19–41, 2000.

[186] J. Ricard and P. Herrera. Morphological sound description : Computational model and usability evaluation. In *Proc. of AES 116th Convention*, Berlin, Germany, 2004.

[187] F. Rigaud, M. Lagrange, A. Roebel, and G. Peeters. Drum extraction from polyphonic music based on a spectro-temporal model of percussive sounds. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, pages 381 – 384, Prague, Czech Republic, May 2011.

[188] T. Rocher, M. Robine, P. Hanna, L. Oudre, et al. Concurrent estimation of chords and keys from audio. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Utrecht, The Netherlands, 2010.

[189] X. Rodet, L. Worms, and G. Peeters. Method for characterizing a sound signal. US 2005/0163325 A1 / EP 1459214 A1 / JP 2005-513576 A / WO 2003/056455, 2003.

[190] A. Roebel. Frequency-slope estimation and its application to parameter estimation for non-stationary sinusoids. *Computer Music Journal*, 32(2) :68–79, 2008.

[191] A. Roebel and X. Rodet. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Madrid, Spain, 2005.

[192] F. Rose and J. Hetrick. Spectral analysis as a ressource for contemporary orchestration technique. In *Proceedings of Conference on Interdisciplinary Musicology*, volume 2005, 2005.

[193] S. Rossignol. *Segmentation et indexation des signaux sonores musicaux*. Phd thesis, Université Paris VI, 2000.

[194] J. Salamon, G. Peeters, and A. Roebel. Statistical characterisation of melodic pitch contours and its application for melody extraction. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Porto, Portugal, October 2012.

[195] G. Sargent, F. Bimbot, and E. Vincent. A regularity-constrained viterbi algorithm and its application to the structural segmentation of songs. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[196] J. Saunders. Real-time discrimination of broadcast speech/ music. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, Atlanta, Georgia, USA, 1996.

[197] P. Schaeffer. *Traité des objets musicaux*. Seuil, Paris, 1966.

[198] E. Scheirer. Tempo and beat analysis of acoustic musical signals. *JASA (Journal of the Acoustical Society of America)*, 103(1) :588–601, 1998.

[199] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/ music discriminator. In *Proc. of IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing)*, volume 2, pages 1331–1334, Munich, Bavaria, Germany, 1997.

[200] E. Scheirer, R. Vaananen, and J. Huopaniemi. Audiobifs : Describing audio scenes with the mpeg-4 multimedia standard. *IEEE Transactions on Multimedia*, 1(3) :237–250, 1999.

[201] E. Scheirer and B. Vercoe. Saol : the mpeg-4 structured audio orchestra language. *Computer Music Journal*, 2, 23.

[202] D. Schnitzer. *Indexing Content-Based Music Similarity Models for Fast Retrieval in Massive Databases.* PhD thesis, Johannes Kepler Universität, Linz, http ://www.schnitzer.at/dominik/phd-thesis/, 2012.

[203] D. Schwarz, G. Beller, B. Verbrugghe, and B. S. Real-time corpus-based concatenative synthesis with catart. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Montreal, Canada, 2006.

[204] X. Serra and J. Bonada. Sound transformations based on sms high level attributes. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Barcelona, Spain, 1998.

[205] K. Seyerlehner. *Content-Based Music Recommender Systems : Beyond simple Frame-Level Audio Similarity.* PhD thesis, Johannes Kepler Universität, Linz, Austria, December 2010.

[206] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie. Design and creation of a large-scale database of structural annotations. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[207] B. Sturm. Two systems for automatic music genre recognition : What are they really recognizing ? In *Proc. of ACM Multimedia/ MIRUM (Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies)*, Nara, Japan, 2012.

[208] D. Tardieu. Transformation par descripteurs de haut-niveau et perceptifs. Master thesis, Mater ATIAM, Université Paris VI, 2004.

[209] D. Tardieu. *Modèles d'instrument pour l'aide a l'orchestration.* Phd thesis, Université Paris VI, 2008.

[210] D. Tardieu, C. Charbuillet, F. Cornu, and G. Peeters. Mirex-2011 single-label and multi-label classification tasks : Ircamclassification2011 submission. In *MIREX (Extended Abstract)*, 2011.

[211] D. Tardieu, C. Charbuillet, F. Cornu, and G. Peeters. Mirex-2011 single-label and multi-label classification tasks : Ircamclassification2011 submission. In *MIREX (Extended Abstract)*, Miami, USA, 2011.

[212] G. Tzanetakis and P. Cook. Marsyas : a framework for audio analysis. *Organised Sound*, 4(3), 1999.

[213] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5) :293–302, 2002.

[214] H. Vinet, G. Assayag, J.-J. Burred, G. Carpentier, N. Misdariis, G. Peeters, A. Roebel, N. Schnell, D. Schwarz, and D. Tardieu. Sample orchestrator : gestion par le contenu d'échantillons sonores. *Traitement du signal*, 2012.

[215] G. Wakefield. Mathematical representation of joint time-chroma distributions. In *Proc. of SPIE conference on Advanced Signal Processing Algorithms, Architecture and Implementations*, pages 637–645, Denver, Colorado, USA, 1999.

[216] A. L.-C. Wang. An industrial strength audio search algorithm. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Baltimore, Maryland, USA, 2003.

[217] B. Whitman and D. Ellis. Automatic record reviews. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Barcelona (Spain), 2004.

[218] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Classification, search and retrieval of audio. In B. Furth, editor, *CRC Handbook of Multimedia Computing*, pages 207–226. CRC Press, Boca Raton, FLA, 1999.

[219] L. Worms. *Reconnaissance d'extraits sonores dans une large base de donnees.* Practical lessons, Ircam, 1998.

[220] A. Wronecki. Application de l'algorithme de dtw dans la detection de structures musicales par approche de type sequence. Master's thesis, Universite Paris XII Val de Marne, 2005.

[221] C. Yeh, N. Bogaards, and A. Roebel. Synthesized polyphonic music database with verifiable ground truth for multiple f0 estimation. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, pages 393–398, Vienna, Austria, 2007.

[222] C. Yeh, A. Roebel, and X. Rodet. Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. In *Audio, Speech and Language Processing, IEEE Transactions on*, volume 18, page 6, 2010.

[223] Y. Yu, M. Crucianu, V. Oria, and E. Damiani. Combining multi-probe histogram and order-statistics based lsh for scalable audio content retrieval. In *Proc. of ACM Multimedia*, pages 381–390, Florence, Italy, 2010.

[224] T. Zhang. Automatic singer identification. In *Proc. of IEEE ICME (International Conference on Multimedia and Expo)*, 2003.

[225] E. Zwicker and E. Terhardt. Analytical expression for critical-band rate and critical bandwidth as a function of frequency. *JASA (Journal of the Acoustical Society of America)*, 68 :1523–1525, 1980.