Un système interactif pour l'analyse des musiques électroacoustiques

Sébastien Gulluni

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-00xx

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Sébastien GULLUNI

le 20 Décembre 2011

Un système interactif pour l'analyse des musiques électroacoustiques

Directeur de thèse : **Gaël RICHARD**

Jury

Mme Myriam DESAINTE CATHERINE	Rapporteur
Mme Anne SÈDES	Rapporteur
M. Michel CRUCIANU	Examinateur
M. Pierre COUPRIE	Examinateur
M. Olivier BUISSON	Encadrant industriel
M. Slim ESSID	Encadrant académique

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

T H È S E

Remerciements

Je tiens à remercier tout d'abord mon directeur de thèse Gaël Richard pour avoir cru en l'intérêt de ce travail qui repose en grande partie sur une application très spécifique ainsi que pour tous les conseils apportés pendant ces années de thèse.

Je remercie également Olivier Buisson et Slim Essid pour le solide encadrement scientifique qu'ils m'ont apporté tout au long de cette thèse ainsi que pour leur grande disponibilité.

Mes remerciements vont également à Emmanuel Favreau pour son encadrement à la fois scientifique et applicatif qui fut très utile pour me permettre de garder en tête les contraintes d'utilisation du système. Merci également à Marie-Luce Viaud pour avoir suivi de près le déroulement de la thèse ainsi que pour son esprit critique.

Je tiens également à remercier Pierre Couprie, François Delalande et Cyrille Delhaye d'avoir accepté de participer aux entretiens réalisés au sujet des pratiques d'analyse des musiques électroacoustiques. De même, je remercie Evelyne Gayou et Yann Geslin pour leurs critiques sur le chapitre concernant les musiques électroacoustiques. Merci également à Alexandre Bazin, Diego Losa et Daniel Teruggi pour m'avoir fourni des "sons électroacoustiques" pertinents m'ayant permis de réaliser mon corpus synthétique. Je remercie également Adrien Lefèvre pour ses conseils experts en développement ainsi que Dominique Saint Martin pour son approche du métier et son goût du débat musical.

Le fait de travailler dans trois lieux différents pendant trois années multiplie forcément les camarades de bureau que je tiens à saluer. Au GRM : Sébastien R., Michael, François, Sébastien M., Antonin, Orianne, Eric, Pierre-Marie, Nicolas, Julien et Raphaël. A l'INA : Hervé, Benjamin, Pierre et Clément. Enfin, à TSI : Benoît, Félicien et François. Je salue également l'équipe de production du GRM, Philippe et François, pour la causticité légendaire de leur humour et pour leur bureau/musée fourni en jouets pour musiciens (quand je pense à ce MS-20 qui prend la poussière...). Je tiens également à remercier toute l'équipe du GRM pour cette passion qu'ils ont pour la musique qui m'a permis d'enrichir mon approche personnelle.

Enfin, je remercie infiniment mes proches qui m'ont toujours soutenu durant ces années de thèse et ont su faire preuve d'empathie dans les moments difficiles.

Résumé

Les musiques électroacoustiques sont encore aujourd'hui relativement peu abordées dans les recherches qui visent à retrouver des informations à partir du contenu musical. La plupart des travaux de recherche concernant ces musiques sont centrés sur les outils de composition, la pédagogie et l'analyse musicale. Dans ce travail de thèse, nous nous intéressons aux problématiques scientifiques liées à l'analyse des musiques électroacoustiques. Après avoir replacé ces musiques dans leur contexte historique, une étude des pratiques d'analyse de trois professionnels nous permet de dégager des invariants pour l'élaboration d'un système d'analyse. Ainsi, nous proposons un système interactif d'aide à l'analyse des musiques électroacoustiques qui permet de retrouver les différentes instances des objets sonores composant une pièce polyphonique. Le système proposé permet dans un premier temps de réaliser une segmentation afin de dégager les instances initiales des objets sonores principaux. L'utilisateur peut ainsi sélectionner les objets qu'il vise avant de rentrer dans une boucle d'interaction qui utilise l'apprentissage actif et le retour de pertinence fourni par l'utilisateur. Le retour apporté par l'utilisateur est utilisé par le système qui réalise une classification multilabel des différents segments sonores en fonction des objets sonores visés. Une évaluation par simulation utilisateur est réalisée à partir d'un corpus de pièces synthétiques. L'évaluation montre que notre approche permet d'obtenir des résultats satisfaisants en un nombre raisonnable d'interactions.

Mots-clés : musiques électroacoustiques, apprentissage interactif, retour de pertinence, apprentissage actif, classification multilabel.

Abstract

Electro-acoustic music is still hardly studied in the field of Music Information Retrieval. Most research on this type of music focuses on composition tools, pedagogy and music analysis. In this thesis, we focus on scientific issues related to the analysis of electro-acoustic music. After placing this music into historical context, a study of the practices of three professional musicologist allows us to obtain guidelines for building an analysis system. Thus, we propose an interactive system for helping the analysis of electro-acoustic music that allows one to find the various instances of the sound objects of a polyphonic piece. The proposed system first performs a segmentation to identify the initial instances of the main sound objects. Then, the user can select the target sound objects before entering an interactive loop that uses active learning and relevance feedback provided by the user. The feedback of the user is then used by the system to perform a multilabel classification of sound segments based on the selected ones sound objects. An evaluation of the system is performed by user simulation using a synthetic corpus. The evaluation shows that our approach achieves satisfying results in a reasonable number of interactions.

Keywords : electroacoustic music, interactive machine learning, relevance feedback, active learning, multilabel classification.

Table des matières

Chapitre 1

Introduction

Sommaire

1.1 Contexte

Cette thèse en convention CIFRE s'est déroulée dans deux départements distincts de l'Institut National d'Audiovisuel (INA). Les membres de l'équipe "Visualisation, Indexation et Fouille de données" (VIF) ont assuré l'encadrement scientifique et le Groupe de Recherches Musicales (GRM) a mené l'encadrement applicatif. La cotutelle académique a été effectuée par le département "Traitement du Signal et de l'Image" (TSI) de l'école TELECOM ParisTech.

Ce travail de thèse s'inscrit dans des problématiques propres au *Music Information Retrieval* (MIR), un domaine de recherche qui vise à retrouver des informations à partir d'un contenu musical (Casey et al. (2008)). Le MIR est un domaine ou les applications concrètes sont très nombreuses. Un exemple d'application qui vient naturellement à l'esprit est la *transcription automatique* d'un enregistrement musical en partition. On peut également imaginer un *système de recommandation musicale* qui vous suggère de nouveaux morceaux en apprenant vos gouts musicaux à partir de ce que vous écoutez. Vous n'êtes pas de bonne humeur et vous comptez sur la musique pour vous apporter la motivation qui vous manque pour attaquer votre journée sereinement ? Il vous suffit de demander à votre application de *classification en humeurs* de rechercher les morceaux adéquats dans la bibliothèque sonore de votre smartphone. Ou encore, pourquoi ne pas personnaliser ou diversifier votre expérience d'écoute ? C'est une application possible de la *séparation de sources* qui vous permet de supprimer les sources instrumentales que vous ne souhaitez pas écouter directement sur vos fichiers musicaux. Enfin, à l'opposé, pouvons nous envisager la *génération automatique* d'une symphonie originale telle qu'elle aurait pu être composée par *Beethoven* en apprenant automatiquement le style du compositeur par des analyses statistiques de ses pièces ?

La plupart des applications et problématiques citées précédemment font encore partie

du domaine de la recherche. Si la "musique conventionnelle"[1] et notamment classique est largement traitée dans la littérature MIR, on ne peut pas en dire autant des musiques plus atypiques. Ainsi, ce travail de thèse est centré autour des musiques électroacoustiques. Une propriété importante de ces musiques est qu'elles ne sont en général pas écrites et qu'il n'existe pas de standard de notation symbolique comme c'est le cas dans la musique conventionnelle. De plus, les musiques électroacoustiques n'utilisent pas les instruments standards et peuvent faire intervenir n'importe quelle source sonore acoustique ou électronique. Il est également important de noter que les musiques électroacoustiques ne disposent pas des mêmes unités sonores de base que la musique conventionnelle qui est centrée sur la notion de note que l'on peut assimiler à une variable discrète ayant une hauteur et une durée relative. Des problèmes de recherche originaux découlent en grande partie de ces dernières remarques. Les travaux de recherches sur les musiques électroacoustiques concernent principalement la création de nouveaux outils de composition et la pédagogie (Desainte-Catherine & Marchand (1999), Sedes et al. (2004), Savage & Challis (2002), Kurtag et al. (2007)). L'analyse musicale des pièces du répertoire est également un sujet d'étude important (Geslin & Lefevre (2004), Couprie (2004), Gayou (2006)).

Dans ce travail de thèse nous nous intéressons en particulier aux problématiques scientifiques liées à l'analyse automatique des musiques électroacoustiques. En effet, dans la plupart des cas, l'analyse des musiques conventionnelles s'appuie en grande partie sur la partition et sur des méthodes établies. C'est, par exemple, le cas de l'analyse des musiques tonales qui passe systématiquement par une étude harmonique. Cependant, pour les musiques électroacoustiques, ne disposant pas de partition, il est nécessaire d'avoir recours à d'autres approches. Nous verrons dans la suite du document que les musicologues passent par une étape de transcription des objets sonores principaux d'une pièce. Ils utilisent ensuite cette transcription afin d'appuyer un *point de vue d'analyse*. Cette étape de transcription étant la base de chaque analyse, dans ce travail, nous proposons une approche MIR pour assister les musicologues dans la transcription des différents objets sonores.

1.2 Objectifs et problématiques

Dans un premier temps, il est important de cerner les besoins réels des musicologues dans le domaine de l'analyse des musiques électroacoustiques. En effet, il semble impératif de passer par cette étape car comme nous le verrons, les approches d'analyse ne sont pas standardisées. Les problématiques scientifiques sont en accord avec cette étape initiale. Nous essayerons de proposer des méthodes qui pourront s'adapter à la diversité des signaux possibles ainsi qu'aux besoins spécifiques de chaque analyse utilisateur. En effet, pour l'analyse des musiques électroacoustiques il est important de ne pas négliger la dimension subjective de la tâche car chaque analyse est fondée sur un point de vue et les objets sonores que l'utilisateur souhaite transcrire sont en accord avec ce point de vue. Ainsi, notre objectif est de proposer des méthodes qui permettent d'assister l'utilisateur dans la sélection des objets sonores qu'il souhaite transcrire et de retrouver les différentes instances de ces objets pour réaliser une transcription adaptée. Pour atteindre cet objectif, nous devons transposer des méthodes devenues classiques en MIR aux musiques électroacoustique qui présentent des configurations sonores se démarquant très nettement de celles des musiques conventionnelles. De plus, il est également nécessaire d'introduire la notion

[1] Dans ce document, nous désignerons ainsi toutes les musiques utilisant des systèmes d'échelles musicales à hauteurs déterminées ainsi qu'un système rythmique qui permet d'exprimer des durées les unes par rapport aux autres.

de subjectivité dans les méthodes développées afin de s'adapter aux nombreux points de vues possibles.

1.3 Contributions

La première contribution de ce travail de thèse est la proposition d'une architecture originale qui utilise le *retour de pertinence* afin de réaliser un système adaptatif (Gulluni et al. (2011*b*,*a*)). La notion de *retour de pertinence* désigne une méthode qui prend en compte le jugement qu'un utilisateur fournit lors de la recherche automatique de documents. A l'origine, le *retour de pertinence* est employé dans les travaux de Rocchio qui l'utilisait pour modifier des requêtes en fonction du jugement apporté par l'utilisateur sur les documents retrouvés par son système (Rocchio & Salton (1971)). Des recherches récentes emploient souvent cette méthode pour retrouver des documents multimédias (photos etc.). L'architecture que nous proposons procède en deux phases principales. La première phase réalise une segmentation de la pièce qui permet d'assister l'utilisateur dans la sélection d'objets sonores. La deuxième phase effectue une classification des objets sonores afin de retrouver les différentes instances des objets sélectionnés dans la première étape. L'architecture proposée est décrite en détail dans la section 2.5.2.2.

Nous proposons une approche de classification *multilabel* des objets sonores (un segment audio peut appartenir à plusieurs classes) et exploitant le *retour de pertinence* adaptée à notre problème (chapitre 4). Ainsi, dans la section 4.6, nous comparons deux approches d'interactions pour la classification multilabel de segments audio sur plusieurs niveaux de polyphonie : une approche par passages multiples et une autre par passage unique. Nous proposons également, dans l'approche par passage unique, une méthode de classification qui s'adapte aux différentes mixtures sonores exprimées par l'utilisateur (Gulluni et al. (2011*b*)). Dans la section 4.7.2.1, nous montrons que cette dernière méthode permet d'obtenir un gain de performances consistant sur plusieurs niveaux de polyphonie par rapport à l'approche directe, tout en conservant des temps de calcul acceptables.

Une autre contribution de ce travail est la proposition d'une méthode de segmentation audio interactive dans laquelle le *retour de pertinence* est propagé sur l'ensemble du signal (chapitre 3). Dans le paragraphe 3.4.2.2, nous démontrons que des interactions simples permettent d'améliorer la segmentation d'un signal audio par rapport à une approche de référence (Gulluni et al. (2009)).

Afin de cerner les besoins réels des professionnels de l'électroacoustique, une étude des pratiques d'analyse a été réalisée auprès de trois musicologues (section 2.4.2). Les renseignements apportés par cette étude sont exploités dans ce travail. De plus, cette étude peut également être utile à la communauté car elle met en évidence des problématiques qui pourraient donner lieu à de nouvelles directions de recherche.

1.4 Présentation du manuscrit

Le manuscrit est divisé en cinq parties incluant la présente introduction. Le chapitre 2 présente le contexte et les notions musicales intervenant dans la thèse et expose l'architecture globale du système. Ainsi, ce chapitre aborde l'émergence des nouvelles pratiques

musicales. Les musiques électroacoustiques sont une conséquence directe de ces pratiques. Nous aborderons ensuite des définitions musicales essentielles propres aux musiques électroacoustiques. Ce chapitre présente également une série d'entretiens avec des musicologues qui permettent de cerner leurs besoins réels. La fin du chapitre présente une vision globale du système d'aide à l'analyse des musiques électroacoustiques que nous proposons et le corpus d'évaluation du système.

Le chapitre 3 porte sur la phase d'initialisation du système qui repose sur la segmentation de la pièce en unités sonores homogènes afin d'obtenir les frontières temporelles qui séparent les différentes mixtures sonores d'une pièce électroacoustique polyphonique. Ce chapitre aborde dans un premier temps l'état de l'art des différents systèmes de segmentation audio puis il propose une solution interactive et compare deux scénarios différents d'interaction avant d'évaluer le système de segmentation.

Le chapitre 4 est focalisé sur la phase de classification des objets sonores. La solution proposée est une classification interactive exploitant le retour fourni par l'utilisateur. Après avoir présenté un état de l'art portant sur les différents domaines connexes au sujet, nous verrons comment exploiter les informations obtenues pendant l'initialisation et nous proposerons différentes approches d'interaction pour réaliser la classification. La dernière partie du chapitre décrit l'évaluation du système complet basée sur des simulations utilisateurs.

Enfin, dans le chapitre 5 nous exposons un bilan des travaux effectués pendant cette thèse et abordons les perspectives et travaux futurs.

Chapitre 2

Musiques électroacoustiques : définitions, analyse et architecture d'un système adapté

Sommaire

2.1 Introduction

Dans ce chapitre, nous présentons le contexte musical lié à ce travail de thèse. L'objectif n'est pas de présenter les musiques électroacoustiques de manière exhaustive mais plutôt d'exposer leurs origines et leurs caractéristiques puis d'expliquer comment les musicologues abordent leur analyse. Les enseignements tirés de ce travail préparatoire nous permettrons de présenter les objectifs ainsi que l'architecture du système proposé. Après avoir abordé la naissance des musiques électroacoustiques, nous donnerons quelques définitions essentielles avant d'aborder leur analyse. La section 2.4 de ce chapitre intègre la présentation d'une synthèse d'entretiens réalisés avec trois musicologues spécialisés dans l'analyse de ces musiques. Le chapitre se termine par la présentation générale du système proposé et du corpus d'évaluation.

2.2 Naissance des musiques électroacoustiques

Cette section est une synthèse tirée de Manning (2004) qui replace les musiques électroacoustiques dans le contexte historique. Nous avons résumé dans cette section les idées principales qui concernent la naissance des musiques électroacoustiques et les expérimentations ayant donné lieu aux premières pièces du genre. Il nous semble important de mentionner que la principale vocation de cette section est didactique et qu'elle ne constitue pas une étude musicologique personnelle. Cependant, la présence de cette section est indispensable, car elle décrit comment les nouvelles pratiques musicales ont émergé et permet ainsi de comprendre les origines et les procédés courants employés dans les musiques électroacoustiques.

2.2.1 Développements avant 1945

2.2.1.1 Apparitions des premiers instruments de musique non acoustiques

Les premiers instruments de musique utilisant des procédés de génération sonore non acoustiques sont apparus au début du 20^e siècle. Le premier de ces instruments est le *Dynamophone* (ou *Telharmonium*) conçu par *Thaddeus Cahill* à partir de 1897 et présenté en public pour la première fois en 1906 (figure 2.1).

Fig. 2.1 – Le premier Telharmonium

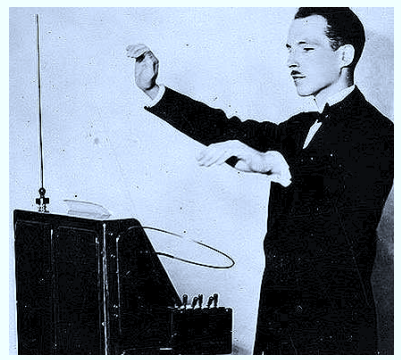

Fig. 2.2 – Léon Thérémin et son invention

Cet instrument est le premier à avoir utilisé un procédé électromécanique pour la génération sonore. Ainsi, cet instrument utilisait principalement une roue phonique placée

devant un microphone pour produire le signal sonore.

Il faudra ensuite attendre l'entre-deux-guerre pour voir apparaitre les premiers instruments électroniques. Les principaux sont le *Thérémin* présenté en 1924 (figure 2.2), le *Spharophon* (1927), le *Dynaphone* (1927), les *Ondes Martenot* et le *Trautonium* (figures 2.3 et 2.4) présentés en 1930. La plupart de ces instruments utilisaient un clavier et ne pouvaient jouer qu'une seule note à la fois (instruments monophoniques).

FIG. 2.3 – Les Ondes Martenot avec leurs diffuseurs

FIG. 2.4 – Le Trautonium

Malgré les contributions de compositeurs établis tel que *Messiaen*, *Koechlin*, *Honegger*, *Hindemith* ou encore *Milhaud*, le répertoire compte un nombre limité de compositions dédiées à ces instruments. Les compositeurs ayant montré le plus d'intérêt pour ces instruments sont ceux qui écrivaient des musiques de films. Cependant, les *Ondes Martenot* ont tout de même réussi à se faire une place relative notamment dans les pièces de *Messiaen* (*Turangalîla-Symphonie*, *Trois Petites Liturgies*). Les *Ondes Martenot* sont aujourd'hui encore enseignées au conservatoire de Paris.

2.2.1.2 Vers de nouvelles formes d'expression

Les nouveaux procédés de génération sonore ont attiré l'attention du *mouvement futuriste* qui cherchait à imiter les sons industriels. Ce mouvement fut initié par le poète italien *Filippo Marinetti* en février 1909 lors de la publication du *Manifesto of Futurist Poetry*[1]. Les intentions musicales de ce mouvement furent par la suite exprimées par *Balilla Pratella* dans *Manifesto of Futurist Musicians*[2] en octobre 1910. Ce document propose un rejet des principes et méthodes traditionnelles musicales d'enseignement pour leur substituer une expression libre inspirée par la nature dans toutes ses manifestations. D'autres ouvrages du mouvement furent publiés dans les mois suivants. Dans *The art of noise*, *Luigi Russolo* proposait d'utiliser des sources sonores environnementales dans la composition musicale : *"Les sons musicaux sont trop limités à des variétés de timbres qualitatives. Les orchestres les plus complexes se limitent à quatre ou cinq catégories d'instruments de timbres différents : les instruments joués à l'archet, les instruments à cordes pincées, la famille des cuivres, la famille des bois et les instruments à percussions ... Nous devons sortir de ce cercle res-*

[1] ce texte fut publié dans *Le Figaro* le 20 février 1909

[2] `http://www.unknown.nu/futurism/musicians.html`

treint des sons musicaux purs et conquérir l'infinie variété des bruits" (Russolo (1913)). Ces propositions furent matérialisées par la construction d'instruments bruitistes : les *Intonarumori*, en collaboration avec le percussionniste *Ugo Piatti*. Ainsi, le premier concert basé sur ces instruments, l'*Art des bruits*, eut lieu à Milan en juin 1913 au théâtre Storchi.

Finalement, le *mouvement futuriste* ne provoqua pas une révolution majeure mais sa remise en cause des relations bien établies entre les sciences de l'acoustique et l'art musical furent prophétiques. D'ailleurs, le futuriste *Busoni* avec son *Sketch of a New Esthetic of Music* (Busoni (1911)) attira l'attention du jeune *Edgard Varèse* qui se rebella contre le conservatoire de Paris afin de pouvoir explorer des nouveaux concepts d'expression musicale. On peut citer *Varèse* comme étant le compositeur de son époque ayant le plus contribué à l'acceptation de sources sonores diverses dans la composition musicale à travers son oeuvre. Malheureusement, il n'eut accès aux moyens techniques qu'il espérait que dans les années 50, vers la fin de sa vie. Les écrits de *Varèse* sur le potentiel des instruments électroniques dans la composition musicale furent approuvés par *John Cage*, un compositeur américain d'une esthétique musicale pourtant bien différente. En 1937, lors d'un congrès à la *Seattle Arts Society*, *John Cage* déclarait : *"Alors que par le passé, les points de divergence se situaient entre la dissonance et la consonance, dans un futur proche cela sera entre le bruit et les sons dit musicaux. Où que nous soyons, tout ce que nous entendons est principalement du bruit... Nous voulons capturer et contrôler ces sons, nous ne souhaitons pas les utiliser comme des traitements de studio mais comme des instruments... De nombreux concepteurs d'instruments musicaux électriques essayent d'imiter les instruments du 18^e et 19^e siècle... Alors que Thérémin proposait un instrument avec des possibilités nouvelles véritables, les Théréministes ont fait de leur mieux pour faire sonner l'instrument comme un vieil instrument en lui donnant avec difficulté un doux vibrato pour interpréter les pièces majeures du passé. Les caractéristiques spécifiques des instruments électriques seront de donner un contrôle total sur la structure harmonique des sons (à l'opposé des bruits) et de rendre ces sons utilisables à n'importe quelle fréquence, amplitude et durée"*. La renaissance des arts d'après guerre fut un terrain plus favorable au développement de la musique électronique. En Europe, deux grands pôles prirent l'initiative de s'investir dans ce domaine : la *Radiodiffusion Télévision française* (RTF) à Paris avec la *musique concrète* et la *Norwestdeutscher Rundfunk* (NWDR) à Cologne avec l'*elektronische musik*. Malgré une grande curiosité réciproque, les deux écoles ont connu quelques divergences à leurs débuts au sujet des pratiques de composition.

2.2.2 Paris et la *musique concrète*

2.2.2.1 Naissance d'un groupe de recherche

Le courant français a pour principal initiateur *Pierre Schaeffer*, un ingénieur polytechnicien ayant commencé des recherches dans le domaine des sciences de l'acoustique musicale en France à partir de 1942 en créant le *Studio d'Essai*. L'équipement de l'époque était rudimentaire puisqu'il s'agissait principalement d'un enregistreur sur disque. En 1951, la RTF accepta de financer un nouveau studio pour les recherches de *Schaeffer*. La nouveauté la plus importante était l'utilisation de l'enregistreur à bande comme outil principal à la place de l'enregistreur sur disque. Un des enregistreurs à bande disponible permettait de reproduire cinq pistes sonores à la fois ce qui ouvrit la porte à la distribution des canaux audio sur un ensemble de plusieurs haut-parleurs. Trois magnétophones particuliers furent également introduits dans le nouveau studio : le *Morphophone* (réverbération basée sur

des échos du son original), deux types de *Phonogènes* qui étaient conçus pour jouer des bandes en boucle à différentes vitesses (le premier type permettait un contrôle continu de la vitesse, le deuxième était associé à un clavier et effectuait des transpositions de hauteurs fixes en variant la vitesse). Le nouveau studio connu une expansion importante des activités et des collaborateurs de *Schaeffer*. Ainsi, le groupe fut renommé "Groupe Recherche de Musique Concrète" pour devenir le "Groupe de Recherches Musicales" (GRM) en 1958.

2.2.2.2 Les débuts de la *musique concrète*

Le premier travail de *Schaeffer*, *Etude aux chemins de fer* (la première des *Cinq études de bruits*), pose une constante de ce qui deviendra la *musique concrète* : la composition à partir d'enregistrements issus de sources sonores diverses. Cette première pièce est composée à partir d'enregistrements effectués à la Gare des Batignolles à Paris. Les sources sonores enregistrées incluaient le sifflement de locomotives à vapeur, leurs accélérations et les wagons passant d'un rail à un autre. La pièce est basée principalement sur des juxtapositions de parties (à l'opposé de la superposition de plusieurs parties), ce qui amplifie le caractère répétitif des sons. Pendant l'été 1949, *Schaeffer* a commencé à se réapproprier les instruments de musique en tant que sources sonores ce qui lui permet de renouer avec les travaux de *Varèse* initiés 20 ans plus tôt. La pièce suivante de *Schaeffer* est *Suite pour quatorze instruments* et a pour caractéristique d'être le point de départ de son travail sur la syntaxe de la *musique concrète*. Cette pièce en cinq mouvements met en valeur divers procédés caractéristiques de la *musique concrète* : *Courante* est une monodie assemblée par juxtaposition de petits extraits de l'ensemble de la librairie d'enregistrements sonores, *Gavotte* utilise l'interprétation par divers instruments d'une petite phrase musicale en juxtaposition pour créer un ensemble de variations. On peut noter un emploi intensif de la transposition de hauteur en jouant les enregistrements à des vitesses différentes. *Schaeffer* ne tarda pas à donner une première définition au concept d'*objet sonore* : *évènement sonore élémentaire qui est isolé de son contexte original et examiné pour ses caractéristiques natives en dehors du continuum temporel normal.*

Symphonie pour un homme seul est la première pièce de *Schaeffer* en collaboration avec le compositeur *Pierre Henry*. Les préoccupations de *Schaeffer* étaient alors l'extension des possibilités des sources sonores instrumentales par l'intermédiaire des nouveaux moyens techniques et également le développement du principe d'*objet sonore* et leurs règles de composition. Dans cette pièce, *Schaeffer* et *Henry* distinguent deux types de sources sonores : celles produites par l'homme (respirations, fragments de voix, cris, fredonnements, sifflements) et celles résultantes de la communication de l'homme avec son environnement (bruits de pas, claquements de portes, percussions, piano préparé[3], instruments orchestraux). Un exemple de divergence entre les courants français et allemands est la première diffusion de *Symphonie pour un homme seul* aux radios de Cologne (NWDR), Hambourg, Baden-Baden et Munich en 1951. Les sympathisants de l'*elektronische musik*, un courant musical allemand qui se développait pendant la même période que la *musique concrète*, accueillirent la pièce avec une certaine hostilité. Malgré cela, cette pièce sera par la suite acceptée et considérée comme un classique.

[3]piano dont le son est modifié par le placement d'objets extérieurs dans ses cordes

2.2.2.3 Formalisation et notation

En 1952, *Schaeffer* publie une syntaxe de la *musique concrète* dans la dernière section de son livre *A la recherche d'une musique concrète* (Schaeffer (1952)). Dans ce chapitre, *Esquisse d'un solfège concret*, il donne entre autres 25 définitions pour l'exploitation des *objets sonores* ainsi que les procédés de base qui leur sont applicables. On peut distinguer notamment des méthodes de classification, des opérations de traitements en amont du travail de composition (altérations de divers paramètres des sons), des procédés de réalisation d'une pièce de *musique concrète* (montage, mixage, spatialisation etc.). L'aboutissement du travail de formalisation de la *musique concrète* par *Shaeffer* est le *Traité des Objets Musicaux* qui parut en 1966 et fixa ainsi les notions esquissées de ses précédents écrits (Schaeffer (1966)).

Henry, pendant la composition de *Concerto des ambiguïtés* et *Suite* en 1950, rencontre des difficultés importantes pour la notation de son travail. *Concerto des ambiguïtés* étant principalement basé sur un piano préparé, l'utilisation d'une notation classique était inadaptée du fait que les résultats acoustiques étaient significativement différents des évènements notés sur la partition. Les premières tentatives de notations incluant des graphiques additionnels à la partition classique, notamment pour représenter la hauteur des sons, furent essayées sans succès. Le principal inconvénient de cette proposition était de ne pas donner d'information sur le timbre. En 1951, *Schaeffer* et *Henry* travaillaient sur le premier opéra concret *Orphée* et *Schaeffer* en cette occasion ressentit la nécessité de créer deux types de partitions : la *partition opératoire* décrivant les procédures techniques et la *partition d'effet* pour le développement des idées musicales sur des portées parallèles associées à chacun des éléments concrets.

2.2.3 Cologne et l'*elektronische musik*

2.2.3.1 Création du studio de Cologne

En Allemagne, les innovations esthétiques et technologiques dans le domaine musical donneront lieu à l'*elektronische musik*. A l'opposé de la *musique concrète* dont *Schaeffer* est le principal initiateur, il s'agit ici du fruit de la collaboration de plusieurs personnes ayant des compétences techniques et musicales.

En 1948, le docteur *Werner Meyer-Eppler* qui est alors directeur du département d'études phonétiques à l'université de *Bonn* reçoit la visite de *Homer Dudley*, un chercheur américain du *Bell Telephone Laboratories*. *Dudley* profite de cette visite pour présenter la machine qu'il venait de concevoir : le *Vocodeur* (Voice Operated reCOrDER). Impressionné par cette invention, *Meyer-Eppler* utilise le Vocoder comme illustration lors d'une conférence sur la production du son par des moyens électroniques qui eu lieu en 1949 à Detmold. Par chance, *Robert Beyer* de la NWDR faisait partie de l'audience. L'intérêt des deux scientifiques pour l'utilisation des technologies électroniques dans un contexte musical aboutit à la réalisation d'une conférence commune sur le *"Le monde sonore de la musique électronique"* en 1950 lors de l'International Summer School for New Music de Darmstadt. Le compositeur *Herbert Eimert* qui était présent exprima un intérêt particulier pour leurs idées et les trois hommes discutèrent d'une association informelle afin de promouvoir l'*electronische music*.

Le 18 octobre 1951, la station radio de Cologne propose un programme intitulé *"Le monde sonore de la musique électronique"* sous la forme d'un forum tenu par *Eimert*, *Beyer* et *Meyer-Eppler*. Le forum était illustré par des démonstrations sonores réalisées à partir d'un instrument électronique, le *Melochord*, qui n'était pas sans rappeler le *Trautonium*. Le jour même, un comité spécial qui incluait entre autres *Fritz Enkel*, le directeur technique de la radio de Cologne, et bon nombre de ses assistants fut formé. Ce comité décida d'établir un studio de musique électronique afin de *"poursuivre les procédés suggérés par* Meyer-Eppler *et de composer directement sur la bande magnétique"*. Le projet prit deux ans avant de devenir complètement opérationnel et *Eimert* fut nommé directeur artistique du studio.

2.2.3.2 Les premières pièces d'*elektronische musik*

Beyer et *Eimert* composèrent leurs premières pièces électroniques entre 1951 et 1953 alors que le studio de Cologne était encore en construction (*Klang im unbegrenzten Raum*, *Klangstudie I*, *Klangstudie II*). Dans la première moitié de l'année 1953, ils composèrent *Ostinate Figuren und Rhythmen* et *Eirmert* composa *Struktur 8* seul. Ces premières pièces du courant de l'*elektronische musik* sont caractérisées par l'application stricte de procédures sérielles[4] que ce soit au niveau de la sélection des timbres ou des traitements. En effet, de nombreux compositeurs d'*elektronische musik* vouent une grande estime à la *seconde école de Vienne*[5] et sont donc de fervents défenseurs de la cause sérielle.

Contrairement à la *musique concrète* qui utilise des sources sonores enregistrées principalement acoustiques comme matériel de base, l'*elektronische musik* utilise plutôt des procédés électroniques pour la génération sonore. Le désir de contrôle total sur le timbre induit le générateur d'ondes sinusoïdales comme étant la source sonore la plus appropriée. En effet, selon *le théorème de Fourier*, on peut décomposer une source sonore périodique en la somme de plusieurs composantes sinusoïdales de fréquences, amplitudes et phases déterminées. Initialement, le studio de Cologne était constitué exclusivement d'un générateur sinusoïdal de haute précision, un générateur de bruit blanc, un *Monochord* électronique et un *Melochord*. Ces deux derniers instruments étaient équipés de claviers et le *Melochord* pouvait générer des ondes caractéristiques que l'on retrouvera plus tard dans les premiers synthétiseurs : onde en dents de scie, onde triangulaire et onde carrée.

Les premières pièces de *Karlheinz Stockhausen* composées au studio de Cologne, *Studie I* en 1953 et *Studie II* en 1954, furent créées uniquement à partir du générateur de sinusoïdes. Ces deux pièces illustrent bien la notion de "mixture de notes" qui désigne la combinaison de sinusoïdes dont les fréquences ne sont pas en rapport harmonique. Cette notion permet de distinguer les spectres harmoniques des spectres inharmoniques. On peut également citer *Gesang der Jûnglinge* (1955-1956) de *Stockhausen* comme un tournant dans le développement artistique du studio de Cologne pour tous les enseignements qu'elle apporte et son intégration de la voix humaine avec des sons électroniques. Parmi les traitements utilisés dans les compositions d'*elektronische musik*, on peut citer entre autres le filtrage et la modulation en anneau. On constate que l'interdépendance entre la synthèse et la composition musicale constitue une constante du courant allemand, la frontière entre

[4] le sérialisme est un courant musical du XX^{e} siècle qui évite toute tonalité en donnant une importance égale à chacune des 12 notes de la gamme chromatique. La musique sérielle est composée autour de la notion de série : succession de sons fixée au préalable et invariable

[5] désigne les compositeurs *Schönberg*, *Berg* et *Webern*, en référence à la "première" école de Vienne, celle d'*Haydn*, *Mozart*, *Beethoven* et *Schubert*

les deux disciplines était d'ailleurs parfois très floue.

Après les quelques divergences entre la *musique concrète* et l'*elektronische musik*, on admet aujourd'hui que les deux écoles constituent deux facettes complémentaires des débuts de la musique électroacoustique.

2.2.4 Milan, un autre studio européen important

Le studio de Milan fut créé en 1955 par la Radio Audizioni Italiane (RAI) et co-fondé par les compositeurs *Luciano Berio* et *Bruno Maderna*. Ce centre, qui a fortement influencé le studio de *Cologne*, a été créé pour les besoins de l'école italienne de composition. La majorité des compositeurs de ce studio ne rentraient pas dans les querelles franco-allemandes sur la production des sons, préférant se consacrer aux caractéristiques perçues des structures sonores.

Une constante des pièces produites dans le studio de Milan pendant les années 50 et au début des années 60 était la préoccupation qu'avaient les compositeurs pour la texture et la sonorité. Un processus de composition courant était la formation de clusters de sons à partir de sinusoïdes et la création de flux sonores à partir de bruits blancs filtrés.

L'école de Milan a donné une réponse pertinente aux problèmes rencontrés par l'*elektronische musik* et la *musique concrète*. Dans *Différences*, *Berio* montre comment des sons naturels peuvent être développés par l'utilisation de traitements sonores. Cette pièce est un quintet pour flûte, clarinette, harpe, alto et violoncelle auxquels s'ajoute une partie sur bande magnétique qui reprend des enregistrements des instruments en les modifiant par des procédés électroniques. La partie électronique sur bande est utilisée comme un moyen de développement des sonorités après une exposition réalisée par les instruments seuls. On peut remarquer que la parole devient une source sonore très utilisée par les compositeurs de Milan. Par exemple, dans la pièce *Thema*, *Berio* utilise principalement de courts extraits du texte *Ulysses* de *James Joyce* qu'il manipule par des procédés électroniques. Le texte est d'abord exposé en intégralité dans une première lecture puis la pièce se développe en désagrégeant le texte original par fragmentations, superpositions et variations du timbre par filtrage.

L'équipement matériel du studio de Milan était composé de neuf générateurs d'ondes sinusoïdales, un générateur de bruit blanc, un générateur d'impulsions, une version modifiée d'*Ondes Martenot* et un ensemble de magnétophones mono, stereo et quatre pistes. La présence des neufs générateurs d'ondes sinusoïdales était un avantage certain pour les compositeurs par rapport au studio de Cologne car cela permettait d'ajuster certaines combinaisons et paramètres en temps réel.

Le studio de Milan, tout comme ceux de Paris et Cologne, a continué de jouer un rôle important dans le développement artistique des années 60. Plusieurs studios se développèrent dans le monde. Ainsi, la Russie, le Japon, le Royaume-Uni, la Suède, la Belgique et les Etats-Unis ont également été des acteurs importants dans le développement des musiques électroacoustiques. On pourra se référer à Manning (2004) qui présente le développement des musiques électroacoustiques de manière exhaustive.

2.3 Définitions

Cette section regroupe des définitions et notions musicales essentielles à la compréhension de la suite du document.

Musiques électroacoustiques

La naissance des pratiques électroacoustiques a engendré plusieurs esthétiques musicales très différentes. Aujourd'hui, il est difficile de donner une définition précise de la *musique électroacoustique*. Selon le Larousse, ce terme a été créé dans les années 50 pour désigner toute musique construite à partir de sons enregistrés (*musique concrète*) ou de synthèses (*elektronische musik*) en références aux deux courants initiés en France et en Allemagne. Aujourd'hui, le Wikipédia recense plusieurs définitions de la *musique électroacoustique* :

1. Le terme "musique électroacoustique" désigne tout type de musique dans laquelle l'électricité a un rôle autre que la simple utilisation du microphone ou de l'amplification pour la production de cette musique ;
2. Désigne tout ce qui utilise la conversion d'un signal acoustique en signal électrique et vice et versa ;
3. Musique utilisant la technologie pour enregistrer, produire, créer, manipuler et diffuser le son ;
4. Désigne toutes les activités utilisant l'électricité pour produire, manipuler, diffuser et étudier le son (correspond au terme "electroacoustics" des pays anglo-saxons).

Ces définitions sont difficiles à utiliser dans le contexte musical actuel ou la quasi-totalité de la production musicale utilise des moyens électroniques à un moment de la chaine de création. Ainsi, si on applique ces définitions, une musique utilisant le langage tonal, entièrement produite à partir d'instruments acoustiques, mais enregistrée par des moyens électroniques devient électroacoustique. Les définitions citées ne prennent pas en compte le paradoxe que nous venons d'exposer. Ainsi, dans ce document, nous ferons principalement référence à une définition stylistique des *musiques électroacoustiques* : *regroupement de courants musicaux aux esthétiques distinctes nés dans les années 40 en réaction aux innovations technologiques de production sonore*. Par conséquent, on considère la musique électroacoustique comme une collection de genres musicaux et non comme une musique utilisant des moyens électroniques pour sa production.

Musique acousmatique

Le terme de *"musique acousmatique"* revient fréquemment dans les écrits consacrés aux musiques électroacoustiques. A l'origine, l'adjectif *acousmatique* est repris par l'écrivain et poète Jérôme Peignot en 1955 pour exprimer la *"distance qui sépare les sons de leur origine"*. Cette expression est par la suite reprise par *Schaeffer* en 1966 dans le *Traité des objets musicaux* (Schaeffer (1966)). En 1974, le compositeur *François Bayle* reprend l'expression afin d'éviter la confusion avec les musiques qui utilisent des instruments ayant recours à l'électricité. La *musique acousmatique* désigne selon *Bayle* une musique qui *"se tourne, se développe en studio, se projette en salle, comme le cinéma"*. Dans l'usage courant, les deux expressions *musique concrète* et *musique acousmatique* sont souvent utilisées pour désigner une même musique, celle créée par *Schaeffer* dans les années 40.

Musique polyphonique et monophonique

Selon le dictionnaire de l'académie française, la polyphonie est un *"chant à plusieurs voix qui se superposent selon les règles du contrepoint (par opposition à monodie) ; par extension, combinaison simultanée de deux ou plusieurs lignes musicales mélodiques qui, tout en formant un ensemble homogène, conservent chacune sa beauté singulière."*. Il faut entendre cette définition dans le contexte de la musique conventionnelle. Ainsi, en adaptant cette définition aux musiques qui nous concernent, les musiques électroacoustiques polyphoniques sont celles qui superposent plusieurs sons. Dans l'usage courant, par abus de langage, on oppose le terme *polyphonique* à *monophonique* (et non *monodique*). Dans ce document, nous opposerons donc les musiques électroacoustiques polyphoniques aux monophoniques qui ne font entendre qu'un seul son à la fois. Un bon exemple de musique électroacoustique monophonique est la pièce *Timbre Durée* de *Messiaen*.

Objet sonore

La notion d'*objet sonore* a été formalisée par *Schaeffer* lors de la naissance de la *musique concrète*. La définition préliminaire citée dans la section précédente laisse place aujourd'hui à la définition suivante qui est admise par la majorité de la communauté : *phénomène sonore perçu dans le temps comme un tout, une unité, quels que soient ses causes, son sens, et le domaine auquel il appartient (musical ou non)*. On peut également se référer à Kane (2007) qui propose d'étudier l'emploi de la notion d'*objet sonore* dans un contexte à la fois contemporain et historique. La notion d'*objet sonore* est suffisamment universelle pour s'appliquer à des esthétiques autres que celle de la *musique concrète* qui a engendré sa définition. L'*objet sonore* est également un outil d'analyse puissant qui permet d'isoler les atomes constituants des musiques électroacoustiques. Le système que nous proposons ne prétend pas convenir à toutes les esthétiques de la grande famille des *musiques électroacoustiques*. **Ainsi, le système présenté sera principalement dédié aux musiques pouvant être décomposées en *objets sonores***.

Analyse poïétique et esthésique

On oppose souvent deux approches d'analyse : l'analyse poïétique et l'analyse esthésique. Molino distingue également le niveau neutre dans son système de tripartition (Molino (2009)). Selon Molino : *"Le poïétique rassemble les processus qui ont mené à la création d'une forme symbolique, et qu'on nomme aussi les stratégies de production"* et *"L'esthésique rassemble les processus de réception, au cours desquels il y a attribution de significations"*. Le niveau neutre correspond à l'oeuvre dans son existence matérielle (la partition dans le cas de la musique écrite).

Unités Sémiotiques Temporelles

Les *Unités Sémiotiques Temporelles* (UST) sont des figures sonores dont la signification musicale s'exprime temporellement (Formosa et al. (1996)). Elles sont issues d'une série de remarques nées de la pratique de la musique électroacoustique :

- Le matériau sonore de ces musiques échappe à une description selon des modèles d'organisation en termes de hauteurs et de durées relatives.
- La pratique des musiciens les mène à appréhender les phénomènes sonores plutôt par des considérations de sens que par des considérations typo-morphologiques, comme le

propose *Schaeffer*. C'est-à-dire à travers ce qu'ils évoquent "en arrière-plan", soit au niveau des images suscitées par les sons, soit au niveau de "l'aventure" de la matière sonore elle-même.
– L'organisation temporelle, la dynamique de l'oeuvre, serait une des composantes importantes d'une musique faite de sons pour lesquels la notion de hauteur vue comme degré est un cas particulier, et dont le travail de composition s'appuie essentiellement sur une réalité sonore, celle des sons enregistrés.

2.4 Analyse des musiques électroacoustiques

Après avoir défini le genre musical concerné par notre système, nous proposons dans cette section de nous intéresser à l'analyse des musiques électroacoustiques telle qu'elle est pratiquée par les musicologues. Dans la première partie, nous décrivons l'état de l'art des approches théoriques puis dans la seconde partie nous apportons des éléments de réponses pratiques à travers trois entretiens réalisés avec des musicologues spécialistes du domaine.

2.4.1 Etat de l'art

Le problème de la méthode est récurent dans l'analyse des musiques électroacoustiques. Simoni et al. (2000) mettent en évidence les problèmes spécifiquement liés à l'analyse des musiques électroacoustiques et proposent une théorie caractérisée par l'interaction d'un modèle perceptif et d'un modèle analytique. Le modèle analytique examine des aspects de la composition aussi bien au niveau macroscopique (la forme) que microscopique (le spectre instantané). Le modèle perceptif, en plus de l'écoute, utilise un spectrogramme sur lequel on peut marquer temporellement les évènements musicaux saillants afin d'informer le modèle analytique. L'auteur distingue deux catégories d'évènements : ceux nécessitant une connaissance théorique musicale et ceux nécessitant des méthodes de traitement du signal (analyse spectrale, reconnaissance de hauteur ...). L'interaction entre les deux modèles n'est pas unidirectionnelle, des aller-retours entre les deux modèles peuvent être nécessaires car ils s'informent mutuellement.

Hist (2004) propose une procédure pour l'analyse des musiques acousmatiques qui est dérivée de la synthèse de l'approche *top-down* "orientée connaissances" et de l'approche *bottom-up* "orientée données". La procédure d'analyse peut être divisée en plusieurs étapes distinctes qui ne sont pas forcément réalisées séquentiellement : ségrégation des objets sonores, intégration horizontale, intégration verticale, assimilation et signification. L'étape de ségrégation vise à identifier les objets sonores. L'intégration horizontale identifie les flux d'objets (motifs) et l'intégration verticale étudie la création et variation de timbres. La dernière étape d'assimilation et signification étudie la nature et le type de discours, l'implication, la réalisation et l'organisation globale dans le temps. Dans Hist (2005), l'auteur propose d'associer sa méthode d'analyse à une représentation graphique sous la forme d'un score d'étude interactif pour l'analyse des musiques électroacoustiques.

Dans Bossis (2006), l'auteur propose de rechercher les invariants des musiques électroacoustiques afin de trouver les conditions d'une méthode d'analyse systématique. L'article est principalement constitué d'une étude approfondie des divers paradigmes propres à la musique électroacoustique : absence d'instruments connus, représentation de la musique, analyse d'une musique dont les variables de hauteur, temps, timbre ne sont pas discrètes (à l'opposé de la musique conventionnelle). Bossis aborde également l'état de l'art des théories d'analyse de ces musiques ainsi que les progrès en traitement du signal permettant

d'apporter des solutions à certains sous-problèmes. La conclusion de ce travail est d'abord l'importance d'une catégorisation des pièces électroacoustiques en fonction de leur époque à la manière du répertoire de la musique occidentale qui distingue la période baroque, classique, romantique etc. Ainsi un groupe de documents choisis peut être étudié afin d'en extraire des invariants. Le musicologue doit ensuite trouver un modèle formel pour chaque pièce étudiée en gardant en tête le contexte global du groupe de documents choisis.

Couprie démontre que la représentation graphique peut être un outil analytique parfaitement adapté aux musiques électroacoustiques (Couprie (2004)). L'auteur conclut que l'association de sons, graphiques et textes permise par les documents multimédias permet d'élargir le champ afin de présenter les oeuvres aussi bien aux néophytes qu'aux spécialistes. Dans Couprie (2006), une catégorisation des concepts importants de la représentation graphique analytique est réalisée afin de mettre en relief les éléments à considérer lors d'une publication multimédia dédiée à l'analyse électroacoustique.

Geslin & Lefevre (2004) présentent l'Acousmographe, un logiciel dédié à la création de représentations graphiques pour les musiques électroacoustiques (ou plus généralement les musiques non écrites au sens traditionnel du terme). Développé depuis 1991 par le *Groupe de Recherches Musicales*, le logiciel en est actuellement à la version 3 et est utilisé dans les écoles et conservatoires. L'Acousmographe permet d'éditer ses propres symboles graphiques afin de réaliser des représentations de timbre, de sons et de structures personnalisées. Les figures 2.5 et 2.6 sont des exemples de représentations (des acousmographies) de deux mouvements de la pièce *Labyrinthe !* de *Pierre Henry*. On peut remarquer la vue globale située en haut de chaque acousmographie qui donne une idée de la structure du mouvement.

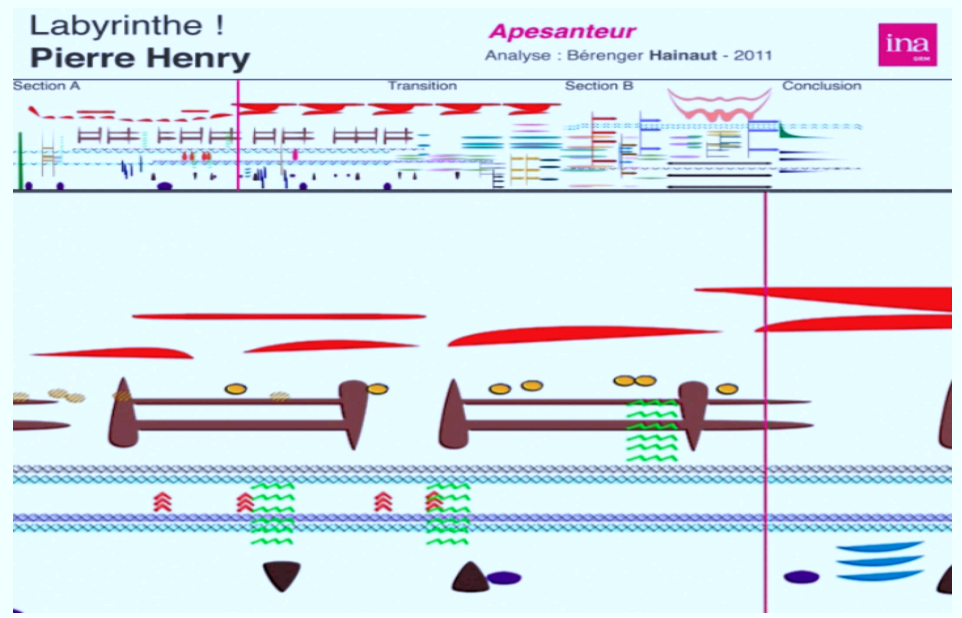

Fig. 2.5 – Acousmographie de la pièce *Labyrinthe !* de *Pierre Henry* (4ème mouvement, *"Apesanteur"*), travail réalisé par Béranger Hainaut.

Dans Gayou (2006), l'auteur présente les portraits polychromes, une série de livres associés à des documents multimédias en grande partie réalisés avec l'Acousmographe et disponibles sur le site internet du *Groupe de Recherches Musicales*[6] depuis 2001. Les

[6] `http://www.inagrm.com/accueil/collections/portraits-polychromes`

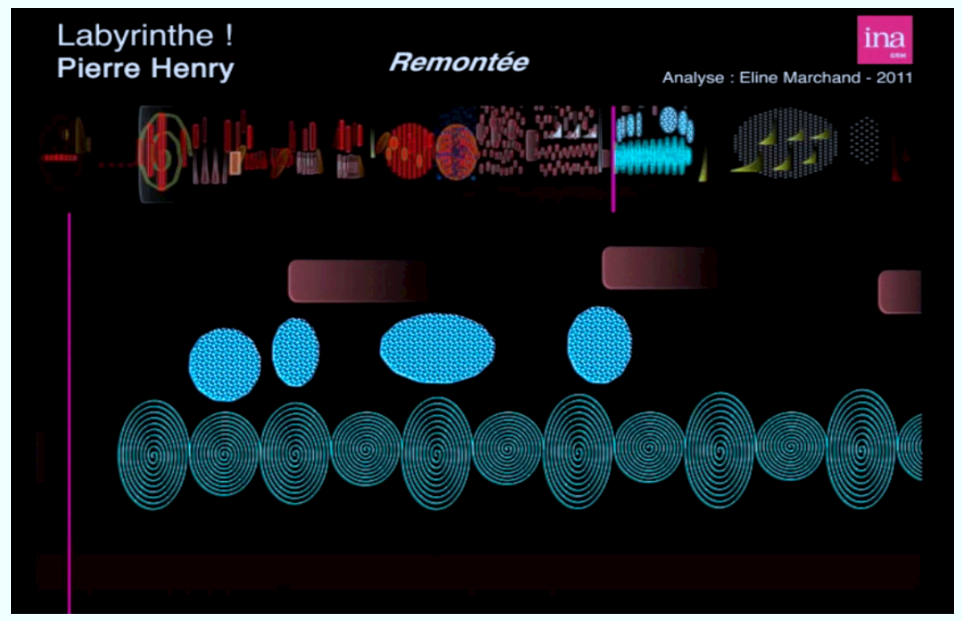

FIG. 2.6 – Acousmographie de la pièce *Labyrinthe !* de *Pierre Henry* (10ème mouvement, *"Remontée"*), travail réalisé par Eline Marchand.

transcriptions multimédias présentées explorent les différentes méthodes d'analyse et de transcription utilisées pour l'analyse de compositions électroacoustiques. Les portraits polychromes abordent également les questions suivantes :

- Quelle est la relation entre la partition et la transcription ?
- Quel est le statut de la transcription dans le processus de communication musicale ?
- La musique électroacoustique peut-elle être écrite sur "partition" ?
- Quelle est la relation entre la représentation graphique et l'écriture ?
- Quelle contribution la transcription graphique apporte-t-elle au genre électroacoustique en particulier ?

Cet état de l'art montre qu'il existe des outils théoriques pensés par des musicologues pour l'analyse des musiques électroacoustiques. **Le but de cette thèse n'est pas de présenter un nouveau modèle d'analyse mais d'apporter une assistance logicielle aux musicologues pour mettre en pratique leurs méthodes personnelles.**

2.4.2 Approche analytique de trois musicologues

Dans la dimension applicative de notre travail, la première préoccupation est d'apporter des solutions logicielles à des problèmes pratiques récurrents rencontrés par les musicologues. Ainsi, trois entretiens ont étés réalisés avec des spécialistes de l'analyse des musiques électroacoustiques :

- *Pierre Couprie*[7] : Maître de conférence à l'IUFM - Université de Paris-Sorbonne.
- *François Delalande*[8] : Groupe de Recherches Musicales de 1970 à 2006, d'abord chef de travaux de recherche, puis directeur de recherche, responsable des recherches en

[7] `http://www.pierrecouprie.fr/`
[8] `http://www.francois-delalande.com/`

Sciences de la Musique.
– *Cyrille Delhaye* : chargé de cours à l'Université de Rouen et chercheur affilié au GRHIS (Groupe de Recherche en HIStoire).

Un questionnaire a été réalisé pour les entretiens, il porte à la fois sur l'analyse pure et sur la représentation. Ces questions correspondent à des interrogations personnelles qui n'ont pas de réponses directes dans la littérature. Nous avons donc écrit les questions dans le but de nous informer sur l'aspect "pratique" de l'analyse, lever certaines ambiguïtés et obtenir des suggestions. Les réponses obtenues permettent d'orienter les choix et spécifications du système. Les trois entretiens ont étés enregistrés et ils durent entre 45 minutes et 2 heures. Dans cette section, nous présentons une synthèse des réponses aux questions et comparons les points de vue afin de repérer les invariants dans les pratiques d'analyse. La synthèse des entretiens a été réalisée de façon à rester focalisé sur les questions posées : éviter les digressions, redondances et les hésitations qui nuisent à la compréhension.

L'entretien est divisé en plusieurs grands thèmes avec une question principale et parfois des sous questions complémentaires. Les grands thèmes abordés sont les suivants :

1. Aspect méthodologique
2. Approche poïétique et esthésique
3. Rapport avec l'analyse tonale
4. L'outil informatique
5. Perception sonore et représentations sonores

1. *Avez-vous une méthodologie générale pour l'analyse des musiques électroacoustiques ?*

- Quelles sont selon vous les grandes étapes d'analyse ?
- Passez-vous forcément par une annotation détaillée de tous les éléments de la pièce ?
- Avez-vous une idée précise de ce que vous voulez mettre en évidence avant de commencer l'analyse ? Ce point de vue peut-il changer en cours d'analyse ?

Pierre Couprie : *J'écoute l'oeuvre plusieurs fois jusqu'à repérer les éléments saillants. Cela peut être des sons, des mouvements dans l'espace, plein de choses différentes. J'utilise des marqueurs sur l'Acousmographe. Chaque fois que j'écoute, je crée une nouvelle couche et je marque des éléments qui me semblent ressortir. Je fais au moins une dizaine d'écoutes, cela prend du temps car je laisse passer 2 ou 3 jours entre chaque écoute pour pouvoir passer à autre chose et masquer ce que j'ai déjà marqué. Ensuite je cherche ce que je souhaite analyser car on ne peut pas tout mettre dans une seule analyse, il faut choisir un point de vue. Si je ne connais pas l'oeuvre que je vais analyser, je pars sans a priori. Je ne sais pas ce que je vais analyser et je commence toujours par l'étape d'écoute pour voir ce qui pourrait être intéressant à analyser. Une fois que j'ai choisi une direction, j'analyse les différents éléments et cela devient relativement classique. Mon point de vue n'est pas influencé par des éléments extérieurs (notes du compositeur, notice etc.), chacun a sa propre vision de l'oeuvre.*

François Delalande : *Je vais d'abord déterminer de quel point de vue je vais analyser la pièce ce qui est valable aussi bien pour les pièces écrites que la musique électroacoustique. Il m'arrive parfois d'utiliser des méthodologies différentes. Par exemple si vous prenez les*

unités sémiotiques temporelles, le point de vue est déterminé par une problématique particulière (en l'occurrence, l'analyse du temps). Il est important de déterminer également les pertinences. Dans le cas de l'étude du temps, on peut par exemple s'intéresser au caractère cyclique. Ensuite, une fois le point de vue et les pertinences déterminées, on utilise presque toujours une transcription. En général, je transcris après la détermination des points de vue pour ne pas être orienté par l'analyse. Je réalise toujours une transcription de repérage la plupart du temps en objets sonores (unités morphologiques). Sur cette toile de fond, je vais par la suite ajouter et décrire des traits qui vont m'aider à analyser par rapport au point de vue que j'ai choisi initialement. Aujourd'hui, on pratique toujours la transcription avec une écoute instrumentée : on utilise un instrument d'écoute (un lecteur de CD, l'Acousmographe etc...) qui nous permet d'affiner la transcription en donnant la possibilité de revenir en arrière, de ralentir ou filtrer dans le cas de l'Acousmographe. Je ne change pas de point de vue en cours d'analyse. Si je veux prendre un autre point de vue, je reprend depuis le départ car il est important pour la clarté de la méthodologie de séparer les points de vus. Il est possible de réaliser plusieurs points de vue pour une même pièce.

Cyrille Delhaye : *Je n'ai pas de méthodologie générale, je pense que chaque pièce est totalement différente et j'essaie d'adapter les outils que j'ai à ma disposition en fonction de ce que je veux analyser. Chaque analyse est différente et j'utilise à chaque fois une méthodologie différente. Par contre, il y a des pratiques qui reviennent souvent : l'écoute acousmatique (écoute "noire") sans représentation qui est très importante, mais cela dépend également de la longueur de la pièce car nos capacités de mémorisation diminuent si la pièce est trop longue. Je réalise plusieurs séries d'écoutes acousmatiques. Pour une pièce de 5 minutes, j'écoute la pièce 4 à 5 fois de suite en prenant des notes à chaque fois. Je construis souvent l'analyse à partir de cette écoute. Il peut également arriver que j'analyse une pièce dont j'ai entendu parler, dans ce cas j'ai déjà étudié de la littérature à son sujet et cela va guider mon écoute. On pourrait dire, si on se place d'un point de vue sémiologique que c'est de la "poïétique externe". Ce sont les écrits des compositeurs qui m'amènent à l'analyse et jamais l'inverse. Je vais chercher dans un premier temps des sections (grandes périodes) dans la pièce. Je ne vais pas tout annoter, mais je vais rechercher les objets sonores avec des factures très fortes et facilement identifiables. J'arrive rarement avec une écoute totalement "blanche", je fais souvent plus attention à certains objets sonores en fonction de ce que j'ai lu en amont et donc je ne suis pas complètement détaché pour mon analyse. Par contre, il est possible que je change mon point de vue de départ, mes hypothèses, en cours d'analyse en fonction de ce que je vais découvrir. Je pense que c'est primordial.*

Pour cette première question de méthodologie, on remarque que les trois musicologues réalisent une transcription partielle des objets sonores les plus saillants. Couprie et Delhaye avouent écouter la pièce de 4 à 10 fois, en laissant passer quelques jours entre les écoutes pour Couprie. Les trois musicologues utilisent une écoute instrumentée à un moment ou à un autre. Ils parlent également tous les trois de l'importance de trouver un *"point de vue"*. Par contre, à ce sujet, il est plus difficile de déterminer un invariant sur l'ordre des étapes méthodologiques car les trois musicologues ont des approches assez différentes : Couprie utilise sa première transcription pour dégager un point de vue pertinent, Delalande a déjà trouvé un point de vue avant de transcrire, Delhaye utilise la transcription ou des écrits pour trouver un point de vue. De plus Delhaye semble accorder de l'importance au changement de point de vue en cours d'analyse alors que les deux autres musicologues préfèrent s'attacher à chaque point de vue séparément.

2. *Une analyse exclusivement esthésique vous semble-t-elle suffisante pour mettre en évidence la construction d'une pièce électroacoustique ?*

Pierre Couprie : *Une analyse exclusivement esthésique me semble tout à fait pertinente pour les musiques électroacoustiques. La plupart du temps on ne peut pas rencontrer le compositeur pour discuter et il n'a d'ailleurs pas forcément envie de parler de sa pièce non plus. De plus, nous n'avons pas non plus accès aux rushs*[9]*. Donc je pense qu'on peut faire abstraction des éléments extérieurs.*

François Delalande : *Oui, je pense qu'une analyse exclusivement esthésique peut permettre de mettre en évidence une construction de la pièce mais bien sur il en existe plusieurs. Il faut séparer les analyses poïétiques et esthésiques, on ne doit pas les mélanger dans une même analyse mais par contre il peut être intéressant de regarder les interférences entre les deux dans un second temps.*

Cyrille Delhaye : *Je pense que les deux approches sont complémentaires. Etant avant tout musicologue, je suis très attaché à l'histoire, aux écrits et par conséquent je commence souvent par la poïétique. J'utilise beaucoup les notes de programme, brouillons du compositeur, les réactions dans la presse.*

Cette question met en valeur l'opposition entre les "écoles d'analyse". On peut noter que Couprie et Delhaye démarrent leur analyse de façons différentes : Couprie utilise uniquement l'enregistrement de la pièce et Delhaye se base souvent sur les écrits en premier lieu. Delalande s'intéresse aux interférences entre les deux approches.

3. *Peut-on transposer les approches d'analyse tonale aux musiques électroacoustiques ?*

- En musique tonale, les motifs mélodico-harmoniques sont répétés pour assurer l'unité et la compréhension/assimilation de la pièce. Selon vous, peut-on retrouver des systèmes de répétitions de motifs dans la musique électroacoustique ? Sur quels aspects s'expriment ces répétitions ?

Pierre Couprie : *Il y a des choses qui sont transposables, notamment au niveau de la structure : on retrouve les structures de type ABA, alternance couplet/refrain, thèmes et variations. Globalement il s'agit des structures qui reviennent le plus souvent dans les pièces électroacoustiques. Il y a également des rapports au niveau du contrepoint, les règles de contrepoint de la musique classique s'appliquent également à la musique électroacoustique. L'analyse harmonique n'est bien sur pas transposable. Par contre, on peut retrouver des cadences dans la musique électroacoustique mais cela ne fonctionne pas de la même façon que dans la musique tonale. Les UST (Unités Sémiotiques Temporelles) arrivent assez bien à décrire ces fonctions. La superposition des plans sonores est un axe d'analyse intéressant pour les musiques électroacoustiques, François Bayle utilise souvent ce principe. Le principe de répétition s'exprime également en musique électroacoustique par exemple avec un son complexe qui revient ponctuellement ou un enchainement de sons. Il ne faut pas prendre cette répétition au sens strict comme la réexposition d'une mélodie : par exemple dans le cas d'un crescendo le mouvement général du crescendo peut être répété mais pas forcément*

[9] analogie au vocabulaire cinématographique dans lequel les *rushs* sont la totalité des plans filmés pendant le tournage, ici ce terme désigne les sons avec lesquels le compositeur travail sur la pièce.

à l'identique. On peut dire que l'analyse de la musique électroacoustique est différente de celle de la musique tonale mais il existe quelques gros "archétypes" qui sont communs.

François Delalande : *On ne peut pas transposer directement les méthodes d'analyse tonale à la musique électroacoustique cependant quand on a déterminé un point de vue et des pertinences, on rentre dans un cas ou la grille d'analyse est à peu près configurée, il ne reste plus qu'à l'appliquer. Les UST sont un bon exemple : on dispose d'une grille d'analyse connue, il s'agit ensuite de l'appliquer. Si vous pensez à la musique écrite (*e.g. *la musique classique), il faudrait plutôt penser à l'analyse d'un enregistrement de la pièce car il n'y a pas que les accords, il y a aussi tout ce que rajoute l'interprète etc... C'est ce qui nous rapproche des musiques électroacoustiques. Dans le cas des musiques électroacoustiques, on trouve parfois des réitérations (répétitions) sous forme de simples copies mais c'est assez rare et surtout la répétition dépend également du point de vue.*

Cyrille Delhaye : *Je ne pense pas que cela soit la direction où il faut chercher. Nous avons besoin d'outils souples qui peuvent s'adapter à chaque pièce. Je pense que la force de la musique électroacoustique c'est justement de s'être libérée de ces carcans théoriques et le fait d'appliquer une méthode très normative peut à mon avis tuer le geste créateur et la liberté apportée au compositeur dans cette musique. Contrairement à la musique tonale, je pense qu'en musique électroacoustique, la répétition pure n'existe pas, j'ai plutôt rencontré des compositeurs qui citaient les mêmes objets sonores mais en les variant. Par contre, il est intéressant de voir que Pierre Henry réutilise des objets sonores, qu'il a enregistré dans les années 50, dans des pièces des années 2000.*

Au sujet du rapport avec la musique tonale, on apprend principalement que la méthodologie standard de la musique tonale n'est pas transposable directement. Il s'agit plutôt de trouver dans un premier temps la bonne méthodologie avant de l'appliquer. En ce qui concerne la répétition de motifs, les avis convergent également : la répétition à l'identique d'un même objet est assez rare, la reprise variée d'un même objet est plus fréquente.

4. *Qu'attendez-vous de l'outil informatique pour vous assister dans votre analyse ?*

Pierre Couprie : *Par exemple, j'aimerais pouvoir repérer les différentes itérations d'un même son dans l'acousmographe, c'est ce qui m'intéresserait le plus. Il pourrait également être intéressant d'essayer de repérer des séries de sons un peu comme dans la musique sérielle. Je souhaiterais également avoir un outil qui me ferait des propositions de segmentation à plusieurs niveaux de précision un peu comme dans les logiciels de musique tel que "Live" qui segmente automatiquement les sons.*

François Delalande : *Il serait intéressant de pouvoir avoir une sorte de fond de carte de la pièce (pour les objets saillants) pour pouvoir ensuite continuer l'annotation à la main. Je pense qu'on peut aller assez loin dans l'automatisation de l'analyse des contrastes, des registres de hauteurs, des grains. Cela pourrait être très utile car certains sonagrammes sont parfois difficiles à exploiter. Je pense qu'on peut automatiser le repérage des sons vu qu'on ne pose pas le problème des pertinences à ce moment. Il s'agit plus d'avoir un certain confort de lecture. Par contre, dans une seconde étape, il me semble important de pouvoir réaliser des symboles graphiques à la main comme dans l'Acousmographe afin de pouvoir*

par exemple étirer ou contracter les symboles si certains objets sont plus longs.

Cyrille Delhaye : *Ce que je recherche dans l'outil informatique c'est une "caution scientifique" : je lui demande une vérification de mes hypothèses analytiques. Par exemple lorsque j'ai essayé* Sound Spotter *(un outil de recherche de sons par similarités), j'ai trouvé des occurrences de sons que je n'avais pas perçues. Evidemment, on rêve tous d'un outil qui permettrait de séparer les différentes voix de mixage d'une pièce pour voir comment le compositeur a réalisé l'assemblage des sons entre eux. Il serait également intéressant d'avoir un outil qui permettrait de pouvoir trouver automatiquement les grandes périodes d'une pièce mais cela me semble un peu moins important. Avoir un outil pour isoler les objets sonores entres eux serait déjà une grande aide pour mes travaux.*

Au sujet des apports de l'outil informatique pour l'analyse, les musicologues ont des demandes assez diverses qui correspondent en fait à leurs habitudes d'analyse qui peuvent être assez différentes. Ainsi, on peut remarquer les propositions suivantes : repérage de grandes périodes ou séries de sons (Delhaye et Couprie), séparation des voix de mixage (Delhaye), utilisation de symboles graphiques personnalisés (Delalande). Les trois musicologues expriment le besoin d'avoir un outil leur permettant de repérer les objets sonores principaux.

5. *Est-ce que vous utilisez une représentation visuelle (forme d'onde, spectrogramme etc.) du signal sonore pour vous aider à démarrer votre analyse ? Les informations de représentation ne risquent-elles pas d'influencer votre analyse ?*

Pierre Couprie : *J'utilise le spectrogramme dès la première écoute de repérage (j'ai déjà écouté la pièce sans support auparavant). Il y a certaines oeuvres pour lesquelles le spectrogramme ne donne rien du tout mais elles sont assez rares. Il est vrai que cela peut influencer l'analyse par rapport à une écoute pure mais cela n'est pas gênant car les deux approches sont complémentaires. Par exemple, le spectrogramme peut révéler la façon dont certains sons complexes sont construits, ce qui est très informatif. Le spectrogramme peut représenter également des détails qu'on n'entend pas mais dans ce cas je n'en tiens pas compte dans l'analyse. Il faut toujours donner la priorité à l'oreille.*

François Delalande : *Je commence toujours par une écoute pure. Dans mes travaux d'analyse esthésique, je mets de côté mon écoute personnelle : je fais écouter à des personnes que j'enregistre et je recoupe les informations afin de repérer les témoignages qui se rejoignent. Je commence la transcription uniquement après avoir distingué les points de vue c'est-à-dire environ trois mois après. J'utilise alors une représentation graphique (le sonagramme). Les perceptions ne sont pas influencées par le support visuel car j'ai déjà des points de vue analytiques en amont et j'utilise les représentations par la suite comme des outils.*

Cyrille Delhaye : *Je commence mon analyse par une écoute sans support visuel. Ensuite j'utilise le spectrogramme et la forme d'onde pour structurer mon analyse et pour m'aider à me repérer dans le document sonore. J'ai fait l'expérience avec mes élèves de leur faire découvrir une pièce en leur montrant le spectrogramme en même temps et ils sont très influencés par le support visuel. Je pense que pour découvrir une pièce, l'écoute*

pure permet une perception plus intéressante. Les outils visuels sont intéressants pour nous aider à comprendre la musique mais dans un second temps.

La comparaison des trois réponses fait apparaitre clairement l'importance de l'écoute pure sans support visuel pour la découverte de la pièce. Le support visuel est par la suite utilisé pour aider la transcription. La réponse à cette question met en avant le fait qu'il serait utile d'avoir accès à une représentation de type spectrogramme et/ou forme d'onde dans le système final.

Les figures 2.7 et 2.8 résument les informations importantes apprises lors des entretiens qui viennent d'être présentés. Ces informations seront utilisées pour l'élaboration du système.

Invariants dans les pratiques d'analyse des personnes interrogées :

- Ecoute de la pièce entre 4 et 10 fois pour localiser les objets sonores les plus saillants ;
- Elaboration de points de vues ;
- Pour certains, la transcription en objets sonores aide à former un point de vue alors que d'autres en ont déjà un en amont ;
- La transcription s'effectue principalement sur les objets saillants ou ceux qui éclairent un point de vue d'analyse.

FIG. 2.7 – Bilan des invariants

Attentes et suggestions des musicologues pour un outil d'analyse :

- Localiser les objets principaux d'une pièce et "cautionner" la transcription de ces derniers ;
- Trouver toutes les instances d'un objet sonore donné ;
- Séparer des différentes voix de mixage ;
- Repérer les grandes périodes ;
- Repérer les séries de sons ;
- Proposer une segmentation de la pièce ;
- Découvrir des instances d'objets qu'ils n'avaient pas remarquées.

FIG. 2.8 – Bilan des souhaits et suggestions

2.5 Un système interactif d'aide à l'analyse des musiques électroacoustiques

Dans la section précédente, nous avons cherché à dégager les besoins réels des musicologues qui analysent les musiques électroacoustiques. Le système proposé constitue une première pierre à l'édifice, nous ne prétendons pas pouvoir répondre à toutes les attentes énoncées. De plus nous nous focalisons sur les musiques qui peuvent être analysées en les décomposant en objets sonores. Dans cette section, nous présentons les systèmes d'analyse existants ainsi que les choix concernant l'architecture de notre système.

2.5.1 Etat de l'art

Il existe des logiciels dédiés à l'annotation musicale dans un but analytique (Couprie (2008), Puig et al. (2005), Geslin & Lefevre (2004)). iAnalyse permet de synchroniser une partition avec un fichier audio, de visualiser des paramètres musicaux et d'annoter la partition à partir de divers objets graphiques (Couprie (2008)). Dans Puig et al. (2005), un logiciel d'éducation musicale qui contient un module d'annotation musicale et de synchronisation audio/partition est proposé (ML-Annotation). Ces deux logiciels ont en commun d'être dédiés aux musiques écrites et d'utiliser la notation musicale traditionnelle. Par conséquent ils ne sont pas adaptés aux musiques qui nous concernent dans ce travail car les méthodes d'analyse sont différentes comme nous l'avons vu dans la section 2.4.2.

L'Acousmographe qui a déjà été cité précédemment est dédié à l'annotation des musiques non écrites (Geslin & Lefevre (2004)). Il constitue un outil d'annotation bien implanté dans la communauté mais ne dispose actuellement pas de modules permettant de réaliser des classifications d'objets sonores par exemple. Park et al. (2009) propose EASY, un système d'aide à l'analyse des musiques électroacoustiques sous la forme d'un programme Matlab qui est principalement dédié à la représentation des descripteurs et qui se focalise en particulier sur le timbre. Outre les représentations classiques de type forme d'onde ou autres spectrogrammes, EASY propose également de visualiser le timbre en trois dimensions. Les trois axes de représentations sont ceux décris dans McAdams et al. (1995) comme étant les plus pertinents pour décrire le timbre (on peut également affecter un attribut quelconque aux axes). On peut d'ailleurs remarquer que l'étude effectuée dans McAdams et al. (1995) ne s'applique que dans un contexte monophonique (notes isolées), ce qui est une limite importante dans le cas des pièces électroacoustiques qui sont majoritairement polyphonique.

Des travaux ont étés proposés afin de réaliser une *description morphologique* du signal audio (Ricard & Herrera (2004), Peeters & Deruty (2008)). La notion de *description morphologique* est introduite par *Schaeffer*, elle désigne la description de la forme d'un objet sonore. Dans sa théorie, *Shaeffer* décrit les critères morphologiques comme des caractères observables dans l'objet sonore, des traits distinctifs ou encore des propriétés de l'objet sonore perçu. Théoriquement le nombre de critères observables est infini mais *Schaeffer* en a limité le nombre à sept :

- *Critères de matière :* masse, timbre harmonique
- *Critères d'entretien :* grain, allure
- *Critère de forme :* dynamique
- *Critères de variations :* profil mélodique, profil de masse

On peut noter que la description morphologique peut constituer une information utile pour l'analyse musicale (François Delalande parle de repérer des unités morphologiques dans la section 2.4.2). Cependant, les travaux proposés s'intéressent à des objets sonores individuels or dans notre cas nous souhaitons traiter des polyphonies d'objets. Dans Nucibella et al. (2005), la méthode de description morphologique de Ricard & Herrera (2004) est appliquée à une pièce électroacoustique. Le temps de calcul pour la description morphologique est important : 3 heures pour un segment de 2 minutes. De plus, les auteurs mentionnent que cette méthode a été conçue pour analyser des objets sonores dans un contexte monophonique et que la pièce testée comporte beaucoup de polyphonie. Ainsi, la

	Contexte musical	Mode d'analyse	Polyphonie	Complexité
iAnalyse	conventionnel	manuel	oui	temps-réel
ML-Annotation	conventionnel	manuel	oui	temps-réel
Acousmographe	électroacoustique	manuel	oui	temps-réel
EASY	électroacoustique	automatique	non	non connue
Descr. morphologique	électroacoustique	automatique	non	importante

FIG. 2.9 – Tableau récapitulatif des systèmes existants

description morphologique donne de bons résultats dans les passages monophoniques mais les résultats se dégradent fortement dans les passages comportant beaucoup de polyphonie. Dans notre cas, cette approche est de toute façon trop lente car nous souhaitons permettre à l'utilisateur de réaliser des interactions avec le système dans un temps acceptable.

Les caractéristiques des systèmes existants sont résumés dans le tableau 2.9. **Il est important de préciser qu'aucun des systèmes existants ne permet d'analyser les objets sonores de manière semi-automatique et dans un contexte polyphonique. L'objectif de cette thèse est de combler ce manque.**

2.5.2 Architecture du système

2.5.2.1 Contraintes fonctionnelles

Nous avons remarqué dans le chapitre 2 que les musicologues passent systématiquement par une première série d'écoutes pendant laquelle ils vont repérer les objets sonores auxquels ils s'intéressent. Il est important de considérer la notion de "point de vue" qui peut être déterminé avant les écoutes de repérage ou bien après quelques écoutes. Le point de vue est également propre à chaque musicologue pour une analyse donnée d'où l'importance de considérer l'aspect subjectif du problème : le système doit s'adapter au point de vue d'analyse de l'utilisateur. On peut également remarquer que les musicologues vont s'intéresser en particulier aux objets sonores saillants et ne vont pas réaliser une transcription complète de la pièce dans un premier temps.

Une des attentes principales des musicologues porte sur l'identification des différentes instances des objets sonores principaux de l'ensemble de la pièce. **Ainsi, dans ce travail de thèse, nous cherchons à assister le musicologue dans le repérage d'objets en l'aidant à retrouver leurs différentes instances à partir d'une instance initiale.** La figure 2.10 illustre le problème pour retrouver les différentes instances de l'objet de couleur verte.

Pour élaborer l'architecture de notre système, nous devons prendre en compte les contraintes suivantes :

- **Nature indéterminée des sons utilisés par les compositeurs :** nous ne pouvons pas nous baser sur des grandes bases de signaux audio pour apprendre les sons. Nous devons donc forcément utiliser des échantillons sonores de la pièce pour apprendre les classes sonores.
- **Polyphonie des pièces musicales :** la plupart des pièces sont polyphoniques et donc il faut considérer la superposition des objets sonores. Autrement dit, un segment audio contenant l'objet sonore cherché peut également en contenir d'autres qui lui

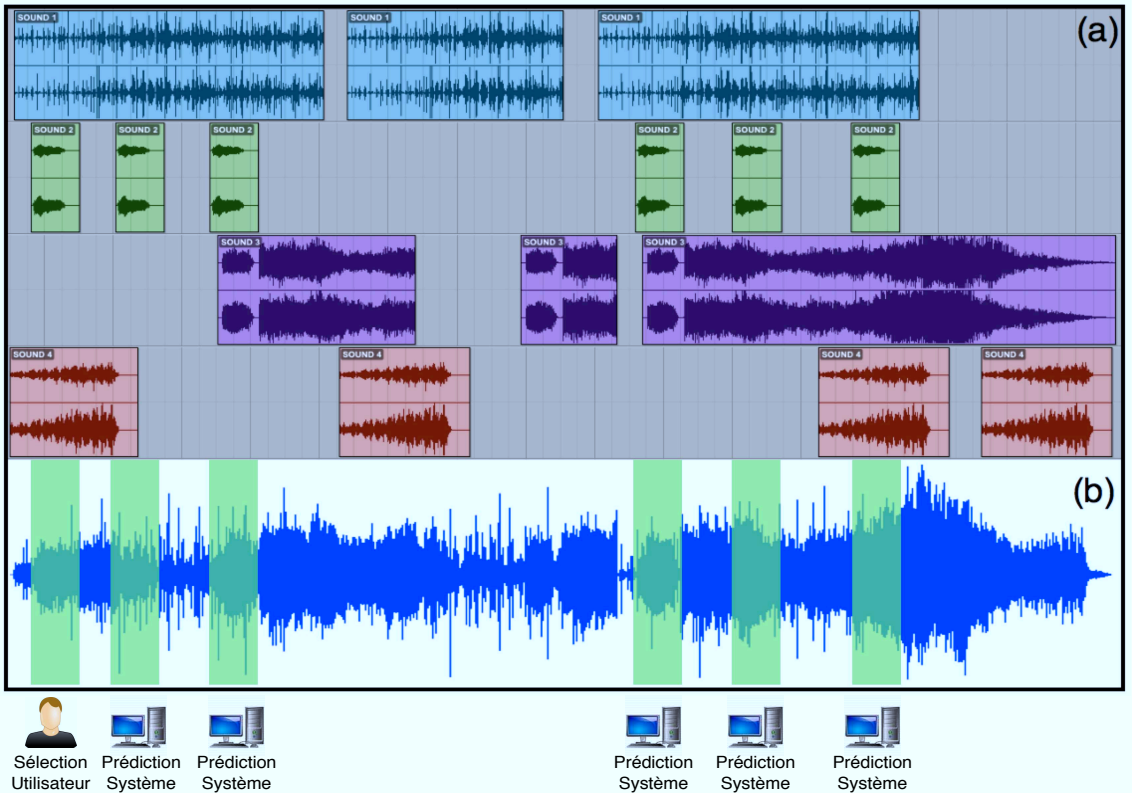

Fig. 2.10 – La figure (a), représente la superposition de diverses sources sonores dans une pièce musicale (un son différent par ligne/couleur), comme c'est le cas dans une pièce électroacoustique polyphonique. La figure (b) est le mixage résultant de toutes les sources sonores, lors de l'analyse nous n'avons accès qu'à ce mélange de sources. Le système doit pouvoir prédire les différentes instances d'un objet donné (en l'occurrence le son vert) à partir de l'instance de la sélection utilisateur.

sont superposés.

- **Adaptation au point de vue d'analyse :** l'utilisateur doit pouvoir exprimer les objets auxquels il s'intéresse.
- **Réactivité :** les composants du système doivent être suffisamment rapides pour que le système soit réactif aux interactions de l'utilisateur.

2.5.2.2 Choix d'architecture

Nous proposons une approche en deux temps pour le système :

1. *Segmentation timbrale* en unités sonores homogènes, pour initialiser le système.
2. *Classification* des objets sonores visés par l'utilisateur.

La figure 2.11 illustre les différents composants de l'architecture du système proposé.

La *segmentation timbrale* effectuée avant la classification est une étape importante de notre approche. Comme nous l'avons vu dans la section 2.4.2, les utilisateurs du système repèrent les objets saillants lors des premières écoutes. L'intérêt de cette *segmentation timbrale* est de leur faciliter le choix et la découverte des instances initiales de classes

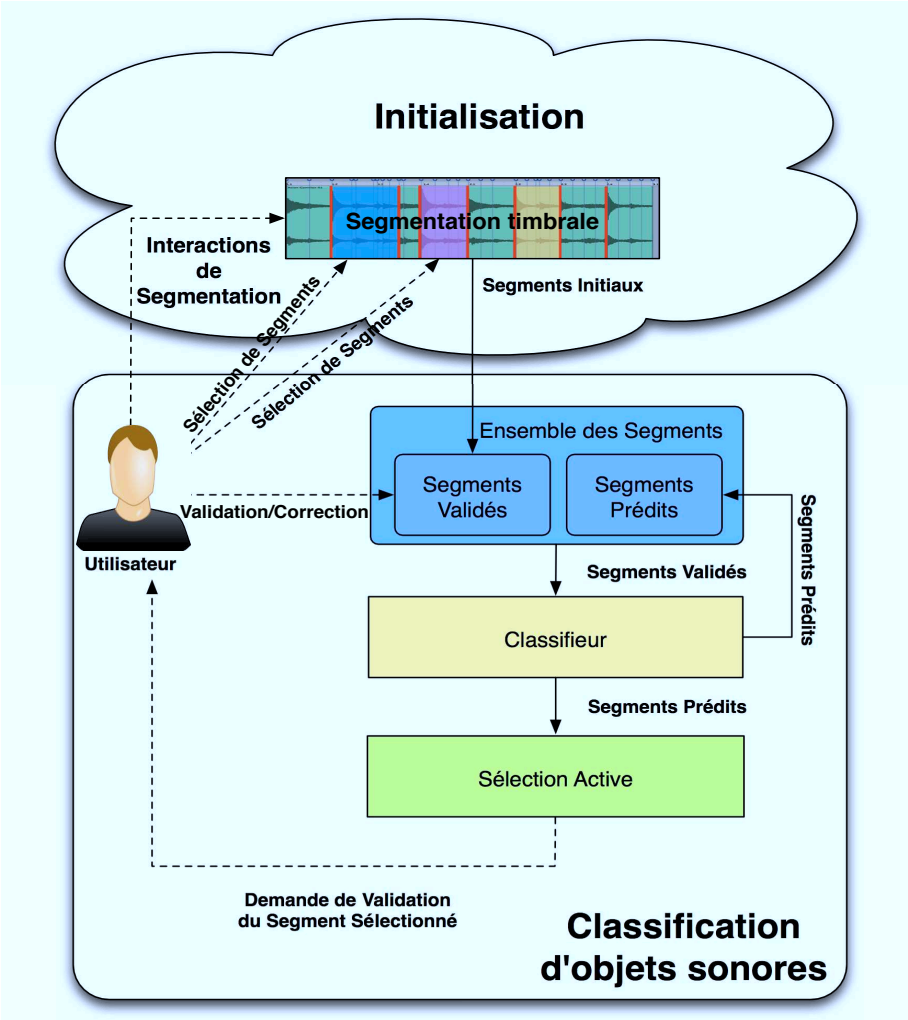

FIG. 2.11 – Architecture globale du système

qui seront utilisées pour initialiser la classification. Nous avons choisi le timbre comme critère de segmentation car il s'agit d'un des aspects les plus structurant des musiques électroacoustiques. Les objets sonores se trouvent à des échelles temporelles différentes et la *segmentation timbrale* permet à l'utilisateur d'écouter des "mixtures sonores" homogènes. De plus, l'approche de segmentation que nous proposons apporte également une information de similarité timbrale entre les segments afin de pouvoir les comparer et de choisir les instances initiales de façon à ce qu'elles soient représentatives. Nous détaillerons l'approche de segmentation dans le chapitre 3.

L'étape de classification des objets permet d'affecter des étiquettes aux différents segments de la pièce. L'étiquette d'un segment correspond aux objets visés par l'utilisateur présents dans le segment. Comme nous l'avons abordé dans les contraintes fonctionnelles, étant donné que les pièces sont polyphoniques, les segments sonores peuvent contenir plusieurs objets visés. L'approche de classification que nous proposons doit donc permettre de réaliser un multi-étiquetage des segments sonores. Autrement dit, le système doit pouvoir prédire pour chaque segment les différentes classes sonores auxquelles le segment appartient. De plus, étant donné la nature indéterminée des sons utilisés par les compositeurs,

la tâche de classification est relativement difficile car nous ne disposons que des segments sonores de la pièce choisis par l'utilisateur pour démarrer l'apprentissage. Une approche par *retour de pertinence* est adaptée au problème car elle permet d'intégrer le jugement de l'utilisateur au fur et à mesure afin de faire progresser les prédictions du *classifieur*. Pour bénéficier du *retour de pertinence*, le système sélectionne des segments que l'utilisateur va pouvoir écouter afin de valider/corriger les prédictions du *classifieur*. La classification est ensuite remise à jour en fonction des informations apportées par l'utilisateur. Ainsi, une boucle d'interaction est mise en place et la classification progresse à chaque itération jusqu'à ce que l'utilisateur soit satisfait des prédictions. La sélection des segments présentés à l'utilisateur par le système se base sur l'*apprentissage actif* qui est une méthode permettant de sélectionner les segments les plus utiles pour l'apprentissage. Nous détaillerons la phase de classification dans le chapitre 4. La figure 2.12 présente un scénario typique d'utilisation du système.

1. Initialisation
   (a) L'utilisateur interagit avec le système afin d'obtenir une segmentation adaptée à la pièce considérée.
   (b) L'utilisateur sélectionne le segment qu'il considère comme caractéristique pour chaque classe sonore.
2. Classification des objets
   (a) Le système réalise une classification en apprenant à partir des segments validés par l'utilisateur. Ainsi, des étiquettes sont prédites automatiquement pour les parties restantes de la pièce.
   (b) Afin d'améliorer la classification, le système réalise la *sélection active* d'un segment et demande à l'utilisateur de valider/corriger les prédictions d'étiquette.
   (c) Les étapes (a) et (b) sont répétées jusqu'à satisfaction de l'utilisateur

Fig. 2.12 – Etapes d'un scénario d'utilisation du système

2.5.3 Corpus synthétique

L'évaluation de notre système n'est pas une tâche simple, notamment en ce qui concerne la recherche d'une vérité terrain. Les annotations de certaines musiques électroacoustiques existent mais la plupart d'entre elles ne font pas la différence entre la description des évènements sonores et l'interprétation musicologique. De plus l'annotation de ce type de musique requiert l'expertise de spécialistes qui sont beaucoup plus rares que les personnes capables d'annoter de la musique classique ou tout autre style plus conventionnel. Ayant connaissance de cette réalité, nous avons décidé de générer un corpus d'évaluation synthétique. Un des grands avantages de ce choix est de pouvoir générer de nombreuses pièces différentes et simultanément l'annotation correspondante ce qui permet de rendre l'évaluation plus robuste. Nous avons ainsi généré deux types de corpus qui seront utilisés pour les évaluations présentées dans les chapitres suivants. Le premier corpus, *Corpus M*, est monophonique et le deuxième, *Corpus P*, est polyphonique et par conséquent plus complexe.

2.5.3.1 Corpus M

Ce premier corpus est le plus simple des deux de par sa nature monophonique. Il a été utilisé en début de thèse pour l'évaluation du système de segmentation timbrale.

Pour la création de ce corpus, nous partons d'une pièce de musique concrète annotée manuellement : un extrait de "Timbre Durée" d'*Olivier Messiaen* a été choisi. Cette pièce peut être considérée comme un archétype car elle utilise des enchainements de timbres et de mixtures pour créer une pièce musicale à la structure complexe. Une propriété importante de cette pièce est qu'elle est monophonique : seulement un timbre/mixture est exposé à la fois. Cette propriété est adaptée à notre problème initial de segmentation : créer des frontières entre les enchainements de mixtures qui constituent une pièce musicale.

Le corpus synthétique a été généré par concaténation de sons extraits à partir de deux banques d'échantillons sonores. La première, fournie par l'INA est une collection de sons environnementaux et sons d'ambiances (applaudissements, ambiances urbaines, sonneries de téléphone etc.). La deuxième est une banque d'échantillons très utilisée en recherche : il s'agit de la partie instrumentale de la base RWC, Goto et al. (2002), qui contient la plupart des instruments de l'orchestre. Une description détaillée des échantillons utilisés pour la génération peut être trouvée en annexe A.1.

En suivant les principes de construction de la pièce "Timbre Durée", un algorithme de génération a été élaboré. Ainsi pour créer une pièce synthétique, un échantillon est choisi arbitrairement et un segment est sélectionné aléatoirement à l'intérieur de l'échantillon avant d'être concaténé au segment précédent. Les segments sont de durée arbitraire : de moins d'une seconde à 5 secondes pour les plus longs. Le processus de génération est répété itérativement pour générer des "pièces" synthétiques de 30 secondes. Le corpus total compte 1000 pièces de 30 secondes dont 200 d'entre elles sont utilisées pour la sélection d'attributs (voir section 3.3.3.2) et les 800 restantes pour le test de l'algorithme de segmentation présenté dans la section 3.4.2.

2.5.3.2 Corpus P

Ce deuxième corpus, à polyphonie variable a été réalisé à un stade plus avancé de la thèse afin d'évaluer le système complet.

Comme nous le savons maintenant, dans les musiques électroacoustiques, on peut rencontrer n'importe quelle source sonore acoustique ou électronique. Ce constat nous amène à nous poser des questions pour le choix des sons constituant les pièces synthétiques. Nous proposons de choisir des sources sonores de difficulté réaliste qui pourraient être utilisées dans des compositions. Pour cette raison, nous avons utilisé des sons sélectionnés par des compositeurs du GRM. Une description détaillée des échantillons utilisés pour la génération peut être trouvée en annexe A.2.

Pour la génération de pièces synthétiques, nous cherchons à créer des polyphonies complexes d'objets sonores. Ainsi, nous choisissons d'utiliser des sons complexes ayant une évolution temporelle. Autrement dit, il ne s'agit pas de notes ou de séquences de notes à hauteur déterminée. Trois compositeurs du GRM ont participé à la sélection des sons qui pour la plupart viennent d'enregistrements personnels et ont été sélectionnés indépendamment, sans intention compositionelle particulière. La contrainte principale pour la sélection était de choisir des sons qui gardent des caractéristiques timbrales relativement stables afin de pouvoir les considérer comme une classe unique. Les trois compositeurs ont

sélectionné un total de 24 sons (par conséquent 24 classes possibles) pour la génération. Les caractéristiques principales des sons sélectionnés sont les suivantes :

- Leur longueur est variable, elle se situe dans des ordres de la seconde à la minute ;
- ils peuvent être construits à partir de l'agrégation de sons élémentaires plus petits ;
- ou encore par la superposition de plusieurs sons élémentaires.

Afin d'étudier l'influence de la polyphonie sur les performances de l'algorithme, pour chaque pièce synthétique, cinq versions ont été générées avec un degré de "difficulté polyphonique" progressif. La première version de chaque pièce est monophonique et la cinquième version superpose un maximum de cinq sons simultanément. Par conséquent, pour la $i^{\text{ème}}$ version d'une pièce P_i, nous avons un maximum de i sons superposés. Dans la suite de ce document, nous définissons le "degré de polyphonie" comme le nombre maximal d'éléments superposés dans une version de pièce. Au total, 100 pièces de 2 minutes ont été générées avec 5 versions polyphoniques différentes pour chacune (soit un total de 500 fichiers synthétiques).

Le processus de génération des pièces est plus complexe que dans le cas monophonique (Corpus M). La figure 2.13 illustre le processus de génération utilisé pour créer des séquences de sons artificielles. Ainsi, pour réaliser les séquences, on choisit arbitrairement 5 sons parmi les 24 disponibles dont on extrait des segments aléatoirement. Cette première étape nous permet d'obtenir différentes instances d'une même classe sonore. Ensuite, cinq couches sonores sont réalisées en alternant les différentes instances de chaque classe et des silences de durée aléatoire. Finalement, les différentes couches sont mixées en tenant compte du degré de polyphonie de chaque version. Dans les pièces synthétiques, on considère les différentes instances d'un même son comme des objets sonores de même classe.

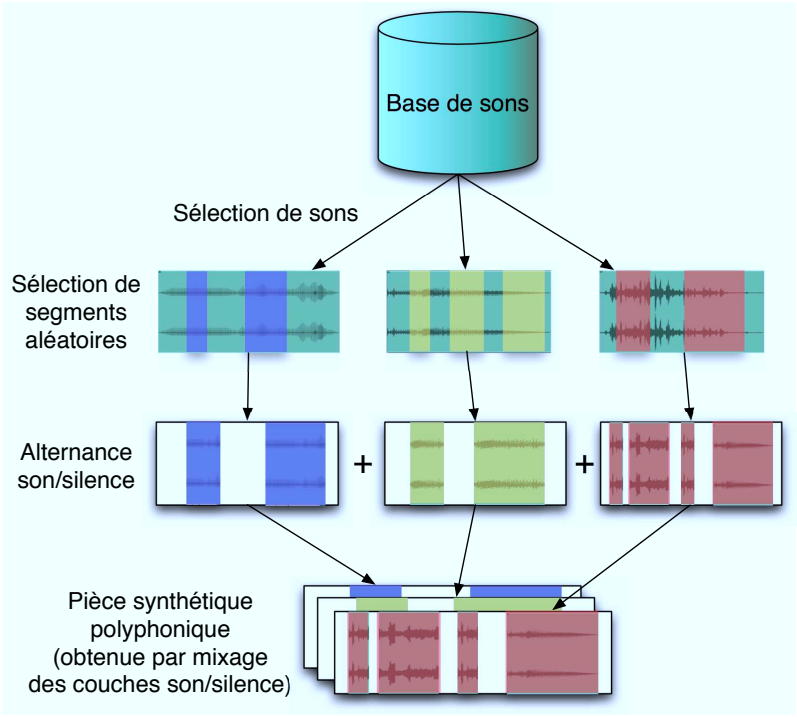

Fig. 2.13 – Processus de génération des pièces synthétiques

2.6 Conclusion

Dans ce chapitre, nous avons décrit le type de musiques auxquelles le système est destiné. Nous avons également abordé l'analyse des musiques électroacoustiques telle qu'elle est pratiquée par des spécialistes afin de mettre en valeur leurs attentes. Certaines pratiques d'analyse sont partagées par les personnes interrogées et nous ont permis de proposer une application utile ainsi que de dégager les contraintes fonctionnelles applicatives. Nous en avons déduit un choix d'architecture en adéquation avec les différentes contraintes et objectifs. Les deux corpus utilisés pour l'évaluation du système ont également étés présentés. Le chapitre suivant expose la première étape d'initialisation du système basée sur une segmentation timbrale interactive.

Chapitre 3

Segmentation interactive de musiques électroacoustiques

Sommaire

3.1 Introduction

Comme nous l'avons vu dans le chapitre précédent, on peut distinguer deux phases interactives dans notre système : la *segmentation timbrale* et la *classification d'objets*. Ce chapitre se concentre sur la première phase qui peut être assimilée à une segmentation en unités sonores homogènes. Pour réaliser cette segmentation, nous axons cette initialisation sur un des aspects les plus structurants de la musique contemporaine : le timbre. La segmentation timbrale a pour objectif de faciliter le choix et la découverte des instances initiales de classes qui seront utilisées pour initialiser la classification. Cette phase nous permet également de connaitre les frontières principales entre les superpositions de timbres d'une pièce polyphonique afin de pouvoir classifier des unités sonores homogènes.

Comme nous l'avons évoqué dans la section 2.5.2.1, nous n'avons pas de connaissances a priori sur les différents sons de la pièce. Par conséquent, nous souhaitons obtenir une segmentation timbrale homogène de manière non supervisée . La taille des segments que l'on souhaite obtenir est variable : de l'ordre de la seconde dans la plupart des cas mais pouvant atteindre la dizaine de secondes dans certaines pièces présentant de longues trames. La nature non supervisée du problème est due à la diversité des timbres qui peuvent être rencontrés. Comme nous l'avons vu dans le chapitre 2 consacré en partie à la présentation des musiques à objets, les sources sonores rencontrées dans ce type de musique ne sont pas préétablies comme c'est le cas dans d'autres styles musicaux : la musique symphonique est orchestrée à partir des grandes familles instrumentales traditionnelles (cordes, bois, vents et percussions), les musiques à tendance rock utilisent souvent une formation de type : guitare, basse, batterie, chant et les musiques à tendance électronique s'appuient principalement sur les synthétiseurs ou autres sonorités d'origine électronique. Dans le cas qui nous intéresse, on peut dire que toute source sonore instrumentale rencontrée dans un style musical quelconque peut être retrouvée dans les musiques à objets. De plus, ces musiques élargissent le spectre des sources possibles aux sons environnementaux ou de manière plus générale à tout ce qui est du domaine du sonore (sources acoustiques ou électroniques).

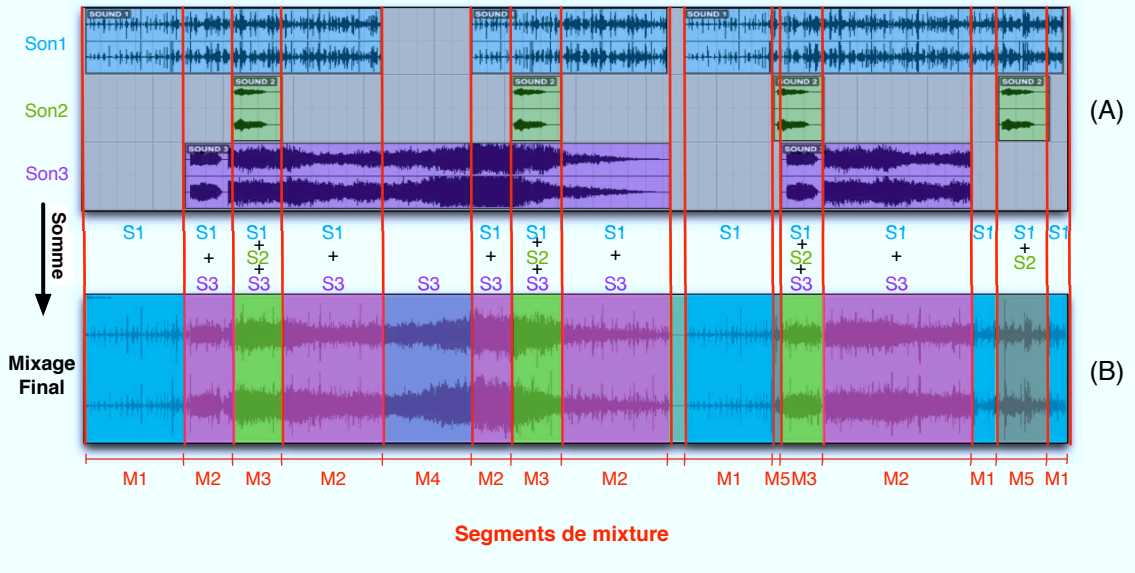

Fig. 3.1 – Segmentation d'un mixage sonore en segments de mixture

La problématique principale de cette phase de segmentation est illustrée par la figure 3.1. A partir d'un mixage de plusieurs sources sonores, on cherche à identifier les différentes superpositions possibles de timbres et marquer les frontières entre ces dernières. Ainsi, la superposition de *Son1* et *Son3* (figure 3.1 (A)) constitue la mixture *M2* qui apparait plusieurs fois dans le mixage (figure 3.1 (B)). Il peut arriver que certaines mixtures ne soient composées que d'un seul son : c'est le cas de la mixture *M1* qui n'est composée que du *Son1* dans la figure 3.1.

Nous commencerons par envisager des segmentations inspirées de l'état de l'art (exposé dans la section 3.2) qui montrera des limites que nous dépasserons par une approche interactive comme décrit dans la section 3.3. Après avoir présenté l'état de l'art et décrit le système, nous étudierons deux scénarios d'interaction différents. Le chapitre se terminera par une évaluation suivie d'une discussion des résultats obtenus.

3.2 État de l'art

Dans cette partie, nous présentons l'état de l'art sur la structuration/segmentation du signal musical. Dans la plupart des méthodes existantes, des descripteurs de timbre bas-niveau sont utilisés pour étiqueter la musique conventionnelle selon des sections de haut niveau (introduction, couplet, refrain, pont). Les différentes approches sont exclusivement automatiques et par conséquent, elles n'utilisent pas le retour utilisateur.

3.2.1 Approches par mesures de similarités

De nombreuses méthodes utilisées pour la structuration musicale exploitent des représentations de type *matrices de similarités* (Foote (2000), Peeters et al. (2002), Cooper (2002), Cooper & Foote (2003), Lu et al. (2004)). La première publication ayant exploité cette voie est Foote (2000). Dans cet article, l'auteur propose d'utiliser une matrice de similarité (figure 3.2) pour mesurer la nouveauté dans le signal audio. Cette matrice est calculée en mesurant les distances euclidiennes entre tous les couples de trames possibles. Ainsi, on obtient une matrice symétrique avec une distance minimale sur la diagonale qui représente la distance d'une trame par rapport à elle-même. Une courbe mesurant la nouveauté dans le signal est déduite de la matrice dont les maximas locaux indiquent les endroits du signal ou la nouveauté est notable. Pour obtenir cette courbe, une fenêtre d'analyse est utilisée le long de la diagonale de la matrice et mesure la nouveauté du signal en se basant sur des propriétés locales de la matrice. Cette méthode a pour avantage d'être totalement non supervisée et de ne faire aucune hypothèse sur la nature des signaux à traiter. Une limite de cette méthode est qu'elle ne permet pas d'étiqueter directement les régions similaires du signal extraites à partir des maxima locaux de la courbe de nouveauté. Autrement dit, cette approche permet uniquement de trouver des frontières temporelles entre les différentes parties du signal.

Dans Goto (2003), Bartsch & Wakefield (2001, 2005), Van Steelant et al. (2002), une représentation de type temps/retard équivalente à la *matrice de similarité* est utilisée. Cette représentation transforme les répétitions représentées par des diagonales dans la matrice de similarité en des lignes de retard horizontales constantes. Dans Goto (2003), l'auteur utilise cette dernière représentation pour découvrir automatiquement la structure de morceaux

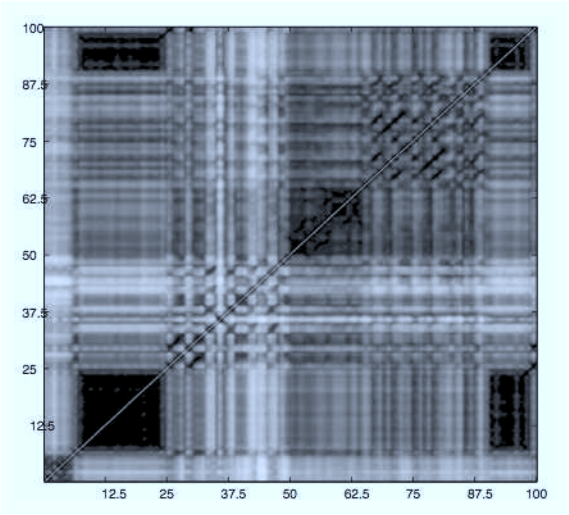

Fig. 3.2 – Un exemple de matrice de similarité. Les deux axes représentent le temps. Les distances entre les trames sont représentées par des niveaux de gris. En l'occurence les grandes distances sont affectées à un niveau sombre et les faibles distances à un niveau clair.

de musique en se concentrant sur la recherche des refrains. Ainsi, des descripteurs de type chromas (Bartsch & Wakefield (2005)) sont extraits du signal à partir desquels la *matrice temps/retard* est calculée. Les sections musicales similaires sont détectées grâce à un critère de répétition. On peut noter que cette méthode prend en compte la modulation de tonalité : deux passages similaires musicalement à une transposition près seront considérés comme proches. Cette méthode obtient des résultats satisfaisants sur une base importante (80 chansons étiquetées correctement sur 100) mais est assez éloignée de notre problème : elle se positionne clairement dans un cadre musical conventionnel en prenant la tonalité et les notes qui la constituent comme hypothèses de base.

3.2.2 Approches par détections de ruptures

Il existe plusieurs méthodes de détections de ruptures : Bayesian Information Criterion (BIC), Kernel Change Detection (KCD), Kernel Fisher Discriminent analysis (KFDA) entre autres (Cettolo & Vescovi (2003), Desobry et al. (2005), Harchaoui et al. (2009)). Ces méthodes sont souvent utilisées pour détecter les changements dans un flux audio (par exemple détecter les changements parole/musique). Le principe général de ces méthodes est de comparer des distributions dans une fenêtre d'analyse : soit une fenêtre d'analyse F contenant n échantillons du signal étudié : $F = (x_1, ..., x_n)$.Pour détecter la présence d'une rupture à un indice τ du signal, les distributions des fenêtre $F1 = (x_1, ..., x_{\tau-1})$ et $F2 = (x_\tau, ..., x_n)$ sont comparées. Les maximas locaux sont considérés comme des ruptures. Cette famille de méthodes souffre des mêmes limites que Foote (2000) et ne permet donc pas d'étiqueter directement les segments obtenus. De plus, ces méthodes sont plus adaptées à la reconnaissance des "grandes sections", or dans notre cas nous cherchons à trouver des changements de mixtures sonores relatifs aux différents évènements musicaux.

3.2.3 Approches par programmation dynamique

D'autres approches de structuration audio utilisent la programmation dynamique Chai (2003), Chai & Vercoe (2003), Chai (2005), Maddage et al. (2004), Maddage (2006). Dans Chai (2003), l'extraction de descripteurs est suivie d'une segmentation à longueur fixe de la suite de vecteurs de descripteurs (par exemple : 200 vecteurs consécutifs, soit 2 secondes de musique, avec un recouvrement de 150 vecteurs, soit 1,5 seconde). La répétition des segments est calculée par une méthode de programmation dynamique qui permet d'obtenir les informations de structuration et les limites temporelles de chaque segment. Ce type d'approche a pour inconvénient majeur d'être coûteuse en temps de calculs : le nombre d'opérations à effectuer augmente rapidement en fonction de la longueur du document à traiter. Dans notre cas, nous souhaitons une remise en forme rapide des données après une intervention utilisateur or cette méthode semble trop rigide pour les applications qui nous concernent.

3.2.4 Approches par clustering

Logan & Chu (2000) proposent d'utiliser une approche de *clustering hiérarchique agglomératif* (nous aborderons cet algorithme en détail dans la section 3.3.5) pour retrouver les refrains dans de la musique conventionnelle. Après avoir extrait les descripteurs, les séquences de vecteurs d'attributs sont divisées en segments contigus de même longueur. Ces segments sont considérés comme étant les clusters initiaux. Une mesure de distorsion est par la suite calculée entre chaque paire de clusters et les deux clusters de la paire ayant la plus faible distorsion sont ensuite fusionnés. Le calcul des distorsions entre les paires de clusters est ensuite répété jusqu'à atteindre un certain seuil. A la fin du processus agglomératif, chaque trame audio dispose d'une étiquette correspondant au cluster à laquelle elle appartient. Le refrain est ensuite trouvé en considérant que le cluster qui contient le plus de trames est celui du refrain. La méthode proposée est comparée à une approche par *modèle de Markov caché*. L'évaluation montre que la méthode proposée est la plus performante.

Peeters et al. (2002) utilise la segmentation de base obtenue par une *matrice de similarité* pour générer des classes potentielles ainsi qu'une approximation du nombre de classes. Ensuite, les résultats obtenus sont utilisés pour initialiser un algorithme de clustering (*K-means*). Enfin les clusters obtenus sont utilisés à leur tour pour initialiser un *modèle de Markov caché* et la représentation optimale du document est déduite par décodage du modèle.

Dans Levy et al. (2006), le problème de structuration est reformulé comme un problème de clustering. Avant d'effectuer le clustering, les vecteurs de description originaux subissent un changement de représentation. Ainsi, un *modèle de Markov caché* est appris sur ces données puis décodé afin d'obtenir une suite d'états. Le nombre d'états est fixe et représente le nombre de catégories de timbre différents dans le document. Une estimation de la longueur d'un temps (unité temporelle de base dépendant du tempo du morceau de musique) est également effectuée. Ensuite, des histogrammes d'états sont calculés à des intervalles réguliers et alignés sur les temps de la musique. Les histogrammes obtenus représentent des distributions de types de timbres décodés. Enfin, une méthode de clustering (*soft k-means*) est utilisée pour regrouper les histogrammes et ainsi en déduire les segments. Cette méthode, de par sa dépendance au tempo n'est pas assez générale pour être appliquée directement à notre problème (nous voulons pouvoir traiter des pièces "arythmiques" comme

c'est souvent le cas dans le style électroacoustique). Cependant, la philosophie générale de cette approche est intéressante car elle fait intervenir un changement de représentation adapté aux données afin de préparer le clustering.

On peut noter qu'EASY, le système présenté dans Park et al. (2009) que nous avons déjà évoqué dans la section 2.5.1 propose également des fonctionnalités de segmentation de la musique. Deux approches simples sont proposées. La première réalise directement un clustering des vecteurs de descripteurs et reporte une couleur différente pour chaque cluster sur la forme d'onde. La deuxième utilise des fenêtres d'analyse longues et compare les distances entre les différentes fenêtres (le système propose plusieurs distances).

3.2.5 Approches issues d'autres domaines

Parmi les approches issues d'autres domaines, on peut mentionner celles concernant l'audio diarisation qui sont intéressantes pour notre problème de segmentation. Dans ce domaine, le problème posé est d'annoter un flux audio en affectant à chaque région temporelle une source sonore spécifique. Les sources peuvent être un locuteur particulier, de la musique, un bruit de fond etc. Un exemple classique d'application est la reconnaissance d'un locuteur particulier dans un flux audio. Tranter & Reynolds (2006) proposent une vue d'ensemble des différents systèmes s'intéressant à ce problème. La plupart sont basés sur les mêmes briques élémentaires. Selon Reynolds et al. (2009), un système typique procède en trois étapes principales. La première étape consiste à détecter les changements dans le signal à partir d'une méthode de détection de rupture (voir section 3.2.2). La seconde étape regroupe les segments de même locuteur ensemble à l'aide d'une méthode de *clustering hiérarchique agglomératif*. Idéalement, le regroupement produit un groupe différent pour chaque locuteur. Le *clustering hiérarchique agglomératif* basé sur un critère d'arrêt de type BIC comporte les étapes suivantes :

0. Initialiser les feuilles de l'arbre avec les segments détectés en amont
1. Calculer les distances entre chaque paire de clusters
2. Fusionner les clusters les plus proches
3. Mettre à jours les distances des paires incluant le nouveau cluster
4. Itérer les étapes 1. à 3. jusqu'à ce que le critère d'arrêt soit atteint

La dernière étape est une re-segmentation itérative basée sur l'algorithme de Viterbi(Viterbi (1967)) pour affiner les points de ruptures et les décisions de clustering.

La structure de cette dernière approche est particulièrement intéressante : d'abord détecter les changements puis regrouper les segments qui sont proches entre eux. Nous utiliserons également une approche en deux étapes dans la phase de segmentation du système en nous adaptant aux signaux concernés.

3.3 Segmentation interactive

L'état de l'art que nous venons de présenter propose un bon nombre de méthodes différentes mais elles sont toutes automatiques. Autrement dit, il n'est pas possible pour ces méthodes de s'adapter à un point de vue utilisateur. A notre connaissance, dans la littérature, il n'existe pas de système de segmentation interactif dans le domaine audio et musical. On pourra cependant noter que certains logiciels audio tel que "Recycle"[1] réalisent

[1] `http://www.propellerheads.se/products/recycle/`

des segmentations interactives de signaux en utilisant les transitoires du signal. Cependant, les applications principales de ces logiciels sont le découpage de boucles de batterie ou autres signaux aux transitoires saillantes. De plus, ces logiciels ne réalisent pas de regroupements par similarité timbrale entre les segments. On peut remarquer que dans les domaines de l'image et de la vidéo, des systèmes de segmentation interactifs ont déjà été proposés (voir Price et al. (2009), Ning et al. (2010) par exemple).

Dans la suite de ce chapitre, nous présentons notre approche de segmentation timbrale. Nous avons remarqué que toutes les méthodes de l'état de l'art sont automatiques et se basent sur des hypothèses fortes propres aux musiques conventionnelles : les répétitions de motifs quasi-identiques sont courantes, les évènements musicaux sont exclusivement des notes issues des échelles musicales standards, la structure est simple et quasiment la même pour tous les morceaux. Dans le cas des musiques électroacoustiques, nous ne pouvons pas prendre ces hypothèses de départ et devons nous situer dans une approche plus générale. Une segmentation adaptative semble essentielle étant donné la diversité des esthétiques sonores potentiellement rencontrées dans les pièces électroacoustiques. De même, la diversité des points de vues possibles pour une même pièce peuvent conduire à des segmentations différentes. Par conséquent, nous proposons un système interactif pouvant s'adapter aux signaux de musiques électroacoustiques qui permet à l'utilisateur d'intervenir sur la segmentation en unités sonores.

3.3.1 Architecture

Le système de segmentation comporte deux phases distinctes comme le montre la figure 3.3 : une *phase d'apprentissage* et une *phase de test* qui font intervenir deux bases de signaux distinctes. Dans la *phase d'apprentissage*, après avoir extrait les descripteurs des signaux de la *base d'apprentissage*, une *sélection d'attributs* est effectuée afin de conserver les plus pertinents. Dans la *phase de test*, les attributs sélectionnés précédemment sont extraits de la *base de test*.

En parallèle, une segmentation de bas-niveau est effectuée à l'aide d'une détection de transitoires. Cette première segmentation permet d'obtenir des segments inter-transitoires dont les attributs seront par la suite intégrés temporellement afin de résumer l'information en un unique vecteur de description pour chaque segment. Les vecteurs ainsi obtenus seront par la suite regroupés par similarité timbrale à l'aide d'un algorithme de *clustering*. L'interaction de l'utilisateur avec le système se situe au niveau des regroupements effectués par l'algorithme de *clustering*. Il est important de noter que les regroupements effectués définissent la segmentation temporelle finale ainsi que l'étiquetage des segments. Les sections suivantes décrivent les détails de chaque étape de la segmentation.

3.3.2 Extraction de descripteurs

Il n'existe pas de descripteur universel pour décrire le timbre. Par conséquent, la stratégie que nous avons adoptée consiste à sélectionner un ensemble d'attributs parmi un grand nombre de descripteurs décrivant les différents aspects du timbre musical. Ne connaissant pas la nature des sources sonores composant les signaux qui nous intéressent, il semble pertinent d'effectuer la sélection d'attributs parmi un grand nombre de descripteurs. Ainsi, nous avons effectué l'extraction de descripteurs Spectraux, Cepstraux, Temporels et Perceptifs. Le tableau 3.3.2 résume l'ensemble des descripteurs extraits pour la phase de segmentation timbrale. Les paragraphes qui suivent décrivent les descripteurs extraits de façon succincte car ils sont pour la plupart standard dans la communauté. Une présentation détaillée des

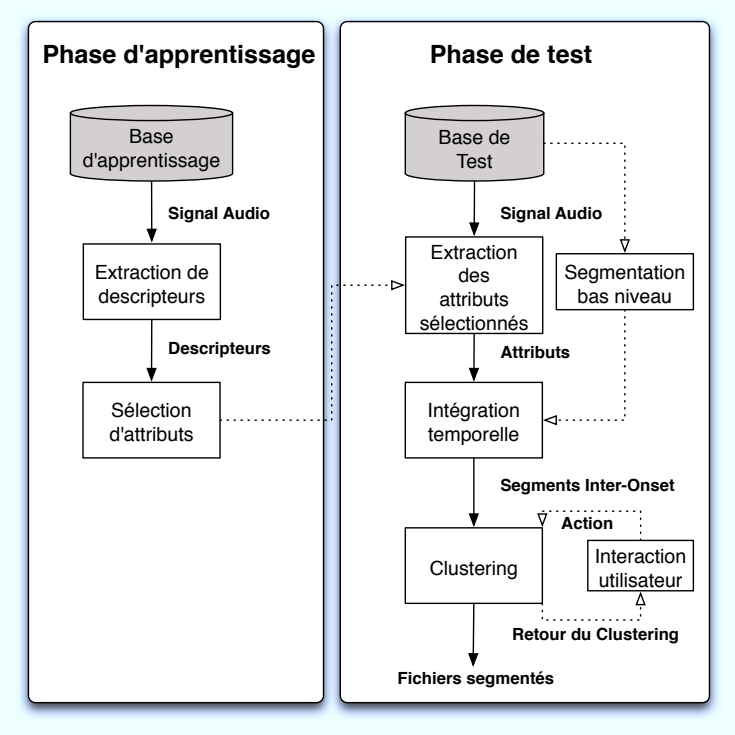

Fig. 3.3 – Architecture du système de segmentation interactif

descripteurs utilisés peut être trouvée en annexe B.

Tous les descripteurs ont été extraits sur des fenêtres d'analyse de 20ms avec un pas de recouvrement de 10ms. On dénombre un total de 279 attributs extraits avant la sélection automatique.

Descripteurs Spectraux

Les descripteurs spectraux sont calculés à partir du spectre estimé par une *Transformée de Fourrier à Court Terme*. Voici ceux qui ont été extraits :

- Les *moments spectraux*, sont calculés en considérant le spectre comme une distribution. Les 4 premiers moments du spectre sont calculés.
  - *centroïde spectral* : barycentre du spectre (valeur moyenne)
  - *largeur spectrale* : étalement du spectre autour de la valeur moyenne
  - *asymétrie spectrale* : mesure l'asymétrie de la distribution autour de la valeur moyenne
  - *platitude spectrale* : mesure la platitude de la distribution autour de la valeur moyenne
- La *platitude d'amplitude spectrale et facteur de crête spectrale par bandes* : mesure les proportions relatives de bruit et de composantes sinusoïdales du spectre sur plusieurs bandes de fréquences.
- La *platitude spectrale globale* : mesure les proportions relatives de bruit et de composantes sinusoïdales sur l'ensemble du spectre.
- La *pente spectrale* : représente le taux de décroissance spectrale.

- La *décroissance spectrale* : mesure la décroissance des amplitudes spectrales.
- La *fréquence de coupure* : fréquence à partir de laquelle 95% de l'énergie du spectre a été mesurée.
- La *modulation d'amplitude* : caractérise les phénomènes de trémolo ou encore la rugosité d'un son.
- Les *coefficients de prédiction linéaire* (ou LPC pour Linear Predictor Coefficients) : permettent de représenter l'enveloppe spectrale d'un signal de façon compressée (voir Makhoul (1975)).
- Les *OBSI (Octave band signal intensity)* : intensité du signal par bande d'octave proposé par Essid (2005).

Descripteur	**Dimension**	**Type**	**Annexe**
Centroïde spectral	1	Spectral	B.1
Largeur spectrale	1	Spectral	B.1
Asymétrie spectrale	1	Spectral	B.1
Platitude spectrale	1	Spectral	B.1
Platitude d'amplitude spectrale par bandes	23	Spectral	B.1
Facteur de crête spectrale par bandes	23	Spectral	B.1
Platitude spectrale globale	1	Spectral	B.1
Pente spectrale	1	Spectral	B.1
Décroissance spectrale	1	Spectral	B.1
Fréquence de coupure	1	Spectral	B.1
Modulation d'amplitude	8	Spectral	B.1
Coefficients LPC	2	Spectral	B.1
OBSI	8	Spectral	B.1
Coefficients MFCC	13	Cepstral	B.2
Coefficients cepstraux à Q constant	114	Cepstral	B.2
Taux de passage par zero	1	Temporel	B.3
Moments statistiques temporels	4	Temporel	B.3
Coefficients d'autocorrélation	49	Temporel	B.3
Loudness spécifique	24	Perceptif	B.4
Acuité perceptive	1	Perceptif	B.4
Etalement perceptif	1	Perceptif	B.4

FIG. 3.4 – Ensemble des descripteurs extraits pour la phase de segmentation timbrale.

Descripteurs Cepstraux

Le *Cepstre* se définit comme la *Transformée de Fourier* inverse du logarithme du spectre d'amplitude. Les descripteurs cepstraux suivants ont été extraits :

- Les *MFCC (Mel Frequency Cepstral Coefficients)* : basés sur l'échelle des fréquences de Mel qui modélise le système auditif humain. Les 13 premiers coefficients sont extraits.
- Les *coefficients cepstraux à Q constant* : calcul du cepste en tenant compte des gammes musicales occidentales tempérées su plusieurs résolutions (résolutions d'une, la moitié, un tiers et un quart d'octave).

Descripteurs Temporels

Ces descripteurs sont calculés directement à partir des trames du signal :
- Le *taux de passage par zero* : nombre de fois que le signal change de signe.
- Les *moments statistiques temporels* : comme pour le spectre, les moments d'ordre 1 à 4 sont calculés sur les trames du signal.
- Les *coefficients d'autocorrélation* : représentent la distribution spectrale dans le domaine temporel.

Descripteurs Perceptifs

- La *loudness spécifique* : coefficients de mesure de l'intensité perceptive à partir des bandes de fréquences de l'échelle de Bark.
- L'*acuité perceptive* : version perceptive du centroïde spectral calculée à partir de la loudness spécifique.
- L'*étalement perceptif* : mesure l'écart entre la loudness spécifique maximale et la loudness totale.

3.3.3 Construction d'un descripteur de timbre adapté

Dans cette étape, nous cherchons à sélectionner un ensemble d'attributs pertinents pour la description du timbre.

3.3.3.1 Algorithme de Fisher

Nous utilisons l'algorithme de sélection dit de Fisher qui exploite un score dérivé de l'*Analyse Linéaire Discriminante* (voir Duda et al. (2001)) pour la sélection automatique d'attributs. Dans cet algorithme, on cherche à conserver les attributs (directions dans l'espace de description) utiles à une bonne discrimination des classes. Dans le cas bi-classe, l'algorithme sélectionne itérativement les attributs qui maximisent le rapport

$$r = \frac{|\tilde{\mu_1} - \tilde{\mu_2}|^2}{\tilde{\sigma_1}^2 + \tilde{\sigma_2}^2}, \tag{3.1}$$

appellé *Discriminant de Fisher* où $\tilde{\mu_q}$ et $\tilde{\sigma_q}$ sont respectivement la moyenne et la variance empirique de la classe C_q ($1 \leq q \leq 2$). Cette méthode vise à maximiser le rapport entre la dispersion inter-classe et la dispersion intra-classe.

Nous utilisons l'implémentation multiclasse de la toolbox Spider[2] qui procède de la façon suivante :

1) Pour l'ensemble des attributs i (avec $1 \leq i \leq D$), des scores f_i^q sont calculés pour chaque classe C_q (avec $1 \leq q \leq Q$) comme il suit :

$$f_i^q = \sum_{p=1}^{Q} \frac{|\mu_i^p - \mu_i^q|}{\sigma_i^p + \sigma_i^q} \tag{3.2}$$

[2] `http://people.kyb.tuebingen.mpg.de/spider/`

f_i^q correspond à la moyenne non normalisée des discriminants de Fisher dans lesquels la classe C_q intervient. Des discriminants de Fisher sont ainsi calculés pour tous les couples de classes possibles et pour chaque attribut.

2) Les valeurs de discriminants sont ensuite triées par ordre décroissant afin de trouver les d attributs distincts correspondant aux valeurs arrivées les premières dans le tri. Les d attributs trouvés seront ceux retournés par la sélection.

3.3.3.2 Sélection d'attributs

Afin d'obtenir un bon rapport entre le nombre d'attributs sélectionnés et les performances, une expérience préliminaire mesurant les performances de *clustering* a été réalisée en faisant varier le nombre d'attributs sélectionnés entre 10 et 40 sur les 279 attributs initiaux. Il est nécessaire de garder un nombre d'attributs relativement bas pour que l'algorithme reste rapide, cette condition est essentielle dans une approche interactive. Les résultats obtenus montrent qu'un optimum local est atteint pour la sélection de 30 et 40 attributs (figure 3.5). Dans un but d'efficacité, nous choisissons de garder les 30 premiers attributs sélectionnés qui sont décrits dans le tableau 3.3.3.2. On remarque dans cette sélection une répartition homogène des différentes familles de descripteurs. Cependant les descripteurs temporels sont un peu en retrait avec la présence unique du *taux de passage par zéro* dans la sélection.

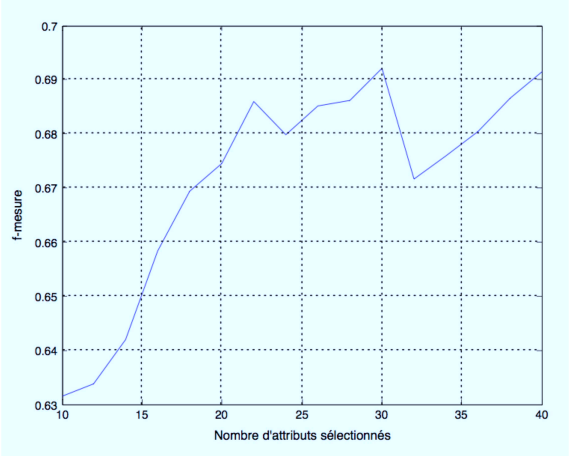

FIG. 3.5 – Choix du nombre d'attributs à garder

On peut noter que la plupart des descripteurs sélectionnés sont parmi les plus employés pour décrire le timbre musical.

Après la sélection d'attributs, on peut associer à chaque trame k un vecteur de description X_k de dimension d (nombre d'attributs choisis)

$$X_k = (x_{k_1}, x_{k_2}, ..., x_{k_d}), \tag{3.3}$$

les suites de vecteurs définissant des segments seront utilisés par la suite pour représenter des unités sonores.

Descripteur	Numéro d'attribut
Centroïde spectral	1
Largeur spectrale	1
Asymétrie spectrale	1
Platitude spectrale	1
Coefficients MFCC	3 à 13
Loudness spécifique	2 à 12, 22
Acuité perceptive	1
Etalement perceptif	1
Taux de passage par zéro	1

Fig. 3.6 – Les 30 attributs sélectionnés pour décrire le timbre

3.3.4 Représentation d'unités sonores

La sélection d'attributs permet d'obtenir une description efficace du timbre pour une trame de signal audio. Cependant, il semble plus pertinent d'un point de vue perceptif de regrouper les trames adjacentes temporellement en unités sonores (Joder et al. (2009)). Dans cette optique, deux étapes sont nécessaires : une étape de segmentation bas-niveau suivie d'une étape d'intégration temporelle pour obtenir une description propre à chaque unité sonore.

3.3.4.1 Segmentation de bas-niveau

Pour la segmentation bas-niveau, nous utilisons la méthode proposée dans Alonso et al. (2005). Il s'agit d'une *détection de transitoire* basée sur le flux d'énergie spectral (dérivée temporelle du spectre). Dans un flux audio, on définit une *transitoire* comme la variation d'une ou plusieurs propriétés psychoacoustiques du signal (timbre, hauteur, amplitude ...). Cette méthode présente l'avantage de ne pas dépendre des descripteurs extraits et donc ne fait pas d'hypothèse sur la nature de la variation qui génère la *transitoire*. En suivant cette approche, le signal est d'abord décomposé en bandes spectrales par une transformée de Fourier à court terme. Chaque bande est par la suite traitée indépendamment pour trouver la position temporelle et l'intensité des transitoires. Les courbes ainsi obtenues sont par la suite sommées pour obtenir une courbe globale à laquelle un seuil est appliqué. La fonction résultante est appelée *fonction de détection* (figure 3.7). Les maximas locaux de la *fonction de détection* sont recherchés par seuillage dynamique.

Des unités sonores sont par la suite déduites de la détection de transitoires en considérant qu'une unité se situe entre deux transitoire (figure 3.7). Chaque segment inter-transitoire définit donc une unité sonore.

3.3.4.2 Intégration temporelle

La segmentation de bas niveau permet de délimiter précisément les unités sonores. Cependant, il faut trouver une stratégie pour représenter ces unités de manière efficace. Le τ^{ime} segment inter-transitoire de longueur L_τ, est définit par l'ensemble S_τ des vecteurs d'attributs qui le constituent :

$$S_\tau = (X_{k_\tau}, X_{k_{\tau+1}}, ..., X_{k_{\tau+L_\tau}}) \tag{3.4}$$

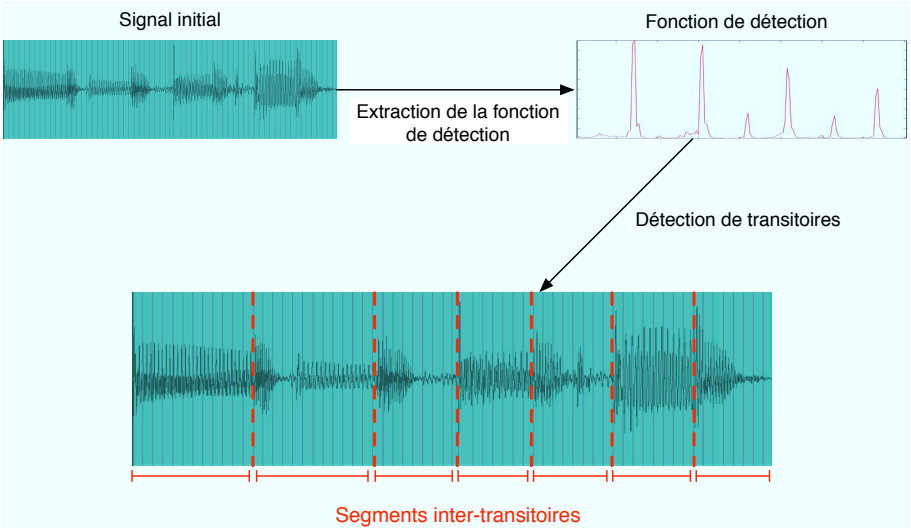

FIG. 3.7 – Détection de transitoires

Une étude spécifique sur le sujet de l'intégration temporelle (voir Joder et al. (2009)) a montré que des opérateurs statistiques simples permettaient d'obtenir des représentations efficaces du signal audio pour décrire des unités sonores. L'intégration temporelle utilisée dans la segmentation en tient compte dans le choix de représentation proposé. Ainsi, le τ^{ieme} segment inter-transitoire est représenté par un vecteur $\tilde{S_\tau}$ de dimension $D = 2d$:

$$\tilde{S_\tau} = (\mu_\tau, \sigma_\tau^2) \tag{3.5}$$

$\tilde{S_\tau}$ est un vecteur constitué de la moyenne μ_τ et de la variance σ_τ^2 du τ^{ieme} segment inter-transitoire :

$$\mu_\tau = \frac{1}{L_\tau} \sum_{k=k_\tau}^{k_\tau+L_\tau-1} X_k, \tag{3.6}$$

$$\sigma_\tau^2 = \frac{1}{L_\tau} \sum_{k=k_\tau}^{k_\tau+L_\tau-1} (X_k - \mu_\tau)^2 \tag{3.7}$$

Par conséquent, chaque unité sonore est représentée par un vecteur $\tilde{S_\tau}$ de dimension D.

3.3.5 Clustering hiérarchique

Nous avons présenté dans la section précédente une approche de représentation des unités sonores. Il s'agit maintenant de les regrouper par similarités afin de faire apparaitre les segments de mixtures illustrés par la figure 3.1. Nous utilisons une approche de *clustering hiérarchique* pour réaliser cette tâche.

Le *clustering* (ou partitionnement de données) est une méthode *non supervisée* qui permet de créer les partitions d'un ensemble en regroupant les données similaires dans les mêmes partitions que l'on assimile à des *clusters*. La méthode de *clustering* que nous utilisons est une approche agglomérative hiérarchisée : chaque vecteur d'unité sonore commence en étant son propre cluster puis les clusters sont fusionnés par paires pour former

un cluster plus gros. Le processus de fusion est répété jusqu'à ce que les deux derniers clusters soient fusionnés. Ainsi, nous obtenons un arbre dont la partie extrême-haute de la hiérarchie (aussi appelée racine) est associée à l'ensemble total des données et symétriquement, les parties extrême-basses (les feuilles) sont associées à un vecteur unique. Le graphe hiérarchique ainsi obtenu se nomme *dendrogramme* (figure 3.8), il peut être vu comme un arbre binaire ou chaque noeud est associé à un ensemble de vecteurs. De plus, le *dendrogramme* a pour avantage de représenter la distance entre les clusters : les longueurs des lignes verticales des branches reliant deux clusters entre eux sont proportionnelles à la distance qui les sépare.

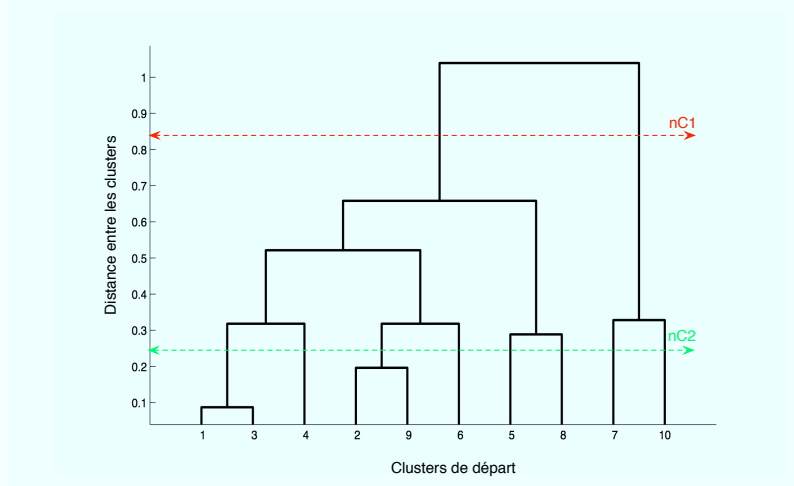

Fig. 3.8 – Exemple de dendogramme

Pour comparer deux clusters, une métrique est nécessaire afin de mesurer la distance entre deux vecteurs de dimension D. Nous utiliserons ici la *distance euclidienne* d_E qui a donné les meilleurs résultats après des essais empiriques :

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{D} |x_i - y_i|^2}; \tag{3.8}$$

Le *clustering hiérarchique* utilise également un *critère de liaison* L_{d_E} pour mesurer la proximité entre deux clusters. Ainsi, pour deux ensembles de vecteurs A et B constituant les clusters à comparer, une distance par paire est utilisée :

$$L_{d_E}(A,B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d_E(a,b), \tag{3.9}$$

avec $|A|$ le cardinal de l'ensemble A.

L'avantage de l'approche hiérarchique est de pouvoir obtenir différents partitionnements des données initiales. En effet, cette propriété se comprend facilement en observant un dendrogramme (figure 3.8) : on peut obtenir un nombre de clusters donné n_C en "coupant" le dendrogramme à un certain niveau de la hiérarchie ($1 \leq n_C \leq n_V$ avec n_V est le nombre

de vecteurs total). Comme le montre la figure 3.8, on obtient un nombre différent de clusters suivant le niveau ou l'on "coupe" le dendrogramme. Si l'on se réfère à la figure 3.8, pour le niveau $nC1$ représenté en rouge on obtient 2 clusters : le cluster de gauche contient les vecteurs 1, 3, 4, 2, 9, 6, 5 et 8 ; celui de droite contient les vecteurs 7 et 10. Pour le niveau $nC2$ représenté en bleu, on obtient les 8 clusters suivants : (1, 3), (4), (2, 9), (6), (5), (8), (7), (10). Nous utiliserons et étendrons cette propriété par la suite afin d'améliorer les performances de clustering.

Une fois le clustering réalisé, nous pouvons obtenir une segmentation timbrale du signal audio en donnant à l'algorithme de clustering le nombre de clusters souhaités en entrée. Les unités sonores sont regroupées par similarité : des segments se créent lorsque plusieurs unités sonores contigües temporellement sont regroupées dans un même cluster. De même, une frontière apparait entre deux unités sonores lorsqu'elles se succèdent temporellement mais appartiennent à des clusters différents.

3.3.6 Clustering interactif

Le but de l'approche interactive est de permettre à l'utilisateur d'intervenir sur les résultats de segmentation obtenus par clustering afin de pouvoir rectifier les clusters erronés. Pour effectuer le clustering initial, nous avons vu dans la section précédente que nous avons besoin de connaitre le nombre de clusters souhaité. Dans ce but, nous considérons que l'utilisateur connait approximativement le nombre de timbres de la pièce musicale et qu'il donnera ce nombre au système afin d'initialiser le clustering qui engendrera la première segmentation.

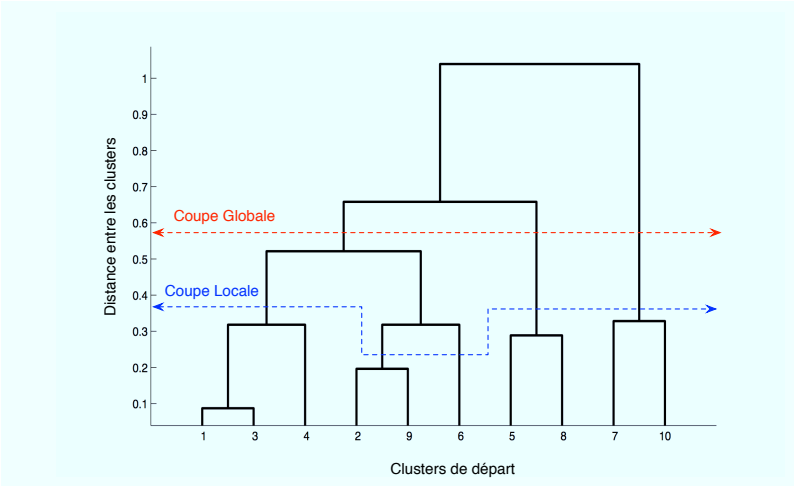

FIG. 3.9 – Comparaison des coupes globale (en rouge) et locale (en bleu)

3.3.6.1 Coupes globales et locales

Comme nous l'avons vu dans la section 3.3.5, il est possible d'obtenir différents niveaux de clustering des données en fonction du niveau ou l'on coupe le dendrogramme (figure 3.8). Ce graphe hiérarchique qui peut être assimilé à un arbre binaire enrichi de l'information de distance entre les noeuds possède la propriété suivante : un cluster associé à un noeud dans un dendrogramme peut être divisé en deux clusters qui sont les deux fils du noeud considéré. C'est sur cette propriété de base que se fondent les interactions que nous

proposons. Nous appellerons la méthode de coupe présentée dans la section 3.3.5 *coupe globale*. Nous proposons d'introduire une approche de *coupe locale* en utilisant le retour utilisateur afin d'améliorer le clustering. La *coupe locale* est illustrée dans la figure 3.9 (en bleu).

3.3.6.2 Comparaisons de deux scénarios d'interaction

Deux scénarios alternatifs ont étés comparés afin d'améliorer la segmentation engendrée par le clustering initial.

Premier scénario : "casser/fusionner"

Dans ce premier scénario, l'utilisateur peut casser ou fusionner les segments proposés par le clustering. Ainsi, l'utilisateur peut choisir le segment qu'il souhaite corriger : l'utilisateur choisit de "casser" un segment lorsqu'il considère que deux timbres différents sont contenus dans un même segment (figure 3.10) et réciproquement, l'utilisateur décide de "fusionner" deux segments de même timbre contigus temporellement lorsqu'ils ont été fragmentés (figure 3.11). Etant donné que chaque segment est associé à un cluster, le retour utilisateur est pris en compte au niveau du clustering, de la façon suivante :

- Scinder un cluster en ses deux fils lorsque l'utilisateur veut "casser" un segment.
- Réunir les clusters considérés quand un utilisateur décide de "fusionner" deux segments.

Deuxième scénario : "casser"

Le deuxième scénario est plus simple car il ne considère que la deuxième action du scénario précédant : à chaque itération, l'utilisateur signale le segment le plus erratique au système qui se charge par la suite de scinder le cluster correspondant en ses deux fils.

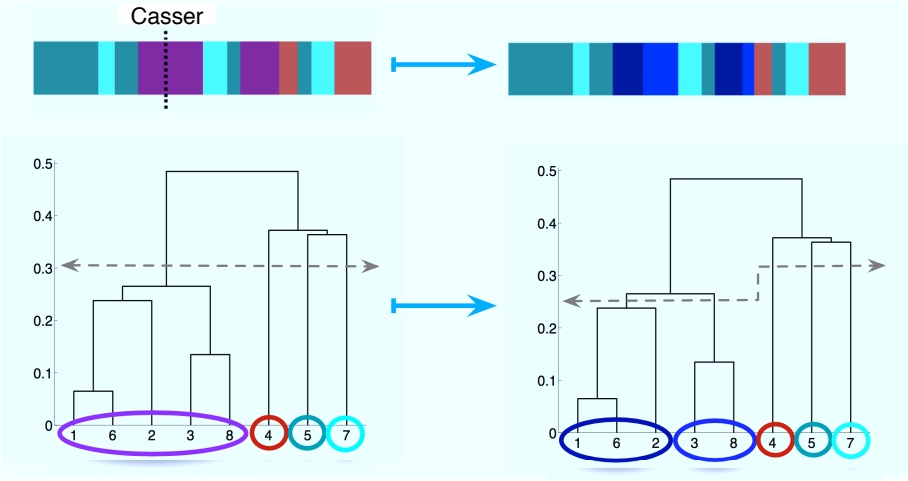

Fig. 3.10 – Casser un segment

On peut noter que chacune des deux stratégies nous permet de propager le retour utilisateur sur toute la durée du signal. Cette propriété est illustrée par les figures 3.10 et 3.11 : les parties hautes représentent les segments temporels sous lesquels on trouve leurs dendrogrammes correspondants. L'approche de "coupe locale" permet d'obtenir des

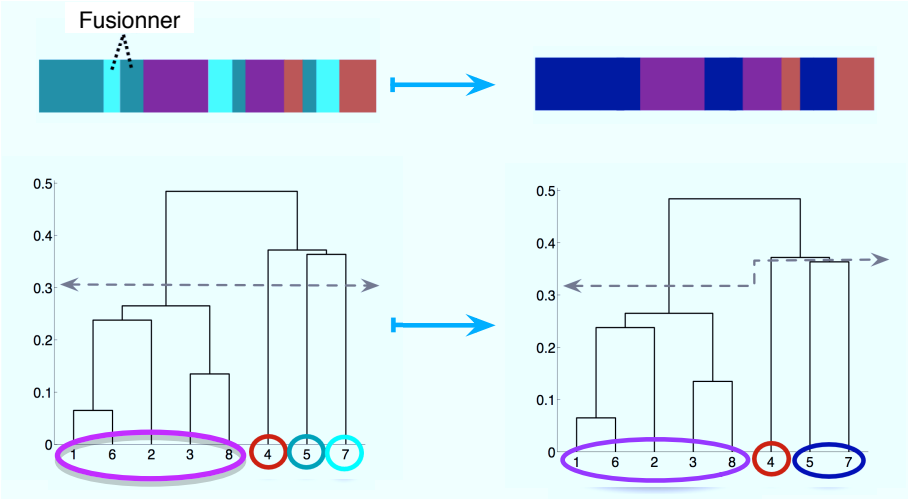

FIG. 3.11 – Fusionner deux segments

partitionnements de données s'adaptant au retour utilisateur. Nous chercherons à évaluer cette approche dans la section suivante.

3.4 Evaluation

Dans cette section, nous présentons l'évaluation de la segmentation timbrale de manière indépendante. Nous avons réalisé cette première évaluation avec le corpus monophonique (*corpus M*) décrit dans la section 2.5.3.

3.4.1 Critères d'évaluation

La comparaison entre la vérité terrain et la segmentation obtenue n'est pas directe. En effet, le nombre de clusters obtenus est la plupart du temps différent du nombre de mixtures réelles présentes dans la pièce synthétique. De plus, il n'y a pas de correspondances a priori entre les mixtures réelles et les étiquettes affectées aux clusters. Par conséquent, il est nécessaire d'associer chaque mixture à un cluster pertinent : chaque mixture M_i est associée au cluster W_j qui contient le plus grand nombre de trames appartenant à la mixture. On peut ainsi reformuler les mesures standard de *rappel* R_i et *précision* P_i de la façon suivante :

$$R_i = \frac{\max_j |M_i \cap W_j|}{|M_i|}, \; P_i = \frac{\max_j |M_i \cap W_j|}{|W_J|}. \tag{3.10}$$

avec

$$J = \arg\max_j |M_i \cap W_j|, \tag{3.11}$$

Nous utilisons ensuite la *f-mesure* pour évaluer les performances :

$$f - mesure = \frac{2RP}{R+P} \tag{3.12}$$

où R et P sont respectivement les moyennes de R_i et P_i sur toutes les classes.

3.4.2 Expériences

3.4.2.1 Simulation utilisateur

Pour évaluer l'influence de l'interaction de l'utilisateur sur les performances du clustering, nous tirons avantage du fait que la segmentation souhaitée ne laisse pas de place à des interprétations subjectives : il existe avec notre corpus synthétique une unique segmentation correcte pour chaque pièce. En connaissant cette vérité terrain, il est possible de simuler les actions d'un utilisateur. Un cluster est un ensemble de segments qui ne sont pas nécessairement contigus temporellement. Nous considérons lors de la simulation que l'utilisateur commence par corriger les segments les plus erratiques : ceux dont le nombre de trames mal étiquetées par rapport à la vérité terrain est maximal. Le clustering est par la suite mis à jour et une nouvelle segmentation est présentée à l'utilisateur. Le même processus est répété par l'utilisateur jusqu'à obtenir une segmentation satisfaisante. De plus, nous considérons que l'utilisateur a accès à la fonction logicielle classique "annuler" (fonction "undo" en anglais) qui permet de remettre le système dans l'état précédent la dernière action effectuée.

3.4.2.2 Comparaison de performances pour les deux scénarios d'interaction

Les résultats des simulations utilisateurs sont d'abord donnés sous la forme de $f - mesure$ moyennes en fonction du nombre d'itérations de l'algorithme (figure 3.12). Le nombre d'itérations correspond au nombre de retours utilisateur.

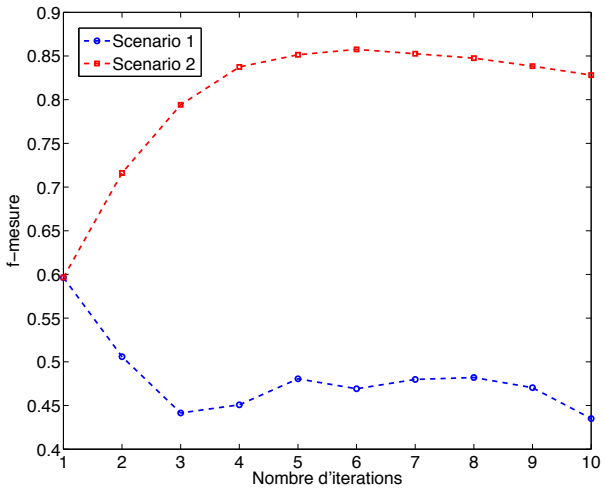

Fig. 3.12 – Comparaison de performances pour deux scénarios d'interaction

Les résultats obtenus montrent que le second scénario donne les meilleurs résultats et que l'interaction permet d'améliorer la segmentation initiale. Nous avons observé expérimentalement que la fusion de clusters ajoutait de l'instabilité au système ce qui explique la dégradation des performances par rapport au premier mode d'interaction. En effet, la fusion entre deux clusters quelconques dans le dendrogramme revient à trouver l'ancêtre commun le plus proche entre ces deux clusters. Par conséquent, la fusion peut être dangereuse : dans un cas extrême, l'ancêtre commun le plus proche des deux clusters peut être la racine du dendrogramme. Dans un tel cas, la fusion entre les deux clusters peut entrainer

la perte de l'information de partitionnement des données car on obtient un unique cluster. Pour ne pas tomber dans de tels cas, nous avons limité la fusion à des clusters dont l'ancêtre commun le plus proche est au maximum à la hiérarchie supérieure d'ordre 2.

La seconde approche, plus stable, donne de meilleurs résultats. Ces résultats sont confirmés par l'évolution des maximas de la *f-mesure* moyenne en ne tenant pas compte du nombre d'itérations. Nous avons comparé les résultats obtenus par "coupe locale" à un score de référence obtenu par "coupe globale". Partant d'un score de référence de 0.82, la première et la seconde méthode obtiennent respectivement 0.71 et 0.9 et améliorent respectivement 34,2% et 92,5% des scores de référence. A titre comparatif, une version "non interactive" du système n'obtient que 0.78.

Compte tenu des meilleurs résultats obtenus avec le second scénario d'interaction, cette approche a été appliquée sur un extrait de notre pièce musicale de référence "Timbre durée" : nous observons une amélioration de 0.04 pour la *f-mesure* par rapport au score de référence (on passe de 0.67 à 0.71). La figure 3.13 illustre le résultat obtenu avec notre algorithme de segmentation pour cette même pièce. Dans cette représentation, un spectrogramme de la pièce est affiché avec une palette de couleurs différente pour une mixture donnée. Ainsi, cette pièce de musique est représentée comme étant un enchainement de mixtures tel que nous l'avons décrit dans l'introduction du chapitre.

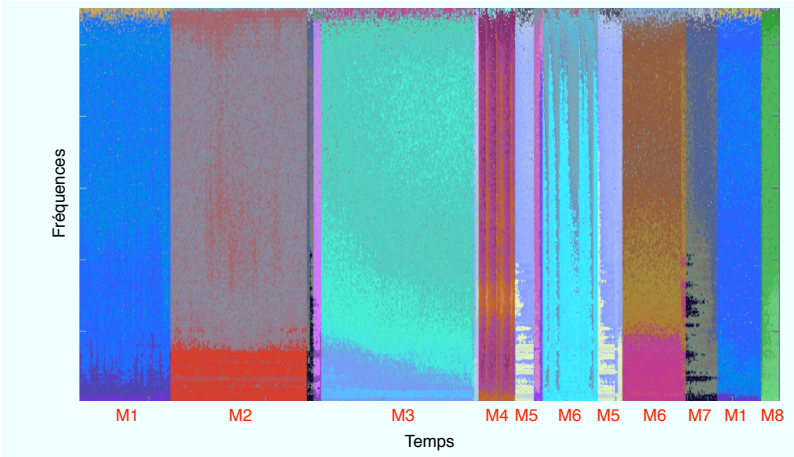

FIG. 3.13 – Segmentation d'une pièce électroacoustique : "Timbre durée"

3.5 Conclusion

Dans ce chapitre, nous avons proposé un système interactif de segmentation adapté aux musiques composées d'objets sonores exploitant le retour utilisateur. Cette méthode permet d'obtenir à la fois des frontières entre les mixtures et également un étiquetage de chaque segment. Deux scénarios d'interaction différents ont étés comparés en générant un corpus synthétique dont la construction est basée sur une pièce de musique concrète d'*Olivier Messiaen*. Les expériences ont été réalisées en simulant l'utilisateur et ses interactions avec le système. Les résultats ont montré qu'une méthode simple qui propose à l'utilisateur de "couper" les segments erronés permet d'améliorer les performances de clustering par rapport à une approche statique.

La segmentation en mixtures ainsi obtenue permet d'initialiser le système d'analyse en donnant la possibilité à l'utilisateur de choisir les segments contenant les classes sonores

qu'il souhaite annoter dans le document : nous les nommerons "segments représentatifs de classe" dans la suite de ce document.

Chapitre 4

Classification interactive d'objets sonores

Sommaire

4.1 Introduction

Ce chapitre se concentre sur la deuxième phase principale du système : la *classification d'objets sonores* qui permet de catégoriser les segments de mixtures obtenus par le procédé de clustering décrit dans le chapitre précédent. Pour obtenir une classification adaptée aux souhaits de l'utilisateur, nous utilisons une approche interactive basée sur le *retour de pertinence* (relevance feedback) et l'*apprentissage actif* (active learning). En effet, les besoins de l'utilisateur sont dépendants du point de vue d'analyse (chapitre 2) et par conséquent, la classification doit pouvoir s'adapter aux différents points de vue. Dans un premier temps, il est nécessaire de définir certains concepts qui nous seront utiles dans la suite du document. On peut distinguer plusieurs types de problèmes de classification (illustrés dans la figure 4.1) :

- **Problème *bi-classes*** : ce type de problème est le plus simple. Dans ce cas, nous avons uniquement deux classes possibles. Par conséquent, l'étiquetage d'un échantillon dans un problème *bi-classe* est binaire.
- **Problème *multiclasses*** : ce type de problème concerne le cas où le nombre de classes possible Q est supérieur à deux. L'étiquette d'un tel échantillon peut donc être représentée par un entier e tel que $1 \leq e \leq Q$ avec $Q > 2$.
- **Problème *multilabel*** : dans ce cas, un échantillon peut appartenir à plusieurs classes en même temps. On parle souvent de *classification non exclusive*. Pour un problème *multilabel* à Q classes, on peut représenter l'étiquette associée à un échantillon x par un vecteur $(v_1, v_2, ..., v_Q)$ avec $v_q = 1$ si $x \in C_q$ et $v_q = -1$ si $x \notin C_q$ $(1 \leq q \leq Q)$.

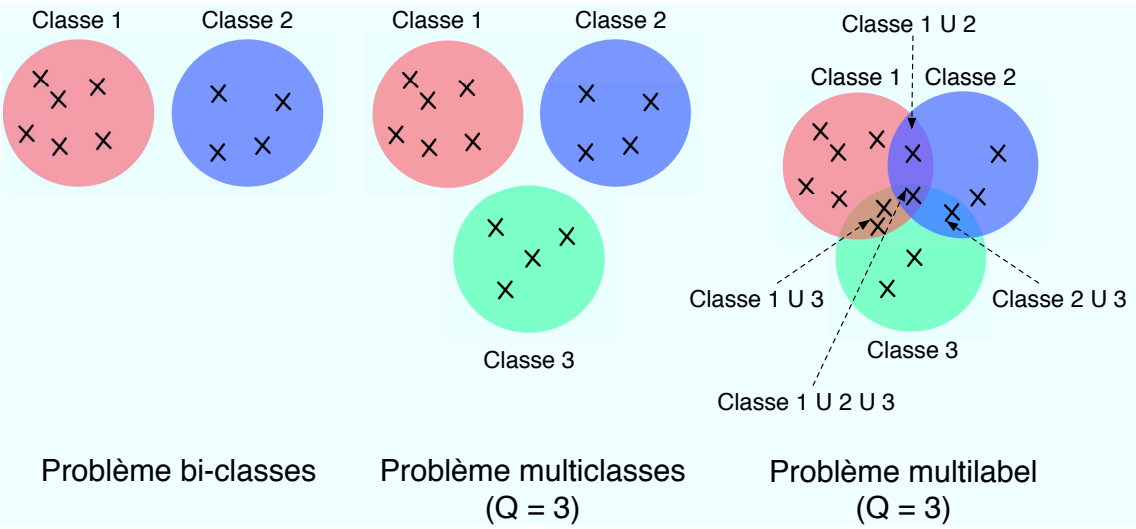

Fig. 4.1 – Les différentes types de problèmes

Dans la phase de classification, l'objectif est d'obtenir un étiquetage multiple des différents segments de mixtures constituant la pièce. Autrement dit, le système doit pouvoir prédire pour chaque segment si il contient les classes sonores exprimées par l'utilisateur ou non. Il s'agit donc d'un problème *multilabel*. Le problème de classification est illustré par la figure 4.2. Dans cet exemple, l'utilisateur s'intéresse à deux classes sonores en particulier ($S1$ et $S3$) et des segments ont été étiquetés manuellement. Le but est d'obtenir l'éti-

quetage des autres segments en apprenant des classifieurs à partir des segments étiquetés manuellement. Il sera ainsi possible d'obtenir l'information de présence d'une classe sonore donnée sur l'ensemble du signal (flèches en pointillé sur la figure 4.2). Il s'agit donc d'un problème de classification supervisée avec un nombre d'échantillons restreint.

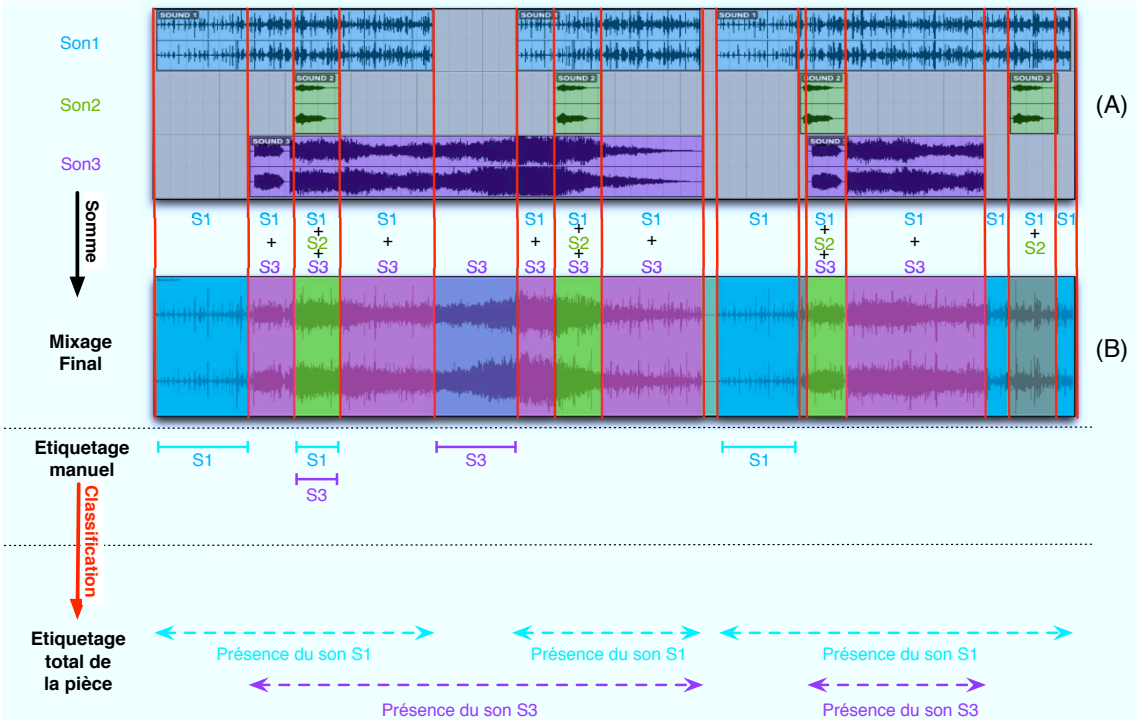

FIG. 4.2 – Classification de deux sons à partir de segments étiquetés manuellement

Le système d'analyse proposé a pour fonction d'assister l'utilisateur dans la tâche d'annotation des objets sonores. Il convient donc de faire en sorte que l'annotation manuelle soit minimale. Cependant, comment minimiser le nombre de segments étiquetés manuellement ? Comment choisir les segments à étiqueter manuellement afin d'obtenir des prédictions efficaces ?

Dans un premier temps, nous présenterons un état de l'art de la classification relatif à notre problème, puis nous verrons comment exploiter les informations obtenues pendant l'initialisation (phase de clustering) et les descripteurs utilisés seront présentés. Nous terminerons le chapitre en présentant les différentes approches d'apprentissage interactif proposées avant d'évaluer le système complet dans la dernière section.

4.2 Etat de l'art

A notre connaissance, il n'existe pas de travaux antérieurs centrés sur la même problématique applicative et scientifique que le système proposé, par contre nous situons ce travail au croisement de divers domaines. Le but de cet état de l'art est principalement d'expliquer les points en commun et les différences avec ces domaines de recherche ainsi que de présenter les concepts importants qui ont inspiré la classification d'objets sonores dans le contexte de pièces électroacoustiques.

4.2.1 Classification d'instruments dans la musique polyphonique

La classification automatique d'instruments dans la musique polyphonique reste un challenge difficile à relever et est de fait moins étudiée que la classification d'instruments pour de la musique monophonique. En effet, dans ce contexte, il s'agit de reconnaitre des instruments de musique dans des mixtures instrumentales. Par conséquent, le problème de la description est plus complexe puisqu'il fait intervenir des superpositions sonores et tous les phénomènes que cela peut engendrer (notamment les chevauchements entre partiels). Certaines méthodes s'attachent à séparer les notes des différents instruments afin d'appliquer des méthodes classiques, d'autres se focalisent sur l'extraction de descripteurs adaptés.

Godsmark & Brown (1999) proposent d'exploiter une trajectoire de timbre dans laquelle le centroïde spectral en fonction de l'amplitude est utilisé pour séparer la musique polyphonique en ses lignes mélodiques constituantes. Le modèle proposé permet d'obtenir un taux de reconnaissance de 80% pour des mélanges piano/contrebasse mais chute de 40% pour des mixtures plus complexes à 4 instruments.

Kinoshita et al. (1999) proposent une extension à un système existant qui n'était pas robuste pour les signaux de mixtures présentant des chevauchements de partiels. Le système est testé avec des mélanges de deux notes créant des chevauchements. La méthode fonctionne par rapprochement avec des templates connus qui seront pondérés en évaluant l'importance des descripteurs.

Eggink & Brown (2003) proposent un système de reconnaissance d'instruments par *Modèle de Mélange Gaussien* (GMM) et utilisent le concept de l'"attribut manquant" (missing feature theory) quand il y a plus d'un son à la fois. Ainsi, les régions fréquentielles qui contiennent des interférences sont exclues du processus de classification car considérées comme non fiables. La méthode proposée est évaluée sur des combinaisons de deux instruments avec des accords de deux notes ainsi que sur des mélanges de phrases instrumentales.

Essid et al. (2006) proposent une nouvelle approche de reconnaissance des instruments basée sur l'apprentissage de taxonomies instrumentales. Cette approche n'utilise pas de sources instrumentales séparées pour l'apprentissage mais exploite des annotations de musiques commerciales. Ainsi, les différents types de mixtures instrumentales sont appris sur des morceaux de jazz (du duo au quartet) et l'algorithme cherche à retrouver directement ces mixtures dans les signaux de test. Cette méthode prend pour hypothèse l'invariabilité de l'instrumentation de certains styles musicaux et apprend à retrouver les mélanges dans des morceaux basés sur les mêmes instruments.

Kitahara et al. (2007) présentent une solution pour les problèmes de variation des descripteurs dus aux superpositions des sons instrumentaux. Pour résoudre ce problème, les auteurs utilisent une pondération des descripteurs basée sur le degré de perturbation introduit par la superposition. Dans cet article, l'influence de la superposition sur chaque descripteur est assimilée au rapport des variances intra-classe et inter-classe. La pondération est réalisée par une *analyse linéaire discriminante* qui permet de minimiser ce dernier rapport.

Little & Pardo (2008) s'intéressent à l'identification d'instruments dans des mixtures en réalisant un apprentissage à partir de segments partiellement étiquetés. Dans ce travail, les échantillons d'apprentissage sont les mixtures qui contiennent l'instrument appris dans une proportion significative. Ainsi, le système apprend à partir de mixtures qui contiennent à la fois l'instrument appris et également d'autres instruments. Une étude comparative est réalisée et montre que l'apprentissage sur des échantillons partiellement étiquetés permet

d'obtenir de meilleurs résultats qu'une approche classique où les modèles sont appris à partir de sources instrumentales isolées.

Dans de nombreux travaux présentés, l'apprentissage des modèles se base sur la connaissance des instruments qui seront utilisés. La plupart du temps, les instruments appris sont "standards" et appartiennent aux grandes familles d'instruments rencontrées dans la musique occidentale. Dans notre cas, nous n'avons pas de connaissance a priori sur les sons qui composent les pièces électroacoustiques car les compositeurs travaillent directement sur le matériau sonore et peuvent utiliser n'importe quelle source sonore acoustique ou électronique. De plus les sources utilisées sont souvent hétérogènes et polyphoniques à l'origine (par exemple : des chants d'oiseaux). Dans les travaux présentés, on peut s'intéresser particulièrement à l'approche de Little & Pardo (2008) qui utilise des échantillons de mixtures partiellement étiquetés pour l'apprentissage. En effet, dans notre cas, comme nous ne disposons pas de sources séparées a priori, nous devrons forcément réaliser l'apprentissage à partir de mixtures sonores. Dans la section suivante, nous présentons un état de l'art du *retour de pertinence* et de l'*apprentissage actif* dans le domaine audio car comme nous l'avons dit dans la section 2.5.2.2, étant donné la nécessité de proposer un système adaptatif (pour les raisons musicologiques évoquées dans la section 2.4.2) et la difficulté du problème (un segment peut avoir plusieurs étiquettes car les pièces sont polyphoniques), le retour utilisateur est une source d'information qui peut fortement aider la classification.

4.2.2 Retour de Pertinence et Apprentissage actif

L'utilisation du *retour de pertinence* a d'abord été introduit dans le domaine de la recherche textuelle (Rijsbergen (1979), Salton (1968)). Ainsi, pendant la recherche de documents, l'utilisateur peut interagir avec le système et sélectionner les documents qui lui semblent pertinents. Les systèmes de classification "orientés audio" exploitant le retour utilisateur sont peu nombreux en comparaison des systèmes purement automatiques.

Hoashi et al. (2003) proposent de retrouver des morceaux de musique selon les préférences propres à un utilisateur qui sont supposées être ambigües en utilisant le *retour de pertinence*. L'approche utilise des arbres de vecteurs quantifiés (TreeQ) pour réaliser la recherche. Pour évaluer la méthode, une base de données a été construite à partir d'une collection de CDs du commerce. Les expériences montrent l'efficacité du *retour de pertinence* pour la recherche dans la base ainsi que pour la constitution de profils utilisateurs personnalisés.

Dans Mandel et al. (2006), un système de recherche par similarité musicale exploitant des *machines à vecteurs supports* (SVM) combiné à l'*apprentissage actif* est présenté. Pour tester le système, 1210 morceaux de musique pop ont étés classés par émotions, styles et artistes. Ainsi, un classifieur est appris pour chaque requête à partir de différentes représentations de descripteurs bas-niveaux d'un ou plusieurs morceaux fournis par l'utilisateur. Le système fonctionne itérativement : à chaque itération, il prédit les étiquettes des morceaux non étiquetés à partir du classifieur courant puis il utilise l'apprentissage actif pour demander à l'utilisateur d'annoter de nouveaux morceaux afin de faire progresser l'apprentissage. Le but du système est d'obtenir une prédiction correcte des étiquettes en un minimum d'itérations. Ce travail vérifie que dans ce contexte, le recours à l'apprentissage actif permet de diminuer de moitié le nombre de morceaux annotés manuellement. De plus, parmi les différentes représentations de descripteurs bas-niveau comparées, les résultats montrent

qu'une représentation simple (moyenne et matrice de covariance des MFCC d'un morceau) permet d'obtenir de meilleurs résultats que des représentations plus complexes (GMM etc.).

Chen et al. (2008) présentent un système de recherche de contenu musical qui intègre le retour utilisateur. Un algorithme d'apprentissage basé sur une fonction de base radiale est utilisé pour la classification et un algorithme de pondération des descripteurs qui utilise à la fois les exemples positifs et négatifs est présenté. Le système est testé pour la classification en genres et en émotions et obtient des résultats comparables à ceux présentés dans la littérature.

Shan et al. (2008) proposent une approche pour la recherche de catégories musicales spécifiques qui partagent un même concept sémantique. Les catégories étant subjectives, ils utilisent le *retour de pertinence* pour apprendre les concepts sémantiques sur de la musique polyphonique représentée symboliquement. Un modèle de segment et une représentation qui intègrent des descripteurs globaux et locaux sont utilisés. La recherche est effectuée via un algorithme de reconnaissance de formes et un algorithme associatif de classification modifié. Trois stratégies sont utilisées pour sélectionner les objets les plus utiles pour l'apprentissage du concept (le plus positif, le plus informatif et une stratégie hybride).

On peut retenir que ces travaux utilisent le *retour de pertinence* et l'*apprentissage actif* afin d'exprimer la subjectivité et l'adaptabilité. Dans ces travaux, la recherche d'un objet particulier est définie par des exemples d'objets considérés comme appartenant à une même catégorie par un utilisateur donné. Ces méthodes sont donc des outils puissants et utiles dans notre contexte car ils peuvent permettre à un utilisateur de définir ses propres *objets sonores* en donnant des exemples choisis. Nous pouvons également mentionner que la phase de classification des *segments de mixtures* se rapproche de Mandel et al. (2006) qui cherche à classifier des représentations de morceaux complets en utilisant un algorithme SVM. Une différence importante avec notre travail est que les *segments de mixtures* sont de l'ordre de quelques secondes (c'est peu par rapport aux 3 minutes d'une chanson standard). Nous avons donc moins d'échantillons à notre disposition pour l'apprentissage des classes ce qui justifie également l'apprentissage actif qui vise à améliorer les performances de classification lorsqu'on a peu d'échantillons étiquetés à notre disposition (section 4.5.4).

4.2.3 Classification multilabel

Le problème de la classification multilabel est également un aspect important de notre problématique de travail. Tsoumakas & Katakis (2007) proposent une vue d'ensemble des différents travaux sur le sujet et distinguent deux types de méthodes : les approches par *transformation du problème* et celles par *adaptation de l'algorithme*. Dans les approches par *transformation du problème*, le problème de classification multilabel est remplacé par un problème ou plusieurs problèmes de classification à simple étiquette. Les méthodes par adaptation d'algorithmes sont celles qui étendent directement un algorithme d'apprentissage spécifique au problème multilabel. Dans le domaine musical, la classification multilabel a été appliquée principalement dans les problèmes de classification en genres ou en émotions.

Wieczorkowska et al. (2006) présentent une approche de classification mutilabel en émotions. La base utilisée comporte 875 extraits de musique de 30 secondes (chansons et pièces de musique classique). On dénombre 13 classes d'émotions différentes. Pour la phase de classification, un algorithme modifié des *k plus proches voisins* est utilisé (l'algorithme

original est décrit dans Duda et al. (2001)). Cet algorithme modifié permet de prendre en compte les étiquettes multiples pouvant être affectées à un échantillon. Pour prédire le multilabel d'un échantillon, l'algorithme calcule un histogramme des étiquettes de son voisinage. Les étiquettes dont le nombre d'occurrences dépasse un certain seuil (déterminé expérimentalement) seront affectées à l'échantillon considéré.

Trohidis et al. (2008) proposent une évaluation de 4 algorithmes de classification multilabel de la musique en émotions, 6 classes d'émotions différentes sont considérées. La base utilisée comporte 593 morceaux et les approches comparées sont les suivantes : pertinence des résultats de classifieurs binaires ou "Binary Relevance" (BR), étiquettes construites avec les parties de l'ensemble initial ou "Label Powerset" (LP), k sous-ensembles aléatoires "RAndom K-labELsets" (RAKEL), k plus proches voisins multilabel ou "MultiLabel k-Nearest Neighbor" (MLkNN). Les trois premières méthodes sont des approches par *transformation du problème* et la dernière est une approache par *adaptation de l'algorithme*. BR considère la prédiction de chaque étiquette comme un problème de classification binaire indépendant. Soit L l'ensemble des étiquettes possibles d'un échantillon, LP considère le problème de prédiction multiclasse des étiquettes définies par les parties de l'ensemble L. RAKEL est une méthode récente qui améliore l'algorithme LP (Tsoumakas & Vlahavas (2007)). MLkNN est une approche performante de type *adaptation de l'algorithme* qui adapte l'algorithme des k plus proches voisins à la problématique multilabel. Les trois premières approches (BR,LP et RAKEL) ont été réalisées à l'aide d'un classifieur SVM. Il ressort de cette étude que la méthode RAKEL est la plus performante au détriment d'un temps de calcul plus long. En effet, RAKEL nécessite de réaliser des validations croisées afin de sélectionner plusieurs paramètres avant la phase d'entrainement. De plus, RAKEL est une méthode de type *ensemble* qui utilise plusieurs modèles d'ou un temps d'entrainement plus long.

Dans Lukashevich et al. (2009), une nouvelle approche pour la classification multilabel des genres musicaux est présentée. Trois expériences différentes sont réalisées sur une base de 430 morceaux de musiques du monde. On distingue 16 sous-genres ou "influences régionales" répertoriées et chaque morceau peut être affecté à une ou plusieurs étiquettes parmi les 16. Dans la première expérience, on considère qu'un multilabel unique est affecté à chaque morceau. Dans la deuxième expérience, chaque morceau est segmenté et chaque segment est affecté à un multilabel. Enfin, dans la troisième expérience, chaque segment de morceau est considéré selon trois aspects (le timbre, le rythme, la mélodie/harmonie) et étiqueté selon une étiquette unique. Un classifieur basé sur le *modèle de mélange de gaussiennes* (ou GMM pour *Gaussian Mixture Model*) est utilisé. Pour la classification multilabel, une approche de type BR est utilisée. Ainsi, chaque classifieur binaire H_C est appris pour prendre une décision binaire : l'échantillon appartient-il à la classe C ou non ? Les auteurs précisent que les résultats obtenus pourraient être améliorés en utilisant un classifieur de type SVM au lieu des GMMs.

Ces travaux nous renseignent notamment sur les différentes approches de classifications utilisées pour résoudre un problème multilabel. On peut retenir en particulier que les approches de type LP semblent être plus performantes car elles prennent en compte les intersections des classes mais elles demandent également plus de temps de calcul car il faut considérer un nombre de classifieurs plus important que le nombre d'étiquettes possibles. La méthode RAKEL obtient de bons résultats mais souffre d'une complexité importante qui augmenterait fortement le temps d'attente utilisateur. Nous pouvons également remarquer que les travaux récents utilisent des classifieurs SVM ou, comme Lukashevich et al. (2009),

souhaiteraient améliorer leurs performances en utilisant ce type de classifieur.

4.2.4 Classification d'images

Le *retour de pertinence* est très utilisé dans le domaine de la classification d'images/photos. La raison de ce constat est simple : les photos, tout comme les sons, sont des objets qui renvoient à des jugements très subjectifs. En effet, un utilisateur peut choisir de classifier de tels objets selon plusieurs axes : description, concept, émotion suggérée, vocabulaire esthétique etc. De plus, en classification d'images, il est courant de vouloir associer une image à plusieurs étiquettes : par exemple une photo d'une personne sur une plage peut être affectée à la fois à la classe "plage" et à la classe "personne". Certains travaux en classification d'images constituent une inspiration importante pour ce travail de classification d'objets sonores car les formalismes possibles pour ces deux types d'objets sont relativement similaires.

Crucianu et al. (2004) et Zhou & Huang (2003) décrivent une vue d'ensemble de la littérature sur le *retour de pertinence* et l'*apprentissage actif* appliqués à la recherche d'images. Certains travaux intègrent à la fois le *retour de pertinence* par *apprentissage actif* et la classification multilabel (Li et al. (2004), Goeau et al. (2008), Goeau (2009), Singh et al. (2009), Qi et al. (2009)).

Li et al. (2004) proposent une méthode multilabel basée sur des SVMs et exploitant l'*apprentissage actif* pour la classification d'images. Dans cette publication, une approche de type BR est utilisée et deux *stratégies d'échantillonnage* originales sont présentées et comparées à une sélection d'échantillons aléatoire.

Goeau et al. (2008) et Goeau (2009) présentent un système de classification d'images basé sur une version évidentielle de l'algorithme des *k plus proches voisins* qui utilise également l'*apprentissage actif*. Ce système permet à l'utilisateur d'initialiser, supprimer ou fusionner des classes et éventuellement de corriger les propositions d'étiquettes du système. L'approche choisie permet de prendre en compte l'imprécision, l'incertitude et les conflits entre les descripteurs visuels. Ainsi, des *stratégies d'échantillonnage* prenant en compte la positivité, l'ambiguité et la diversité sont présentées. Dans cette approche, les sorties des classifieurs sont exprimées sous forme de probabilités pignistiques (Smets (2005)) qui permettent de sélectionner les échantillons en fonction de la stratégie. L'évaluation du système est réalisée par simulation utilisateur et permet de comparer les résultats obtenus avec les différentes *stratégies d'échantillonnage*. Le classifieur proposé permet également de gérer la classification multilabel.

Singh et al. (2009) proposent une approche de classification multilabel exploitant l'apprentissage actif qui permet de réduire le nombre d'images présentées à l'utilisateur. Cet article propose d'utiliser un classifieur SVM dans une approche BR et compare trois *stratégies d'échantillonnage* afin de réduire le nombre d'images que l'utilisateur doit annoter manuellement. La première stratégie est aléatoire, elle sert de point de référence. Pour toutes les stratégies, l'image nouvellement annotée est ajoutée à l'ensemble d'apprentissage (elle est donc retirée des images non étiquetées) et le processus est répété itérativement. La deuxième stratégie est dite "annotation monolabel" : il s'agit d'une stratégie souvent utilisée en *apprentissage actif* qui consiste à chercher l'échantillon le plus informatif. Dans le cas des SVM, dans un problème bi-classe, l'échantillon le plus informatif est l'échantillon le plus ambigu (celui le plus proche de l'hyperplan séparateur). Dans la dernière stratégie, dite "annotation multilabel", la distance à la marge est probabilisée et effectué pour

chaque étiquette possible et pour chaque échantillon. Une moyenne est ensuite calculée pour chaque échantillon et celui qui maximise cette moyenne est sélectionné. Les expériences réalisées démontrent que la stratégie “annotation monolabel” est plus performante que la stratégie “annotation multilabel”.

Qi et al. (2009) proposent une nouvelle approche de sélection d'échantillons pour des problèmes multilabels. Cette approche est dite à “2 dimensions” (ou 2DAL pour “2 Dimensional Active Learning”), elle sélectionne des paires d'étiquettes à annoter manuellement pour un échantillon sélectionné. En considérant à la fois la redondance des échantillons et des étiquettes, l'annotation manuelle est minimisée. On peut noter que cette approche, si elle semble performante, introduit des calculs supplémentaires lors de la sélection afin de calculer les redondances.

De nombreux travaux comme Hong et al. (2000), Tong & Chang (2001), Joshi et al. (2009), Singh et al. (2009) utilisent des classifieurs SVM couplés à des *stratégies d'échantillonnage* pour sélectionner des exemples utiles. En effet, les SVM apparaissent comme des classifieurs de choix car en *apprentissage actif*, nous avons besoin de mesurer l'appartenance relative d'un échantillon à une classe. Or, cette information peut être obtenue naturellement à partir de la distance d'un échantillon à l'hyperplan séparateur dans le cas des SVM. On peut remarquer que les travaux exploitant l'apprentissage actif dans des problèmes multilabels sont encore assez rares.

4.3 Exploitation des informations d'initialisation

Pendant la phase d'initialisation, le système réalise une segmentation timbrale afin d'obtenir des unités sonores homogènes (chapitre 3). Cette tâche est effectuée par un algorithme de clustering hiérarchique qui permet, en plus de l'information de segmentation, d'obtenir un étiquetage des segments : les segments proches timbralement ont la même étiquette. Cette dernière information est importante car elle permet à l'utilisateur de comparer facilement les segments proches et d'effectuer un choix entre eux. Comme nous l'avons vu dans la figure 4.2, nous représentons les différents segments par des couleurs identiques lorsqu'ils appartiennent au même cluster afin de repérer facilement les segments de même timbre.

Pour initialiser l'apprentissage des différentes classes sonores, l'utilisateur doit choisir un segment de démarrage pour chaque classe visée. Il est souhaitable d'initialiser l'apprentissage avec des segments représentatifs de chaque classe afin de ne pas obtenir des résultats contradictoires pour les premières itérations du système. En effet, comme nous l'avons expliqué dans le chapitre précédent, les segments obtenus après segmentation sont pour la plupart des mixtures composées de plusieurs sons superposés car la plupart des pièces électroacoustiques sont polyphoniques. Si nous reprenons notre exemple et la segmentation correspondante (figure 4.3), il est facile d'observer ce phénomène de superposition propre à la musique polyphonique. La figure 4.3 oppose les segments représentatifs d'une classe sonore aux segments ambigus. On considère que les mixtures $M1$ et $M4$ sont représentatives respectivement des classes de sons $S1$ et $S3$. En effet, ces deux segments contiennent des sons “isolés” et par conséquent ils ne sont pas sujets aux phénomènes de masquage sonore que peut engendrer la superposition de sons (Fastl & Zwicker (2007)). A l'opposé, les segments $M3$ et $M5$ sont ambigus car ils sont constitués du mixage de plusieurs sons.

Du point de vue de l'utilisation du système, les remarques qui précédent conduisent

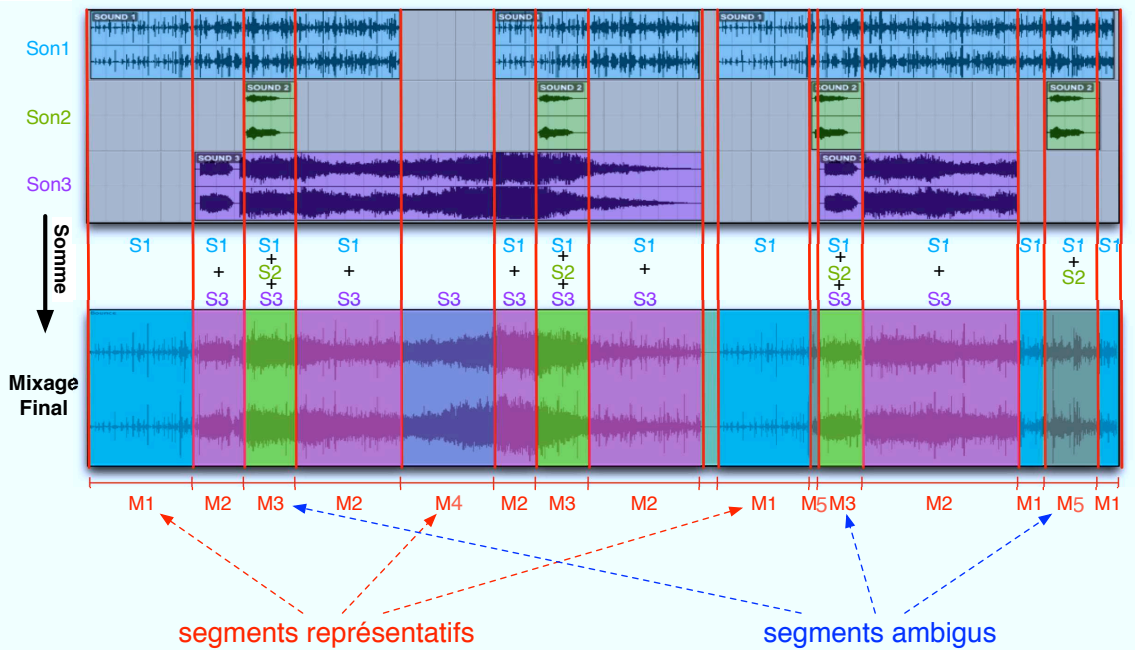

Fig. 4.3 – Segments caractéristiques et ambigus

à spécifier à l'utilisateur qu'il doit choisir un segment représentatif (dans la mesure du possible) pour initialiser une classe sonore. Cependant, on peut observer dans certains cas qu'il n'existe pas de segments représentatifs pour certaines classes (c'est le cas pour la classe $S2$). Dans ce cas, il est nécessaire que l'utilisateur choisisse un segment dans lequel la classe sonore visée est distinguable d'un point de vue sonore.

4.4 Descripteurs utilisés

Les descripteurs employés dans cette phase du système sont nombreux et couvrent divers aspects du son. Au total, on dénombre 217 attributs pour 26 descripteurs extraits en utilisant le logiciel d'extraction *YAAFE*(Mathieu et al. (2010)). Certains descripteurs sont similaires à ceux utilisés dans le chapitre 3 alors que d'autres sont propres à la phase de classification. Le tableau 4.4 présente l'ensemble des descripteurs extraits pendant la phase de classification des objets sonores.

Une présentation détaillée des descripteurs est proposée en annexe B). La phase de classification des objets sonores utilise également les descripteurs spectraux et temporels suivants qui n'ont pas été présentés dans le chapitre précédent :

Descripteurs Spectraux :

- Les *OBSIR* : mesure de la différence entre les valeurs OBSI de bandes consécutives (Essid (2005)).
- Les coefficients *LSF (Line Spectral Frequency)* : utilisés pour représenter les coefficients de prédiction linéaires Bäckström & Magi (2006), Schussler (1976).

Descripteur	Dimension	Type	Annexe
Centroïde spectral	1	Spectral	B.1
Largeur spectrale	1	Spectral	B.1
Asymétrie spectrale	1	Spectral	B.1
Platitude spectrale	1	Spectral	B.1
Platitude d'amplitude spectrale par bandes	23	Spectral	B.1
Facteur de crête spectrale par bandes	23	Spectral	B.1
Platitude spectrale globale	1	Spectral	B.1
Pente spectrale	1	Spectral	B.1
Décroissance spectrale	1	Spectral	B.1
Flux spectral	1	Spectral	B.1
Fréquence de coupure	1	Spectral	B.1
Coefficients LSF	10	Spectral	B.1
OBSI	10	Spectral	B.1
OBSIR	9	Spectral	B.1
Variation spectrale	1	Spectral	B.1
Coefficients LPC	2	Spectral	B.3
Coefficients MFCC (+dérivées d'ordres 1 et 2)	39	Cepstral	B.2
Taux de passage par zero	1	Temporel	B.3
Moments statistiques temporels	4	Temporel	B.3
Coefficients d'autocorrélation	49	Temporel	B.3
Energie	1	Temporel	B.3
Enveloppe d'amplitude	6	Temporel	B.3
Moments de l'enveloppe	4	Temporel	B.3
Loudness spécifique	24	Perceptif	B.4
Acuité perceptive	1	Perceptif	B.4
Etalement perceptif	1	Perceptif	B.4

FIG. 4.4 – Ensemble des descripteurs extraits pendant la phase de classification des objets sonores.

Descripteurs Temporels :

- L'*énergie* : calculée à partir de la moyenne quadratique des trames du signal.
- L'*enveloppe d'amplitude* : obtenue par transformée de Hilbert et filtrage passe-bas.
- Les *moments de l'enveloppe temporelle* : calculés à partir de l'enveloppe temporelle de la même manière que les moments spectraux (voir 3.3.2).

4.5 Apprentissage interactif

Nous avons rappelé en Annexe C les techniques d'apprentissage que nous exploitons dans notre système. Cette section expose les grandes étapes de l'apprentissage interactif, les méthodes propres à chaque approche d'interaction seront présentées et comparées dans la section suivante.

4.5.1 Architecture de la boucle d'interaction

La boucle d'interaction avec l'utilisateur démarre après la phase d'initialisation pendant laquelle une segmentation en unités homogènes est effectuée. Ensuite, l'utilisateur choisit un segment de démarrage pour chaque classe sonore visée afin d'amorcer l'apprentissage (voir section 4.3). La figure 4.5 présente les grandes parties de l'architecture de la boucle d'interaction.

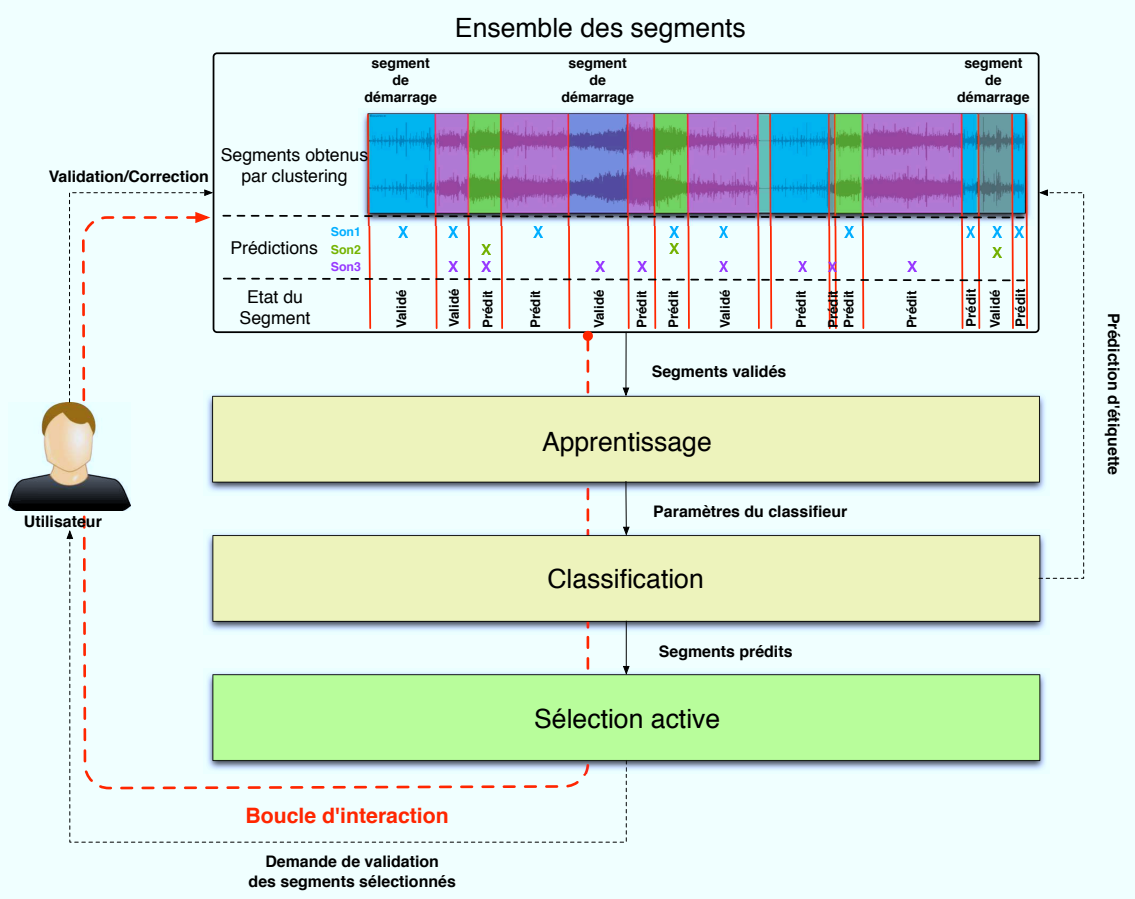

Fig. 4.5 – Architecture de la phase de classification des objets sonores

Les segments qui rentrent en compte dans l'apprentissage sont uniquement ceux validés par l'utilisateur. Par conséquent, pour la première itération, seuls les segments de démarrage sont pris en compte. La phase d'apprentissage regroupe deux tâches distinctes : la sélection d'attributs et l'apprentissage des modèles pour la classification. Ensuite, la phase de classification est effectuée : elle réalise une classification bas-niveau (pour chaque trame) suivie d'une intégration temporelle permettant la prise de décision au niveau du segment. Une fois la prédiction effectuée, les étiquettes de segments sont mises à jour dans l'interface. La dernière phase de la boucle d'interaction est la sélection active de segments. Dans cette dernière, les segments non validés par l'utilisateur sont considérés pour une sélection active (voir section 4.5.4). Le segment sélectionné est par la suite présenté à l'utilisateur pour la validation ou la correction. L'algorithme peut ensuite démarrer un nouveau cycle en prenant en compte le segment validé/corrigé par l'utilisateur.

4.5.2 Sélection dynamique d'attributs

Dans le cas de la classification d'objets sonores dans des pièces électroacoustiques, nous n'avons pas de connaissances a priori sur la nature des sources sonores à classifier. Ce cas est complexe car nous ne savons pas ce que l'utilisateur cherche à décrire. Pour répondre à ces difficultés, nous proposons de réaliser la sélection d'attributs à l'intérieur de la boucle d'interaction afin d'adapter la sélection au retour utilisateur. Ainsi, les attributs choisis par l'algorithme de sélection peuvent varier au cours des itérations successives. La stratégie d'apprentissage utilisée dans le système proposé est du type *un contre tous* (ou OVA pour "One Versus All"), nous décrivons cette stratégie dans l'Annexe C.3. Tout comme pour les classifieurs, des sélections indépendantes sont effectuées pour les Q classes sonores afin de construire un espace de description propre à chacune. La sélection d'attributs est réalisée à l'aide de l'algorithme de Fisher comme précédemment (section 3.3.3.1). Ainsi, les D meilleurs attributs qui maximisent le critère de Fisher sont gardés. Comme nous l'avons précisé dans la sélection d'attributs pour l'initialisation, il est nécessaire que la classification soit efficace afin de ne pas faire attendre l'utilisateur entre chaque itération. Après un test préparatoire où nous faisons varier le nombre d'attributs choisis dans un intervalle acceptable d'un point de vue de la complexité, nous observons que les meilleurs résultats sont obtenus pour $D = 10$ attributs.

Nous présenterons dans la section 4.7.2.3 l'évolution de la sélection des descripteurs pour différents niveaux de polyphonie et deux approches d'interaction différentes.

4.5.3 Prédiction au niveau des segments de mixtures

Dans cette section, nous cherchons à prendre une décision de plus haut niveau dans la hiérarchie temporelle : nous exposons comment décider de l'appartenance d'un segment à des classes sonores.

Dans notre problème d'apprentissage, nous avons Q classes $\{C_q\}_{1\leq q\leq Q}$. Après avoir réalisé la sélection d'attributs, nous disposons d'un vecteur de description à D dimensions $X_k = (x_{k_1}, x_{k_2}, ..., x_{k_D})$ pour chaque trame k. La classification bas-niveau est réalisée par un classifieur SVM (voir description en Annexe C.2) qui nous permet d'estimer une probabilité a postériori $p(C_i|X_k)$ de la classe C_i en sortie de chacun des Q classifieurs. Nous utilisons les SVM avec une stratégie d'apprentissage OVA, donc $p(C_i|X_k)$ estime la probabilité de l'hypothèse d'appartenance à la classe C_i. Or, l'unique hypothèse alternative dans la stratégie OVA est que X_k n'appartiennent pas à la classe C_i. La somme des deux hypothèses étant égale à 1 comme il s'agit de probabilités on obtient la deuxième hypothèse facilement :

$$p(\overline{C_i}|X_k) = 1 - p(C_i|X_k) \tag{4.1}$$

Dans cette phase de prédiction, nous utilisons également l'information des frontières de segments de mixtures obtenue dans la phase de clustering (chapitre 3). Ainsi, nous cherchons à estimer pour chaque segment de texture τ, une probabilité d'appartenance à chaque classe $P(C_i|X_{k_\tau}, ..., X_{k_\tau+L_\tau-1})$. Pour réaliser cette estimation, nous faisons l'hypothèse simplificatrice classique que les observations $(X_{k_\tau}, ..., X_{k_\tau+L_\tau-1})$ sont indépendantes ce qui nous permet d'écrire :

$$P(C_i|X_{k_\tau}, ..., X_{k_\tau+L_\tau-1}) = \prod_{k=k_\tau}^{k_\tau+L_\tau-1} p(C_i|X_k) \tag{4.2}$$

Il peut arriver que le produit décrit dans 4.2 devienne trop petit pour des petites valeurs de $p(C_i|X_k)$ donc nous préférons utiliser :

$$\log\left(P(C_i|X_{k_\tau},...,X_{k_\tau+L_\tau-1})\right) = \sum_{k=k_\tau}^{k_\tau+L_\tau-1} \log p(C_i|X_k) \tag{4.3}$$

Connaissant $\log\left(p(\overline{C_i}|X_k)\right)$, nous obtenons $\log\left(P(\overline{C_i}|X_{k_\tau},...,X_{k_\tau+L_\tau-1})\right)$ de façon similaire :

$$\log\left(P(\overline{C_i}|X_{k_\tau},...,X_{k_\tau+L_\tau-1})\right) = \sum_{k=k_\tau}^{k_\tau+L_\tau-1} \log\left(1-p(C_i|X_k)\right) \tag{4.4}$$

Finalement, l'hypothèse retenue $h_{\tau,i}$ de l'appartenance du segment de mixture τ à la classe C_i est celle qui obtient la probabilité maximum :

$$h_{\tau,i} = \arg\max_{C_i,\overline{C_i}} \left(\log\left(P(C_i|X_{k_\tau},...,X_{k_\tau+L_\tau-1})\right), \log\left(P(\overline{C_i}|X_{k_\tau},...,X_{k_\tau+L_\tau-1})\right)\right) \tag{4.5}$$

Ainsi, pour chaque segment de mixture, nous disposons d'une estimation de son appartenance à chaque classe.

4.5.4 Apprentissage actif

4.5.4.1 Présentation

En apprentissage automatique, l'"apprenant" peut être vu comme une entité passive qui est entrainée à partir de données étiquetées par un utilisateur expert. Le but de l'apprentissage actif est d'améliorer les performances d'apprentissage en donnant un rôle actif à l'"apprenant". Afin de réaliser cet objectif, le domaine de l'apprentissage actif étudie les actions de sélection d'échantillons ainsi que les requêtes qui influencent l'introduction de nouvelles données d'apprentissage dans l'ensemble d'entrainement. Les principales motivations de l'apprentissage actif viennent de la difficulté d'obtenir des échantillons étiquetés. En effet, selon les différents domaines, l'obtention d'échantillons peut être coûteuse en temps et demander l'intervention d'un ou plusieurs experts. L'hypothèse de base sur laquelle s'appuie l'apprentissage actif est que lorsque les exemples devant être étiquetés manuellement sont sélectionnés de façon utile, la quantité de données requises pour un apprentissage efficace diminue fortement (Cohn et al. (1996)).

Pour réaliser cette sélection utile des échantillons à étiqueter, l'apprentissage actif utilise des *stratégies d'échantillonnage*, il en existe de nombreuses dans la littérature. Nous étudierons les différentes *stratégies d'échantillonnage* dans la section suivante. La procédure standard d'une méthode d'apprentissage actif peut être illustrée par l'algorithme 1. On peut noter que cette méthode d'apprentissage est également interactive car elle fait intervenir l'utilisateur.

On définit variables et fonctions suivantes :

- M : un modèle de prédiction.
- L, U : les ensembles de vecteurs respectivement étiquetés et non étiquetés du problème.
- n : le nombre d'exemples pour lesquels on souhaite demander une annotation manuelle par l'utilisateur.

Algorithme 1 Procédure de base d'apprentissage actif et interactif

```
1  Variable
2      M : modèle
3      L,U : ensembles de vecteurs
4      n,e : entiers
5  Début
6      Répéter
7          M ← train(L)
8          e = arg max_{x∈U} u(x, M)
9          label(e) ← input()
10         U ← U \ e
11         L ← L ∪ e
12     TantQue |L| < n
13 Fin
```

- e : l'échantillon sélectionné par la procédure d'échantillonnage.
- $train(L)$: fonction d'entrainement qui retourne le modèle appris avec un ensemble d'échantillons étiquetés L.
- $u(x, M)$: fonction d'échantillonnage qui retourne le degré d'utilité d'un échantillon $x \in U$ étant donné le modèle M.
- $label(e)$: fonction qui retourne l'étiquette d'un échantillon e.
- $input()$: fonction qui permet à l'utilisateur d'entrer une étiquette.

Dans la procédure d'apprentissage actif, M est entrainé avec les exemples de L puis l'échantillon e qui maximise la fonction d'utilité u est recherché. Ensuite on demande à l'utilisateur l'étiquette de e. Enfin, l'échantillon e est enlevé de l'ensemble des échantillons non étiquetés U puis ajouté à celui des étiquetés L.

4.5.4.2 Adaptation à notre problème

Dans notre cas, il y a deux aspects principaux sur lesquels nous devons réaliser une adaptation pour appliquer l'apprentissage actif à notre problème :

- Nous ne connaissons pas le nombre d'échantillons qui devront être annotés manuellement
- Nous souhaitons réaliser une sélection de *segments utiles* à l'apprentissage

Pour le premier point, il nous suffit de remplacer la condition d'arrêt de l'algorithme. En effet, au lieu de fixer le nombre d'échantillons à annoter manuellement, on peut considérer que l'utilisateur est le seul "maître à bord" et qu'il peut décider de la terminaison de l'algorithme quand il est satisfait du retour qu'il obtient.

La seconde adaptation revient à réaliser une intégration temporelle sur les *segments de mixtures* des *scores d'utilités* $u(k)$ obtenus pour les différentes trames/échantillons. Après des essais empiriques avec plusieurs opérateurs statistiques de base (médiane, moyenne, écart type etc.) il s'est avéré que l'opérateur de moyenne des *scores d'utilités* $u(k)$ d'un

segment constituait un meilleur choix. Le calcul de $u(k)$ dépend de la stratégie d'échantillonnage utilisée et sera abordé dans les sections 4.6.1.2 et 4.6.2.2. On calcule donc le score d'utilité $S(\tau)$ pour chaque segment :

$$S(\tau) = \frac{1}{L_\tau} \sum_{k=k_\tau}^{k_\tau+L_\tau-1} u(k) \tag{4.6}$$

Le segment $\mathcal{S}_0$ choisi est celui qui maximise le score $S(\tau)$:

$$\mathcal{S}_0 = \arg\max_{\tau} S(\tau) \tag{4.7}$$

Ainsi, le segment choisi est présenté à l'utilisateur expert qui peut valider ou corriger la prédiction du modèle.

4.6 Comparaison de deux approches interactives

Cette section présente les deux approches d'interaction proposées pour la classification d'objets sonores dans des pièces polyphoniques.

4.6.1 Approche par passages multiples (PM)

4.6.1.1 Concept

Cette première approche est inspirée d'une pratique courante qui consiste à écouter une pièce dans son intégralité pour réaliser la transcription d'un objet particulier. En procédant ainsi, on peut focaliser toute son attention uniquement sur cet objet et réaliser une transcription précise (on peut considérer que cette transcription est horizontale car elle suit l'évolution d'un objet particulier). Par analogie à cette pratique, nous proposons une première *boucle d'interaction* par *passages multiples* qui invite l'utilisateur à se concentrer sur un objet unique à la fois. Une nouvelle boucle d'interaction est réalisée pour chacune des Q classes d'intérêt. Cette approche a pour particularité de proposer un retour utilisateur simple car binaire que l'on peut résumer par la question suivante : *le segment proposé appartient-il à la classe courante ?* Le système propose une prédiction à l'utilisateur qui peut soit la corriger soit la valider. La *boucle d'interaction* par passages multiples est décrite dans l'algorithme 2.

On définit les variables et fonctions suivantes :
- $\mathcal{M}$: ensemble des modèles de prédiction pour chaque classe.
- $\mathcal{L}, \mathcal{U}$: les ensembles des segments de mixtures respectivement étiquetés et non étiquetés du problème.
- $\mathcal{C}$: ensemble des classes de sons visées.
- $\mathcal{S}_0$: le segment de mixture sélectionné par la procédure d'échantillonnage de segments.
- $train(\mathcal{L}, \mathcal{C}_i)$: fonction d'entrainement qui retourne le modèle appris pour la classe $\mathcal{C}_i$ avec un ensemble d'échantillons étiquetés $\mathcal{L}$.
- $E(\tau, \mathcal{M}_i)$: fonction d'échantillonnage qui retourne le degré d'utilité d'un segment $S_\tau \in \mathcal{U}$ étant donné le modèle $\mathcal{M}_i$.
- $label(\mathcal{S}_0)$: fonction qui retourne l'étiquette d'un segment $\mathcal{S}_0$.

Algorithme 2 Boucle d'interaction par passages multiples

```
1  Variable
2      M : ensemble de modèles
3      L ,U : ensemble de segments de mixtures
4      C : ensemble de classes
5      S_0 : segment de mixtures
6      i, Q, τ : entiers
7  Début
8      Pour i variant de 1 à Q Faire
9          Répéter
10             M_i ← train(L, C_i)
11             S_0 = arg max_{τ∈U} E(τ, M_i)
12             label(S_0) ← correction_pm(S_0)
13             U ← U \ S_0
14             L ← L ∪ S_0
15         TantQue l'utilisateur n'est pas satisfait
16 Fin
```

- *correction_pm*() : fonction qui demande à l'utilisateur de valider/corriger la prédiction.

Dans la *boucle d'interaction* par *passages multiples*, pour chaque classe i, $\mathcal{M}_i$ est entrainé avec les exemples de $\mathcal{L}$ puis le segment $\mathcal{S}_0$ qui maximise la fonction d'utilité E est recherché. Ensuite on demande à l'utilisateur de valider ou corriger $\mathcal{S}_0$. Enfin, le segment $\mathcal{S}_0$ est enlevé de l'ensemble des segments non étiquetés $\mathcal{U}$ puis ajouté à celui des étiquetés $\mathcal{L}$. La boucle *tant que* est répétée jusqu'à ce que l'utilisateur soit satisfait de la prédiction pour la classe i avant de passer à la classe suivante.

4.6.1.2 Stratégies d'échantillonnage

Dans cette section, nous présentons les différentes *stratégies d'échantillonnage* qui ont été testées dans le cadre de l'approche par *passages multiples*. En effet, nous avons défini une fonction de *score d'utilité* $u(k)$ dans la section 4.5.4.2 qui est utilisée pour calculer l'utilité d'un segment dans l'équation 4.6. Les *stratégies d'échantillonnage* sont différentes pour les deux approches d'interaction. Dans cette première approche, l'utilisateur se focalise sur une classe unique et le classifieur SVM est bi-classe. Nous cherchons à résoudre un problème OVA classique et par conséquent, il n'y a pas de recombinaison de classifieurs : soit l'échantillon appartient à la classe, soit il appartient à une autre classe. Les *stratégies d'échantillonnage* suivantes sont les plus courantes :

Stratégie de l'échantillon le plus positif ou "Most Positive" (MP)

Cette stratégie sélectionne les échantillons qui sont considérés par le classifieur comme étant les plus pertinents (Crucianu et al. (2004)). Dans notre cas, nous disposons d'une probabilité a postériori estimant l'appartenance à la classe visée. Par conséquent, les échantillons les plus pertinents sont ceux qui maximisent la probabilité a postériori. Géométriquement, ces échantillons sont les plus éloignés de la *surface de décision* et du côté positif. Cette stratégie a pour avantage de proposer à l'utilisateur rapidement des échantillons de la classe d'intérêt. L'utilisateur est donc rassuré car il n'a pas beaucoup de corrections à effectuer mais en contrepartie la généralisation peut prendre plus de temps.

Stratégie de l'échantillon le plus négatif ou "Most Negative" (MN)

Cette stratégie est le contraire de la précédente : elle sélectionne les échantillons qui sont considérés par le classifieur comme étant les moins pertinents (Wu et al. (2006)). Géométriquement, ces échantillons sont les plus éloignés de la *surface de décision* et du côté négatif (n'appartenant pas à la classe visée). Cette stratégie a pour avantage d'introduire de la diversité dans les données d'apprentissage en sélectionnant des échantillons considérés comme différents de la classe visée.

Stratégie de l'échantillon le plus ambigu ou "Most Ambiguous" (MA)

Cette dernière stratégie (également appelée stratégie de l'échantillon le plus informatif ou "Most Informative" dans la littérature) a pour but de sélectionner les échantillons qui apportent le plus d'informations au classifieur (Tong & Chang (2001)). Ainsi, les échantillons sélectionnés sont ceux qui sont les plus proches de la *surface de décision* dans l'espace des attributs. Cette stratégie a pour avantage de permettre au classifieur d'"affiner" la *surface de décision*. En théorie, cette approche doit permettre de généraliser rapidement dans le cas d'un classifieur discriminatif.

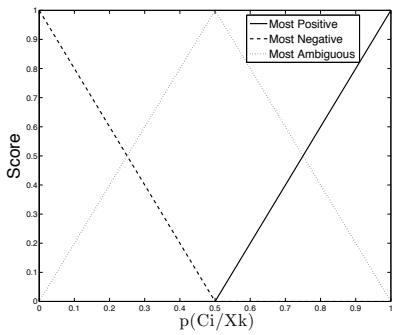

Fig. 4.6 – Courbes de calcul des scores d'utilités pour différentes stratégies

Le score d'utilité $u(k)$ est obtenu à partir de la probabilité a postériori en utilisant la courbe correspondante de la figure 4.6. On peut interpréter facilement ces courbes.

Dans le cas de la stratégie MP, les échantillons dont la probabilité est inférieure à 0.5 n'appartiennent probablement pas à la classe en question et auront donc un score nul. Par contre, les échantillons ayant une probabilité forte d'appartenance à la classe auront un score fort comme on peut le voir sur la courbe. Dans le cas de la stratégie MA, les échantillons les plus ambigus sont ceux ayant une probabilité incertaine située autour de la valeur 0.5. Donc, la courbe correspondante admet un score d'utilité maximum pour cette valeur de probabilité et des scores nuls pour les valeurs de probabilités certaines (c'est-à-dire 0 ou 1).

Nous avons réalisé une expérience préliminaire sur les pièces synthétiques du corpus polyphonique afin de comparer les différentes stratégies d'échantillonnage : échantillon le plus ambigu (MA), le plus positif (MP), le plus négatif (MN). Dans cette expérience, nous simulons les interactions avec l'utilisateur pendant la phase de classification. Le corpus synthétique utilisé pour cette expérience est polyphonique (corpus P) et nous utilisons la segmentation "parfaite" obtenue lors de la génération. Ainsi, nous évaluons la *f-mesure* à chaque itération de l'algorithme pour 100 pièces synthétiques de difficulté intermédiaire (3 sources sonores peuvent apparaitre simultanément au maximum). La simulation boucle jusqu'à ce que le score maximum soit atteint ($f - mesure = 1$). Nous calculons le score de *f-mesure* F_i pour la classe C_i en utilisant les prédictions au niveau de chaque segment :

$$F_i = \frac{2R_iP_i}{R_i + P_i} \tag{4.8}$$

où R_i et P_i sont les mesures de rappel et de précision de la classe i.

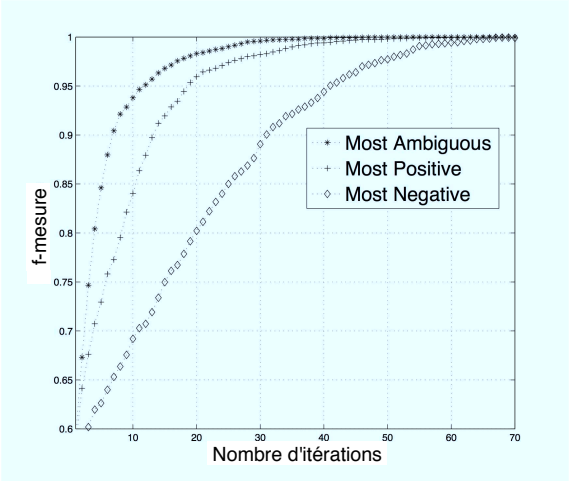

FIG. 4.7 – $f - mesure$ en fonction du nombre d'itérations pour 3 stratégies d'échantillonnage

La figure 4.7 présente une vue globale de la moyenne de *f-mesure* pour toutes les itérations de l'expérience. La figure 4.8 montre les performances détaillées des 3 stratégies d'échantillonnage sur les premières itérations. Dans cette dernière figure, la marque centrale des boites en fil de fer correspond à la médiane, les bords des boites sont les 25^{eme} et 75^{eme} centiles, les fils de fer s'étendent jusqu'aux minimas et maximas des données. Les deux figures nous permettent de voir que la stratégie MA est la plus performante suivie de la stratégie MP et de MN. La stratégie MA permet d'obtenir un score de *f-mesure* de 0.95 en 12 itérations en moyenne comme le montre la figure 4.7. Les deux autres stratégies,

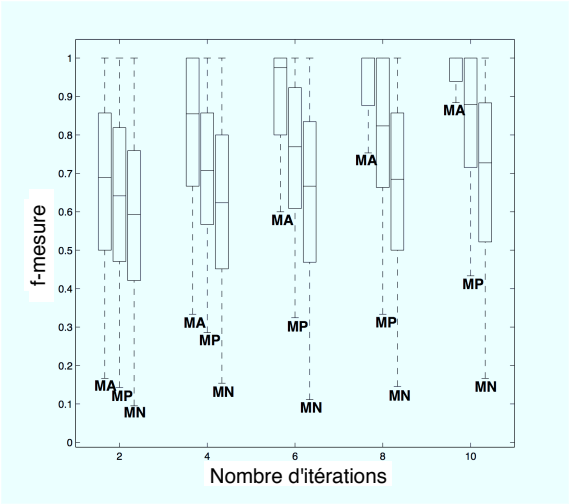

Fig. 4.8 – Comparaison de performances détaillées pour les premières itérations

respectivement MP et MN permettent d'obtenir les mêmes résultats en 19 et 41 itérations. L'efficacité de la stratégie MA avec les classifieurs SVM est un résultat attendu compte tenu du fait que cette stratégie tend à préciser la surface de décision en sélectionnant les échantillons ambigus. Cette expérience nous a permis de vérifier ce résultat sur nos données.

4.6.2 Approche par passage unique (PU)

4.6.2.1 Concept

Dans cette deuxième approche, le scénario d'interaction différents : nous considérons que l'utilisateur est capable de donner une information polyphonique sur les segments d'écoute sélectionnés par le système. En effet, il est difficile de suivre plusieurs sons à la fois sur toute la durée d'une pièce mais dans notre cas le système assiste l'utilisateur en lui proposant des segments de mixtures à écouter. Or, un musicologue a l'oreille entrainée et il sera capable de dire au système quelles sont les classes d'intérêt entendues dans un segment de mixture. Partant de cette hypothèse, nous proposons un deuxième scénario d'interaction par *passage unique*. La dénomination de cette approche vient du fait qu'une seule boucle est effectuée pour toutes les classes. Dans cette approche, plusieurs classifieurs sont utilisés puis leurs sorties sont combinées afin de proposer des prédictions d'étiquettes pour le problème *multilabel*. Le retour utilisateur, plus informatif que dans l'approche par *passages multiples*, peut être résumé par la question suivante : *le segment proposé contient-il les classes suivantes ?*. Le système propose à l'utilisateur une prédiction *multilabel* pour le segment en question et il doit corriger ou valider les différentes étiquettes. La *boucle d'interaction* par *passage unique* est décrite dans l'algorithme 3.

Ce dernier algorithme est proche de l'algorithme 2 que nous avons déjà présenté. Cependant certaines fonctions diffèrent quelque peu :

- $train(\mathcal{L}, \mathcal{C})$: fonction d'entrainement qui retourne l'ensemble des modèles appris pour l'ensemble des classes $\mathcal{C}$ avec un ensemble d'échantillons étiquetés $\mathcal{L}$.
- $E(\tau, \mathcal{M})$: fonction d'échantillonnage qui retourne le degré d'utilité d'un segment $S_\tau \in \mathcal{U}$ étant donné l'ensemble des modèles $\mathcal{M}$.

- $label(\mathcal{S}_0)$: fonction qui retourne l'ensemble des étiquettes d'un segment $\mathcal{S}_0$.
- $correction\_pu()$: fonction qui demande à l'utilisateur de valider la prédiction des différentes étiquettes.

Algorithme 3 Boucle d'interaction par passage unique

```
1  Variable
2      M : ensemble de modèles
3      L ,U : ensemble de segments de mixtures
4      C : ensemble de classes
5      S_0 : segment de mixtures
6      τ : entier
7  Début
8      Répéter
9          M ← train(L, C)
10         S_0 = arg max_{τ∈U} E(τ, M)
11         label(S_0) ← correction_pu(S_0)
12         U ← U \ S_0
13         L ← L ∪ S_0
14     TantQue l'utilisateur n'est pas satisfait
15 Fin
```

4.6.2.2 Stratégies d'échantillonnage

Cette section, présente des *stratégies d'échantillonnage* adaptées à l'approche par *passage unique* qui expose un problème différent puisqu'elle se situe dans un contexte d'apprentissage *multiclasse* à l'opposé de l'approche précédente qui était *bi-classe*. En effet, les critères d'échantillonnage présentés précédemment se basent sur les surfaces de décision des SVM qui séparent les échantillons en deux catégories mais dans l'approche par *passage unique* nous utilisons les résultats de plusieurs classifieurs indépendants qui sont combinés. Par conséquent, nous ne disposons pas de *surfaces de décisions* qui partitionnent les échantillons en Q classes. Cependant, l'estimation de probabilités a postériori pour chacun des classifieurs permet d'utiliser de nouveaux critères.

Entropie

L'entropie est utilisée pour mesurer l'incertitude d'une variable aléatoire. Cette information peut être très utile dans notre cas car elle permet de présenter à l'utilisateur un segment dont la classification est incertaine (Settles (2010)). Le calcul de l'entropie se fait à partir des estimations de probabilités a postériori p_i des Q classes :

$$H = -\sum_{i=1}^{Q} p_i \log p_i \tag{4.9}$$

Les valeurs d'entropie les plus grandes impliquent plus d'incertitude sur la distribution. Par conséquent, si un échantillon a une distribution de probabilité avec une forte entropie cela signifie que le classifieur n'est pas certain de l'appartenance aux classes. L'inconvénient de ce critère est qu'il est fortement influencé par les valeurs des classes peu importantes (celles ayant des probabilités faibles). Ainsi, un échantillon ayant des probabilités très proches pour deux classes différentes pourra se retrouver avec une entropie relativement faible alors que l'incertitude sur son appartenance aux différentes classes est grande.

Best Versus Second Best

Ce deuxième critère proposé par Joshi et al. (2009) tente de résoudre les problèmes du critère d'entropie. Dans ce critère, la mesure d'incertitude est calculée par la différence des probabilités des deux classes ayant les valeurs de probabilités les plus grandes. Cette approche est une mesure plus directe de l'incertitude d'appartenance à des classes. Le calcul de ce critère est fait à partir d'une fonction $\mathcal{D}$ qui utilise $\max 1(p)$ et $\max 2(p)$, respectivement la première et deuxième plus grande probabilité de p :

$$\mathcal{D}(p) = \max 1(p) - \max 2(p) \tag{4.10}$$

Ainsi les petites valeurs de $\mathcal{D}$ seront les plus incertaines. Pour rester conforme à l'équation 4.9 nous souhaitons que les valeurs grandes soient les plus incertaines donc dans la pratique, on obtient le critère *Best Versus Second Best* (BVSB) de la manière suivante :

$$BVSB(p) = 1 - \mathcal{D}(p) \tag{4.11}$$

Nous utilisons ce critère qui a déjà fait ses preuves dans le domaine de la classification d'images (voir Joshi et al. (2009)) et se révèle plus efficace que l'entropie en pratique.

4.6.2.3 Gestion de classifieurs

Nous proposons de comparer deux méthodes de classification différentes dans l'approche par *passage unique*. Les deux méthodes impliquent une gestion des classifieurs différente mais cela reste transparent pour l'utilisateur.

Méthode directe

Dans cette première méthode, la gestion des classifieurs est une version multiclasse de celle par *passages multiples*. La différence principale est que pour les Q classes considérées, à chaque itération, Q classifieurs sont considérés (soit un par classe). Ainsi, les échantillons positifs d'une classe C_q sont ceux qui contiennent cette classe et les négatifs ceux qui ne la contiennent pas. La décision d'appartenance d'un segment de mixture S_τ à la classe $\mathcal{C}_q$ est donnée par la règle de combinaison que nous avons présentée (équation 4.5). On peut considérer que cette méthode est une stratégie de classification de type *"Binary Relevance"* tel que nous l'avons décrit dans l'état de l'art (section 4.2.3).

Méthode par classifieurs de mixtures

Cette méthode diffère de la précédente car elle repose sur des *classes de mixtures*. Une classe de mixture est une classe potentiellement composée de la superposition de plusieurs sons. Ainsi, l'ensemble des *classes de mixtures* $\mathcal{C}_m$ est composé de l'ensemble des parties de l'ensemble des classes $\mathcal{C}$. Pour un cas simple à trois classes ($Q = 3$) nous avons donc les énumérations suivantes :

$$\mathcal{C} = \{a, b, c\}$$

$$\mathcal{C}_m = \{\emptyset, a, b, c, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

Pour un problème à Q classes, on dénombre $|\mathcal{C}_m| = 2^Q$ classes de mixtures potentielles. D'un point de vue sonore, la classe de mixture $C_{a \cup c}$ comprend les segments qui font entendre les sons a et c en même temps. Cependant, dans la musique électroacoustique, il est relativement rare d'entendre dans une même pièce toutes les superpositions possibles de sons. Par conséquent, dans cette méthode, nous proposons d'introduire dynamiquement les classifieurs de mixtures en fonction du retour utilisateur. De ce point de vue, la méthode proposée se rapproche des stratégies de classification multilabel de type *"Label Powerset"* (voir état de l'art en section 4.2.3), cependant notre approche ne considère que les mixtures directement exprimées par l'utilisateur. Par exemple, dans l'étape d'annotation manuelle du segment sélectionné par apprentissage actif, si l'utilisateur exprime que le segment appartient à la fois aux classes a et b, une nouvelle classe de mixture $C_{a \cup b}$ est créée et le classifieur correspondant est entrainé avec ce segment. L'algorithme commence avec un classifieur pour chaque classe puis le nombre de classifieurs M augmente au cours des itérations ($Q \leq M \leq 2^Q$). L'appartenance d'un segment S_τ à la $j^{ième}$ classe de mixture C_j parmi les M classes de mixtures exprimées est prédite de façon similaire à la *méthode directe* :

$$h_{\tau,j} = \arg \max_{C_j, \overline{C_j}} \left(\log \left(P(\mathcal{C}_j | X_{k_\tau}, ..., X_{k_\tau + L_\tau - 1} \right), \log \left(P(\overline{\mathcal{C}_j} | X_{k_\tau}, ..., X_{k_\tau + L_\tau - 1} \right) \right) \quad (4.12)$$

4.7 Evaluation

4.7.1 Simulation utilisateur

Pour réaliser l'évaluation, nous avons simulé les interactions de l'utilisateur avec le système au cours des différentes étapes. Dans cette section nous présentons la simulation de cette suite d'interactions ainsi que les évaluations réalisées sur le corpus P que nous avons décrit précédemment (section 2.5.3).

4.7.1.1 Segmentation

La segmentation a été présentée dans le chapitre 3. Nous rappelons qu'elle est utilisée pour trouver les frontières qui séparent les segments de mixtures dans le signal. De plus, la segmentation permet de regrouper les mixtures proches entre elles timbralement afin que l'utilisateur puisse les comparer. Dans le chapitre 3, nous tirons comme conclusion que le meilleur scénario à utiliser parmi les deux comparés est celui qui permet à l'utilisateur de "couper" les segments. Nous avons vu que la méthode de clustering hiérarchique offrait des possibilités d'interaction intéressantes pour adapter la segmentation au signal considéré mais il est également important que l'ensemble du système soit intuitif pour un utilisateur

qui n'a dans la plupart des cas pas de connaissances sur son fonctionnement interne. Par conséquent, nous décidons de garder un mode d'interaction simple lors de l'initialisation afin de rendre le système plus facile d'utilisation et de bien distinguer les étapes de segmentation et de classification. Le moyen d'interaction choisi est l'utilisation d'un slider qui permet d'obtenir une segmentation plus ou moins dense en fonction de sa position. La position d'origine du slider correspond au niveau le plus haut du *dendrogramme* à savoir la racine (pas de segmentation) et la position la plus haute du slider correspond au niveau des feuilles dans le *dendrogramme* (segmentation maximale). Ainsi, à chaque incrément du slider, on descend d'un niveau dans la hiérarchie du *dendrogramme* (cela correspond à une *coupe globale*). Cette approche permet à l'utilisateur d'obtenir un compromis sans grand effort puisque l'interaction est simple et il sera d'autant plus disponible pour la phase de classification pendant laquelle il est activement sollicité.

L'utilisateur a la capacité de trouver le bon positionnement du slider en réalisant un compromis entre le *rappel* et la *précision*. Pour arriver à ce résultat, l'utilisateur regarde les frontières positionnées sur le signal et en écoute les segments résultants. Pour simuler le positionnement du slider par l'utilisateur, nous évaluons la *f-mesure* pour chacune des valeurs possibles de ce dernier. Or, la position optimale du slider correspond au score de *f-mesure* le plus élevé qui représente le meilleur compromis entre le *rappel* et la *précision*. Nous garderons cette valeur optimale pour réaliser la simulation du choix des segments représentatifs.

4.7.1.2 Choix des segments les plus représentatifs

Comme nous l'avons vu dans la section 4.3, l'utilisateur a pour consigne de choisir les segments les plus représentatifs comme classes de départ pour amorcer l'apprentissage actif. Pour la sélection d'un segment d'initialisation d'une classe C_i, on souhaite choisir celui dans lequel l'instance de C_i est la plus dominante du point de vue du volume sonore en tenant compte des autres classes présentes. Ainsi, pour chaque segment de mixture S_τ, nous calculons des rapports d'énergies $r_\tau(i)$ entre les différentes sources sonores présentes dans le segment :

$$r_\tau(i) = E_\tau(i) / \sum_{l \neq i} E_\tau(l), \tag{4.13}$$

avec

$$E_\tau(i) = \sqrt{\frac{1}{L_\tau} \sum_{k=k_\tau}^{k_\tau + L_\tau - 1} x_i^2(k)} \tag{4.14}$$

et x_i est le signal de la classe C_i. Pour la classe C_i, le segment de mixture T_i qui maximise le rapport $r_\tau(i)$ est utilisé comme initialisation :

$$T_i = \arg\max_\tau r_\tau(i) \tag{4.15}$$

4.7.1.3 Classification des objets sonores

Une fois la sélection des segments représentatifs de chaque classe effectuée, nous cherchons à simuler la phase de classification. Le chapitre précédent illustre les différentes

étapes de classification avec la figure 4.5. La seule simulation intervenant dans ce processus semi-automatique est la correction des prédictions ainsi que les décisions d'arrêt de la boucle. La correction des prédictions est simulée facilement du fait que nous disposons de la *vérité terrain* construite pendant la génération des pièces synthétiques. Pour simuler les décisions d'arrêt, on considère que l'utilisateur interrompt le processus interactif lorsqu'il est satisfait ce qui se traduit par le dépassement d'un certain seuil de *f-mesure*. Par conséquent, la boucle d'interaction se termine lorsque chaque classe a atteint le seuil de satisfaction Fm_0. Dans les expériences, nous considérons que la prédiction est acceptable lorsque la valeur seuil $Fm_0 = 0.85$ est atteinte. Nous avons également simulé la fonction logicielle standard "annuler" : si le seuil de satisfaction Fm_0 est atteint pour une classe donnée, les résultats ne doivent pas décroître dans les itérations suivantes. Par conséquent nous supposons que l'utilisateur utilisera la fonction "annuler" si les résultats décroissent et la classe correspondante sera verrouillée afin de conserver les prédictions précédentes afin de les réutiliser à l'itération suivante sans mise à jour. De plus, lors de la sélection d'un segment par apprentissage actif, nous filtrons les segments de longueurs inférieures à 0.5 s car ils pourraient être mal jugés lors de la demande de retour utilisateur étant donné les limites de la perception humaine.

Ainsi, nous calculons la *f-mesure* pendant toutes les itérations de l'algorithme complet du système pour les 500 signaux synthétiques.

4.7.2 Résultats

4.7.2.1 Performances

Approche par passages multiples

Pour cette première approche, nous avons conservé la stratégie d'échantillonnage du plus ambigu pour réaliser l'évaluation. Ce choix a été motivé dans la section 4.6.1.2. La figure 4.9 présente la *f-mesure* moyenne obtenue pour 20 itérations de l'algorithme sur des classes individuelles pour les 5 niveaux de polyphonie.

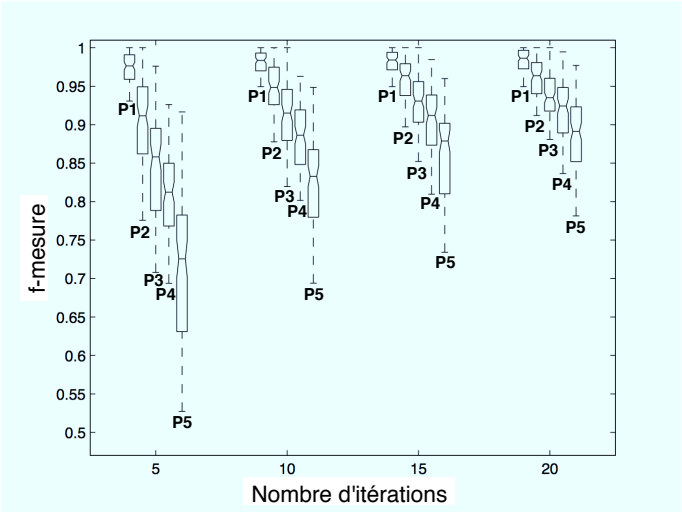

FIG. 4.9 – Score de *f-mesure* pour l'annotation d'une classe en fonction du nombre d'itérations pour une approche par passages multiples sur les 5 niveaux de polyphonie

Compte tenu de la nature de l'algorithme, les résultas sont donnés pour la prédiction

d'une classe unique. La figure 4.9 montre que les résultats décroissent en fonction de la difficulté polyphonique (c'est un résultat attendu compte tenu de la complexité croissante). Nous observons que de bons résultats sont obtenus après 10 itérations de l'algorithme pour un niveau de polyphonie acceptable : on obtient une *f-mesure* moyenne de 0.87 pour une complexité polyphonique de degré 4. Il est important de noter que compte tenu de la nature de l'approche qui permet à l'utilisateur de se concentrer sur une classe à la fois, le nombre d'itérations doit être multiplié par le nombre de classes visées présentes dans la pièce.

Approches par passage unique

Pour cette deuxième approche interactive, nous limitons le nombre d'itérations à 30 pour les évaluations car nous souhaitons obtenir des résultats acceptables en un nombre raisonnable d'interactions. La figure 4.10 compare la méthode de classification directe (PU1) à la méthode par classifieurs de mixtures (PU2) pour des complexités polyphoniques de 2 (figure de gauche) et de 4 (figure de droite).

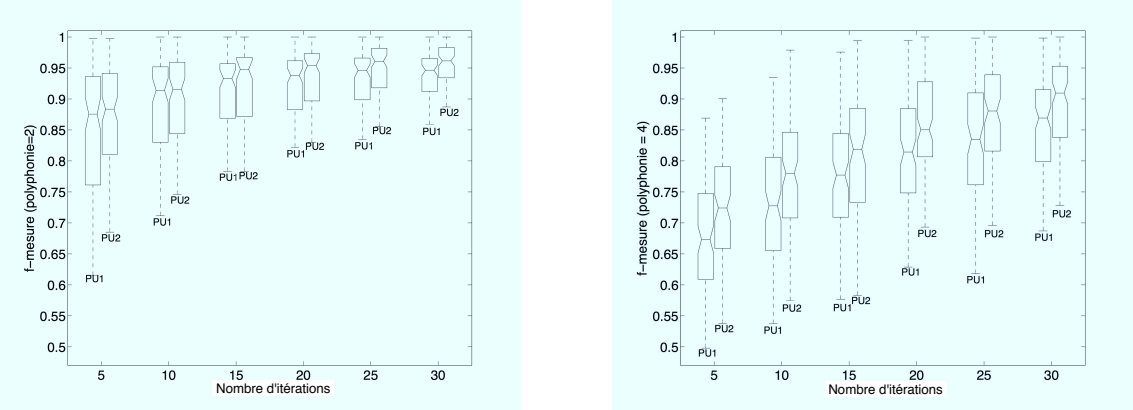

Fig. 4.10 – Score de *f-mesure* pour l'annotation d'une pièce complète en fonction du nombre d'itérations pour les deux méthodes par passage unique (la figure de gauche présente les résultats pour un degré de polyphonie de 2 et celle de droite pour un degré de polyphonie de 4).

Les résultats montrent que la méthode qui introduit des classifieurs de mixtures dynamiquement en fonction du retour utilisateur (PU2) permet d'obtenir un gain faible mais consistant sur l'ensemble des niveaux de polyphonie. Ces deux méthodes considèrent en même temps toutes les classes d'intérêt et nous observons qu'elles permettent de réduire le nombre d'itérations en comparaison de l'approche par *passages multiples* dans laquelle l'utilisateur doit répéter le processus de classification autant de fois qu'il y a de classes. On remarque que l'on atteint une *f-mesure* médiane acceptable (elle dépasse le seuil de satisfaction de 0.85) en moins de 5 itérations pour une complexité polyphonique de 2. Pour obtenir une *f-mesure* médiane similaire avec un degré de complexité polyphonique de 4, il faudra réaliser 25 itérations.

4.7.2.2 Complexité des méthodes

Dans cette section nous vérifions que les approches proposées sont utilisables dans le cadre d'une application interactive. Une contrainte fonctionnelle que nous avions citée dans la section 2.5.2.1 est de proposer des approches réactives. Autrement dit, il ne faut pas que

l'utilisateur attende trop longtemps entre chaque itération de l'algorithme principal. Les interactions pendant la phase de segmentation sont quasiment instantanées sur une machine standard compte tenu de l'efficacité de la structure hiérarchique. Nous nous intéresserons donc à la phase de classification des objets sonores qui est la plus complexe.

Dans cette phase, nous nous intéressons au temps d'attente de l'utilisateur entre chaque itération de l'algorithme. Ce temps d'attente dépend principalement de l'algorithme de classification basé sur les SVMs car le reste de l'algorithme est constitué d'opérations négligeables. La complexité de l'algorithme SVM dépend grandement de l'implémentation utilisée. Nous avons utilisé celle de Chang & Lin (2011) pour ce travail qui est une librairie efficace, écrite en C. Au sujet de la complexité de cet algorithme, Chang & Lin (2011) précise que de nombreux travaux ont étudié la complexité des méthodes SVM (voir List & Simon (2005)) mais que ces travaux sont consacrés à des méthodes différentes. De plus, il n'y a actuellement pas de résultats théoriques sur le nombre d'itérations de la méthode utilisée mais Chang & Lin (2011) affirme qu'empiriquement il est reconnu que la complexité de cet algorithme est polynomiale.

La figure 4.11 mesure le temps d'attente total imposé par le système pour l'annotation d'un fichier avec la méthode PM pour différents niveaux de polyphonie.

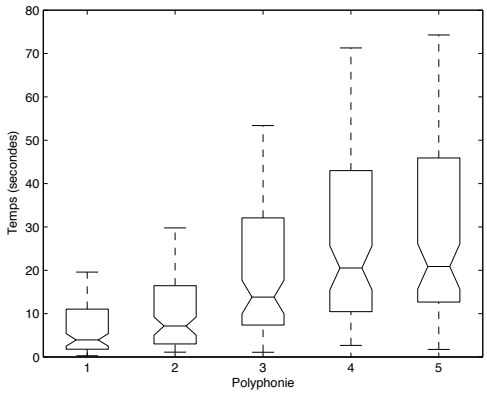

Fig. 4.11 – Temps d'attente total pour l'annotation d'un fichier avec la méthode PM en fonction du niveau de polyphonie

On peut interpréter facilement cette figure en considérant que le temps mesuré correspond au temps d'annotation totale d'une pièce sans compter le temps de réflexion et les interactions de l'utilisateur. On constate que cette méthode est tout à fait utilisable : on observe des médianes à 4 secondes pour les pièces monophoniques, 14 secondes pour un degré polyphonique intermédiaire de 3 et 20 secondes pour un degré polyphonique de 5. Cependant, comme nous l'avons évoqué dans la section précédente, le nombre d'interactions utilisateur demandé par le système est plus important que dans les autres méthodes. Pour le temps d'exécution d'une itération, dans le cas des pièces de degré polyphonique de 5, on mesure un temps minimum de 0,07 secondes et un temps maximum de 0,8 secondes.

La méthode PU2 (figure 4.12) expose des temps qui peuvent être considérés comme acceptables compte tenu du fait que toutes les classes sont gérées en même temps : on observe des médianes à 18, 42 et 44 secondes pour des polyphonies respectives de 1, 3 et 5. Cette efficacité s'explique par le fait qu'à chaque itération, on introduit les nouveaux échantillons dans un unique classifieur (dans cette méthode nous avons un classifieur pour

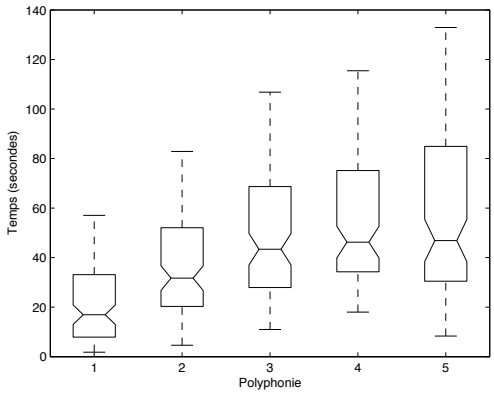

Fig. 4.12 – Temps d'attente total pour l'annotation d'un fichier avec la méthode PU2 en fonction du niveau de polyphonie

chaque mixture et un segment ne peut être affecté qu'à une unique mixture). Par conséquent, chaque itération ne nécessite de lancer qu'une unique tâche de classification. Cette méthode a tendance à faire augmenter le nombre de classifieurs mais chaque classifieur a peu d'échantillons : cela permet de garder des temps de calculs acceptables pour l'unique tâche de classification lancée à chaque itération. Le temps d'exécution d'une itération, dans le cas des pièces les plus complexes, est compris entre 0,53 et 5,12 secondes.

On peut préciser que la méthode PU1 est la moins exploitable dans le contexte d'une application interactive. En effet, pour un segment de mixture contenant n classes, ses échantillons seront introduits dans les classifieurs des n classes correspondantes. Cette méthode n'est pas efficace car elle a tendance à augmenter la quantité d'échantillons dans chaque classifieur et elle impose souvent d'effectuer plusieurs tâches de classification à chaque itération.

4.7.2.3 Analyse des descripteurs sélectionnés

Comme nous l'avons vu dans la section 4.5.2, la sélection d'attributs est réalisée dans la boucle d'interaction, pour chaque itération. Cette section expose les résultats de sélection des descripteurs ainsi que les variations de cette sélection.

La figure 4.13 présente un classement des 20 descripteurs les plus sélectionnés pour les deux approches d'interaction. On peut remarquer que les descripteurs sélectionnés par les deux méthodes varient quelque peu : les coefficients de loudness, MFCC et moments spectraux sont plus présents dans l'approche PM alors que l'approche PU sélectionne un bon nombre de coefficients OBSI dont le rôle est de *"capturer de façon sommaire la distribution de puissance des différentes harmoniques du son"* (Essid (2005)). Etant donné que l'approche PU crée des classifieurs de mixtures, et que les sons utilisés pour les pièces synthétiques sont en bonne partie des sons harmoniques, on peut effectivement s'attendre à ce que les différentes mixtures aient des distributions harmoniques relativement différentes. La sélection automatique de coefficients OBSI pour distinguer les mixtures sonores semble donc pertinente. De façon générale certains descripteurs tels que les coefficients de loudness sont bien représentés dans les deux approches. On peut remarquer que parmi les 20 descripteurs les plus sélectionnés présentés, 12 sont en commun entre les deux approches : le coefficient MFCC :1 (qui est proche de l'énergie du signal), des coefficients de loudness, des

Descripteur :attribut
Loudness :1
MFCC :1
Loudness :2
OBSI :2
Largeur spectrale
OBSIR :2
OBSI :1
Platitude spectrale
MFCC :3
OBSIR :4
Loudness :3
OBSI :3
Etalement perceptif :1
Asymétrie spectrale
Loudness :19
Loudness :4
Acuité perceptive :1
Fréquence de coupure :1
Pente spectrale :1
MFCC :2

Descripteur :attribut
MFCC :1
Acuité perceptive :1
Loudness :1
Asymétrie spectrale
OBSIR :4
OBSI :2
Loudness :2
LSF :1
OBSI :3
OBSI :4
Loudness :3
OBSI :1
Etalement perceptif :1
OBSI :8
Largeur spectrale
OBSI :5
Platitude spectrale
OBSI :9
Taux de passage par zéro :1
OBSIR :2

FIG. 4.13 – Classement des 20 descripteurs les plus sélectionnés pour l'approche PM (à gauche) et PU (à droite). Chaque descripteur est présenté dans le format *Nom du descripteur : Numéro de l'attribut.*

descripteurs perceptifs, des coefficients OBSI/OBSIR. Parmi ces 12 attributs en commun, la différence principale entre les deux méthodes est le rang dans le classement qui diffère d'une méthode à l'autre.

Les figures 4.14 et 4.15 présentent la variation des descripteurs sélectionnés pour différentes itérations de l'algorithme. On remarque une tendance commune entre les deux approches : les attributs sélectionnés à partir de l'itération 10 varient très peu. En effet, pour l'approche PM, entre les itérations 10 et 30, seul un attribut diffère (le coefficient Loudness :3 est remplacé par MFCC :3) dans la sélection (les attributs communs ont cependant des rangs différents). On remarque également pour l'approche PU que seulement 2 attributs diffèrent entre les itérations 10 et 30. Ce résultat est intéressant car il nous permet de déduire que la sélection "utile" se fait pendant les premières itérations de l'algorithme.

Les figures 4.16 et 4.17 présentent la variation des descripteurs sélectionnés pour les différents niveaux de polyphonie. On remarque que les attributs sélectionnés varient de manière importante entre les différents niveaux de polyphonie. En effet, dans les deux approches, il n'y a que 4 attributs en commun sur 10 pour des niveaux de polyphonie extrêmes. Cette observation met en valeur l'importance de choisir des descripteurs spécifiques lors d'un problème de classification polyphonique.

Itération 1	**Itération 10**	**Itération 20**	**Itération 30**
OBSIR :4	Loudness :1	Loudness :1	Loudness :1
Loudness :1	MFCC :1	MFCC :1	MFCC :1
MFCC :1	Loudness :2	Loudness :2	Loudness :2
Asymétrie spectrale	OBSI :2	OBSI :2	OBSI :2
Loudness :2	OBSIR :4	Largeur spectrale	Largeur spectrale
OBSI :2	Largeur spectrale	OBSIR :2	OBSIR :2
Acuité perceptive :1	Loudness :3	OBSI :1	OBSI :1
OBSIR :2	Platitude spectrale :1	OBSIR :4	Platitude spectrale :1
Etalement perceptif :1	OBSI :1	Platitude spectrale :1	MFCC :3
Loudness :3	OBSIR :2	Loudness :3	OBSIR :4

Fig. 4.14 – Variation des descripteurs sélectionnés pour l'approche PM pour les itérations 1, 10, 20 et 30. Chaque descripteur est présenté dans le format *Nom du descripteur : Numéro de l'attribut*.

Itération 1	**Itération 10**	**Itération 20**	**Itération 30**
OBSIR :4	MFCC :1	MFCC :1	MFCC :1
Loudness :1	Loudness :1	Acuité perceptive :1	Acuité perceptive :1
MFCC :1	OBSIR :4	Loudness :1	Loudness :1
Asymétrie spectrale	Asymétrie spectrale	Asymétrie spectrale	Asymétrie spectrale
Loudness :2	Acuité perceptive :1	OBSIR :4	OBSIR :4
Acuité perceptive :1	Loudness :2	Loudness :2	OBSI :2
OBSI :2	LSF :1	OBSI :2	Loudness :2
OBSIR :2	OBSI :2	LSF :1	LSF :1
Etalement perceptif :1	Loudness :3	OBSI :3	OBSI :3
LSF :1	Etalement perceptif :1	Loudness :3	OBSI :4

Fig. 4.15 – Variation des descripteurs sélectionnés pour l'approche PU pour les itérations 1, 10, 20 et 30. Chaque descripteur est présenté dans le format *Nom du descripteur : Numéro de l'attribut*.

4.8 Conclusion

Dans ce chapitre, nous avons présenté une solution adaptée au problème de classification des objets sonores. La méthode proposée permet d'obtenir un multilabel pour chacun des segments de mixtures. Ainsi, on peut connaitre l'appartenance relative d'un segment à chacune des classes. L'approche proposée est interactive et s'adapte aux choix de l'utilisateur. De plus, nous avons comparé deux approches différentes d'interaction. Il est important de noter que la première approche, par *passages multiples*, a pour avantage de demander un retour simple à l'utilisateur. La deuxième approche, par *passage unique*, demande un effort plus important à l'utilisateur qui doit corriger les prédictions pour toutes les classes visées dans les segments sélectionnés par le système ce qui peut constituer un facteur de fatigue.

Une évaluation par simulation utilisateur nous a permis de comparer les deux approches d'interaction. L'évaluation montre que l'approche par *passages multiples* est plus adaptée à un petit nombre de classes : si le nombre de classes à annoter est important, des résultats satisfaisants peuvent être obtenus en un nombre d'itérations inférieur avec la méthode

Polyphonie 1	**Polyphonie 2**	**Polyphonie 3**
MFCC :1	MFCC :1	Loudness :1
Loudness :2	Loudness :1	Loudness :19
Loudness :1	Etalement perceptif :1	MFCC :1
Loudness :3	Loudness :19	Largeur spectrale
Asymétrie spectrale	OBSIR :4	Loudness :2
Etalement perceptif :1	Loudness :2	OBSIR :2
Largeur spectrale	Acuité perceptive :1	OBSIR :4
OBSIR :4	Loudness :3	Etalement perceptif :1
LSF :4	OBSI :2	Platitude spectrale :1
Platitude spectrale :1	MFCC :2	Fréquence de coupure :1

Polyphonie 4	**Polyphonie 5**
Loudness :1	OBSI :2
OBSIR :2	Loudness :1
OBSI :2	OBSIR :2
Largeur spectrale	Largeur spectrale
Etalement perceptif :1	OBSI :1
Loudness :2	Loudness :2
MFCC :3	MFCC :1
Asymétrie spectrale	MFCC :3
OBSI :3	Loudness :4
Loudness :4	OBSI :3

FIG. 4.16 – Variation des descripteurs sélectionnés pour l'approche PM pour les différents niveaux de polyphonie. Chaque descripteur est présenté dans le format *Nom du descripteur : Numéro de l'attribut.*

par classifieurs de mixtures qui est l'approche par *passage unique* la plus efficace. Nous avons également montré que les approches PM et PU2 sont tout à fait exploitables dans le contexte d'une application interactive car les temps de calculs mesurés pour ces méthodes sont acceptables. Nous avons également réalisé une étude des descripteurs sélectionnés dans la boucle d'interaction. Cette étude met en évidence le fait que les descripteurs sélectionnés varient en fonction du niveau de difficulté polyphonique. De plus, la sélection de descripteurs varie principalement pendant les premières itérations de l'algorithme.

Polyphonie 1	**Polyphonie 2**	**Polyphonie 3**
OBSIR :4	MFCC :1	MFCC :1
Loudness :1	OBSIR :4	Acuité perceptive :1
MFCC :1	Loudness :1	OBSIR :4
Loudness :2	Acuité perceptive :1	Asymétrie spectrale
Loudness :3	Loudness :2	Loudness :1
Acuité perceptive :1	Asymétrie spectrale	Loudness :2
Asymétrie spectrale	OBSI :2	LSF :1
OBSIR :2	LSF :1	Loudness :3
OBSI :2	Loudness :3	OBSI :3
Centroïde spectral	Etalement perceptif :1	OBSI :2

Polyphonie 4	**Polyphonie 5**
MFCC :1	MFCC :1
Acuité perceptive :1	OBSI :3
Loudness :1	OBSI :4
Asymétrie spectrale	OBSI :2
LSF :1	OBSI :9
OBSI :2	OBSI :8
OBSI :8	Asymétrie spectrale
OBSI :9	Acuité perceptive :1
OBSI :3	OBSI :1
Loudness :2	LSF :1

Fig. 4.17 – Variation des descripteurs sélectionnés pour l'approche PU pour les différents niveaux de polyphonie. Chaque descripteur est présenté dans le format *Nom du descripteur : Numéro de l'attribut.*

Chapitre 5

Conclusion

5.1 Bilan

Dans cette thèse, nous avons proposé un système interactif destiné à aider l'analyse des musiques électroacoustiques. Le système procède d'abord par une segmentation interactive à l'issue de laquelle l'utilisateur pourra sélectionner les segments qui contiennent les objets sonores qu'il vise pour illustrer le point de vue de son analyse. Ensuite, le système entre dans une boucle d'interaction dans laquelle il présentera des segments à l'utilisateur pour qu'il puisse les valider ou les corriger. Ce processus itératif prend en compte les informations apportées par l'utilisateur afin d'améliorer la classification des objets sonores.

Nous avons réalisé une étude sur les pratiques d'analyse des musicologues ce qui nous a permis de dégager des pistes de travail adaptées aux besoins réels et de comprendre la philosophie générale de l'analyse des musiques électroacoustiques qui se distingue de l'analyse musicale "traditionnelle". Il est également important de considérer la nature subjective du point de vue de l'analyse. Il ressort de cette étude que les transcriptions de pièces en objets sonores sont rarement réalisées entièrement car les musicologues se focalisent principalement sur les objets saillants et sur ceux qui illustrent le point de vue de leur analyse. Parmi les souhaits des musicologues, nous nous sommes concentrés principalement sur la segmentation et la classification d'objets sonores afin de retrouver leurs différentes instances dans une pièce.

La première phase du système est une segmentation timbrale interactive qui repose sur une segmentation temporelle obtenue par détection de transitoires suivie d'un clustering hiérarchique. Nous avons comparé deux approches d'interactions qui ont été évaluées par simulation de l'utilisateur à l'aide d'un corpus synthétique. L'évaluation a montré qu'il est possible d'améliorer les performances de segmentation en réalisant des "coupes locales" de dendrogrammes qui exploitent le retour de pertinence. Cette phase de segmentation nous permet d'obtenir à la fois des frontières entre les mixtures sonores ainsi que de regrouper les segments similaires timbralement. Ainsi, l'utilisateur peut choisir un segment de mixture représentatif pour chaque classe sonore qu'il souhaite étudier.

La deuxième phase du système propose une approche de classification interactive des objets sonores que l'utilisateur souhaite étudier dans une pièce électroacoustique. A chaque itération de l'algorithme, un nouveau segment est sélectionné automatiquement par apprentissage actif et l'utilisateur corrige ou valide les prédictions du système. Nous pouvons ainsi obtenir un ensemble d'étiquettes donnant l'appartenance relative aux classes étudiées pour les segments de mixtures définis dans l'étape de segmentation. Deux approches d'interactions ont été comparées en simulant l'utilisateur sur un corpus synthétique à polyphonie

variable. La première approche, dite par passages multiples, a pour avantage de demander un retour très simple à l'utilisateur mais prend plus de temps pour réaliser l'annotation d'une pièce. La deuxième approche, par passage unique, est plus efficace mais demande à l'utilisateur plus d'attention car le retour de pertinence est plus complexe et la gestion de "verrouillage des classes" est également contrôlée par l'utilisateur. Nous avons également proposé une méthode de classification de type "multilabel", dans une approche "Label Powerset", orientée "mixtures exprimées par l'utilisateur" et par passage unique qui permet d'obtenir une amélioration consistante des performances sur l'ensemble des niveaux de polyphonie évalués.

Ces travaux sont une première pierre à l'édifice. En effet, le sujet étant relativement inexploré à l'origine, de nombreux problèmes abordés par les musicologues pourraient être explorés. De plus, la démarche proposée pourrait encore être améliorée sur certains points.

5.2 Perspectives

Une partie des souhaits qui ont été exprimés par les musicologues dans la section 2.4.2 ne sont pas pris en compte par le système proposé qui constitue une base sur laquelle il serait possible de greffer des fonctionnalités additionnelles. Couprie et Delhaye souhaiteraient par exemple pouvoir trouver les grandes périodes dans une pièce ce qui pourrait être réalisé en analysant le repérage des instances d'objets sonores retrouvées par notre système. En effet, dans certaines pièces, les suites d'objets récurrentes pourraient constituer des motifs donnant des indices sur la structure. Delhaye parle également d'un système qui permet de "séparer les différentes voix de mixage d'une pièce". Or, une fois les objets sonores principaux repérés par notre système, cette connaissance pourrait informer un algorithme de séparation de sources pour réaliser le "démixage" de ces objets (Hennequin et al. (2011)). On peut également citer Delalande qui considère comme important le fait de "pouvoir réaliser des symboles graphiques à la main". Ainsi, une intégration de notre système à une architecture logicielle ouverte telle que l'acousmographe permettrait d'obtenir le "meilleur des deux mondes".

Nous avons montré l'intérêt de la coupe locale du dendrogramme pour une tâche de segmentation pure. Cependant, cette approche introduit un nombre d'interactions supplémentaires non négligeable. Il pourrait être intéressant de tester une approche qui laisse à l'utilisateur la possibilité de modifier manuellement des frontières de segmentation. Ainsi, si certains segments caractéristiques de classes ne sont pas segmentés correctement, l'utilisateur pourrait intervenir directement et corriger la segmentation pour la rendre plus adaptée. De plus, nous nous sommes limités au timbre pour la segmentation mais il pourrait être envisageable de considérer d'autres aspects (enveloppes temporelles, hauteurs etc.) en utilisant des descripteurs différents. Il serait également possible d'essayer d'exploiter l'information d'étiquetage donnée par le clustering : les segments proches timbralement ont la même étiquette. Pour démarrer l'apprentissage d'une classe, il serait donc possible d'utiliser les échantillons des segments de même étiquette que l'objet sonore visé afin d'accélérer le processus d'annotation. Cependant, ces segments n'étant pas confirmés par l'utilisateur, il ne sont pas aussi fiables que ceux directement sélectionnés par ce dernier.

Pour la phase de sélection d'attributs qui est réalisée à chaque itération de l'algorithme de classification, la méthode testée dans notre système est une des plus simples et a pour intérêt principal d'être très rapide. Cependant, l'emploi d'une méthode plus évoluée tout en restant rapide permettrait sans doute d'obtenir de meilleures performances générales. De plus, en ce qui concerne les deux approches d'interaction comparées, il serait envisageable de

considérer une approche hybride par passage unique et passages multiples afin de minimiser l'effort fourni par l'utilisateur.

Un bon nombre de pièces électroacoustiques utilisent des effets de spatialisation étant donné qu'elles sont la plupart du temps diffusées sur un "orchestre de hauts-parleurs"[1]. Ainsi, la composition de l'espace à une importance capitale dans ce type de pièces. Par conséquent, il serait pertinent d'envisager des descripteurs multicanaux qui puissent décrire les effets de spatialisation afin de les intégrer dans le système d'analyse.

Certains compositeurs de musiques électroacoustiques utilisent les différents effets et traitements sonores pour présenter les instances d'un même objet sonore de façons différentes. Il pourrait être intéressant de prendre en compte ces traitements et altérations des différentes instances d'un même objet dans la phase de classification.

Le système a été évalué à partir d'un corpus synthétique ce qui semble être une étape indispensable afin d'obtenir des résultats sur une quantité raisonnable de données et d'observer le comportement des méthodes en fonction de la complexité polyphonique. Cependant, une évaluation avec des utilisateurs réels pourrait sans doute mettre en valeur des améliorations possibles des modes d'interactions du système. De plus, nous sommes conscients qu'on ne peut pas tout simuler et que la confrontation à des pièces réelles serait un indicateur précieux pour des améliorations possibles du système.

[1] http ://www.inagrm.com/categories/un-orchestre-de-haut-parleurs

Annexe A

Echantillons sonores utilisés

Cette annexe présente les différents échantillons sonores utilisés lors de la création des deux corpus.

A.1 Corpus Monophonique

La figure A.1 présente les échantillons sonores utilisés pour la phase de sélection d'attributs et la figure A.2 ceux utilisés pour la phase de test. La catégorie "environnemental" correspond aux sons fournis par l'INA et la catégorie "instrumental" aux sons de la base RWC (Goto et al. (2002)). On peut noter qu'il y a plus de sons instrumentaux mais l'équilibre entre les deux types de signaux a été pris en compte lors de la génération. Il est difficile de donner la durée des sons car chaque source sonore dispose de multiples instances mais pour la génération une unique instance est sélectionnée arbitrairement puis utilisée.

A.2 Corpus Polyphonique

La figure A.3 présente les différents échantillons sonores utilisés pour la création du corpus polyphonique. Le nom de chaque échantillon correspond aux initiales du compositeur l'ayant fourni suivi du numéro de l'échantillon. On remarque que la moitié des échantillons sont polyphoniques et que leurs durées se situent entre 1 et 48 secondes.

Source sonore	**Catégorie**
Applaudissements	environnemental
Sonnerie	environnemental
Sifflet	environnemental
Contrebasse	instrumental
Cornet à piston	instrumental
Bugle	instrumental
Piano	instrumental
Trombone	instrumental
Saxophone ténor	instrumental
Tuba	instrumental
Alto	instrumental

Fig. A.1 – Echantillons sonores utilisés pour la création du corpus monophonique lors de la phase de sélection d'attributs.

Source sonore	**Catégorie**
Ambiance urbaine	environnemental
Rires	environnemental
Sirènes	environnemental
Saxophone alto	instrumental
Clarinette	instrumental
Cor anglais	instrumental
Flûte	instrumental
Cor d'harmonie	instrumental
Hautbois	instrumental
Flûte de pan	instrumental
Saxophone soprano	instrumental
Trompette	instrumental
Violoncelle	instrumental
Violon	instrumental

Fig. A.2 – Echantillons sonores utilisés pour la création du corpus monophonique lors de la phase de test.

Nom	Source sonore	Polyphonique	Durée
AB1	Synthétiseur	oui	2s
AB2	Synthétiseur	oui	2s
AB3	Synthétiseur	non	4s
AB4	Synthétiseur	non	17s
AB5	Instrument électrique	non	3s
AB6	Instrument électrique	non	4s
AB7	Synthétiseur	oui	1s
AB8	Instrument acoustique	oui	20s
DL1	Acoustique	oui	25s
DL2	Acoustique	oui	6s
DL3	Acoustique	non	7s
DL4	Acoustique	non	13s
DL5	Cymbales	non	15s
DL6	Acoustique	non	21s
DL7	Vocale	non	1s
DT1	Instruments acoustique	oui	41s
DT2	Acoustique	oui	2s
DT3	Synthétiseur	oui	11s
DT4	Acoustique	oui	31s
DT5	Synthétiseur	non	44s
DT6	Synthétiseur	non	17s
DT7	Synthétiseur	oui	6s
DT8	Synthétiseur	oui	48s
DT9	Acoustique	non	19s

Fig. A.3 – Echantillons sonores utilisés pour la création du corpus polyphonique. Chaque ligne du tableau correspond à un échantillon unique.

Annexe B

Descripteurs utilisés

B.1 Descripteurs Spectraux

Les *moments spectraux* (centroïde spectral, largeur spectrale, asymétrie spectrale, platitude spectrale) permettent de décrire différentes caractéristiques spectrales. Ces descripteurs ont été utilisés avec succès notamment dans Gillet & Richard (2004) pour la transcription de boucles de batterie. On utilise les moments μ_i pour le calcul des 4 *moments spectraux* :

$$\mu_i = \frac{\sum_{k=0}^{K-1} (f_k)^i a_k}{\sum_{k=0}^{K-1} a_k}, \tag{B.1}$$

avec a_k est l'amplitude de la $k^{\text{ième}}$ composante fréquentielle du spectre et $f_k = \frac{k}{N}$ est la fréquence correspondante.

Centroïde spectral

Il s'agit du barycentre du spectre calculé en considérant le spectre comme une distribution. Il est souvent utilisé pour caractériser la "brillance" d'un spectre en mesurant l'équilibre entre les basses fréquences et les hautes fréquences :

$$M_1 = \mu_1 \tag{B.2}$$

Largeur spectrale

Etalement du spectre autour de la valeur moyenne :

$$M_2 = \sqrt{\mu_2 - \mu_1^2} \tag{B.3}$$

Asymétrie spectrale

Mesure l'asymétrie de la distribution autour de la valeur moyenne (correspond au moment statistique d'ordre 3). Une valeur d'asymétrie nulle correspond à une distribution symétrique, une valeur négative indique qu'il y a plus d'énergie dans la partie gauche du spectre, une valeur positive indique qu'il y a plus d'énergie dans la partie droite du spectre :

$$M_3 = \frac{2\mu_1^3 - 3\mu_1\mu_2 + \mu_3}{M_2^3} \tag{B.4}$$

Platitude spectrale

Mesure la platitude de la distribution autour de la valeur moyenne, elle est calculée à partir du 4$^{\text{ème}}$ moment statistique :

$$M_4 = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{M_2^4} - 3 \tag{B.5}$$

Platitude d'Amplitude Spectrale par bandes

Mesure des proportions relatives de bruit et de composantes sinusoïdales du spectre. Ce critère est calculé par le rapport des moyennes géométriques et arithmétiques de l'énergie du spectre dans différentes bandes de fréquences :

$$PAS(bf) = \frac{(\sum_{k\in bf} A_k)^{1/k}}{\frac{1}{K}\sum_{k\in bf} A_k}, \tag{B.6}$$

ou A_k est l'amplitude de la $k^{ième}$ bande de fréquences et bf est l'ensemble des bandes.

Généralement, on distingue les 4 bandes de fréquences suivantes :
- de 250 à 500 Hz
- de 500 à 1000 Hz
- de 1000 à 2000 Hz
- de 2000 à 4000 Hz

Pour un signal bruité, la valeur PAS est proche de 1. Le cas échéant, pour un signal essentiellement composé de sinusoïdes, la valeur PAS est proche de 0.

Platitude Spectrale Globale

Mesure des proportions relatives de bruit et de composantes sinusoïdales sur l'ensemble du spectre :

$$PLASG(bf) = \frac{\exp\left(1/N\sum_k \log\left(a_k\right)\right)}{1/N\sum_k a_k}. \tag{B.7}$$

Facteur de Crête Spectral par Bandes

Un autre descripteur relatif à la platitude est le facteur de crête spectral qui se calcule à partir du rapport de la valeur d'amplitude maximale des bandes et de la moyenne arithmétique de l'énergie du spectre. Les bandes de fréquences considérées ainsi que les variables sont les mêmes que pour le critère PAS.

$$FCSB(bf) = \frac{\max A_{k\in bf}}{\frac{1}{K}\sum_{k\in bf} A_k} \tag{B.8}$$

Pente Spectrale

La pente spectrale représente le taux de décroissance spectrale. Il est calculé par régression linéaire de l'amplitude spectrale : $PS(f) = pente.f + c$ avec

$$pente = \frac{1}{\sum_k a_k} \frac{N \sum_k f(k).a_k - \sum_k f(k) - \sum_k a_k}{N \sum_k f^2(k) - (\sum_k f(k))^2} \tag{B.9}$$

Décroissance Spectrale

Mesure la décroissance des amplitudes spectrales. Il se calcule de la façon suivante :

$$DS = \frac{1}{\sum_{k=2...K} a_k} \sum_{k=2...K} \frac{a_k - a_1}{k-1} \tag{B.10}$$

Variation Spectrale

Facteur de variation du spectre en fonction du temps. Il est calculé à partir de la corrélation croisée entre les amplitudes spectrales successives $a(t-1)$ et $a(t)$. Le flux spectral tend vers 0 quand les contenus spectraux successifs sont similaires, vers 1 quand ils sont différents.

$$VS = 1 - \frac{\sum_k a_k(t-1).a_k(t)}{\sqrt{\sum_k a_k(t-1)^2}\sqrt{\sum_k a_k(t)^2}} \tag{B.11}$$

Fréquence de coupure

Fréquence à partir de laquelle 95% de l'énergie du spectre est contenue :

$$\sum_0^{fc} a^2(f) = 0.95 \sum_0^{fe/2} a^2(f), \tag{B.12}$$

avec f_c est la fréquence de coupure et f_e la fréquence d'échantillonnage.

Flux Spectral

Mesure la variation du spectre entre des trames consécutives conformément à Scheirer & Slaney (1997) :

$$FS = \frac{\sum_k (a_k(t) - a_k(t-1))^2}{\sqrt{\sum_k a_k(t-1)^2}\sqrt{\sum_k a_k(t)^2}} \tag{B.13}$$

Modulation d'Amplitude

Caractérise les phénomènes de trémolo (entre 4 et 8 Hz) ou encore la rugosité d'un son (entre 10 et 40 Hz). Les 4 critères sont détaillés dans Martin (1999), Eronen (2001), Essid (2005) :

- Fréquence MA : fréquence du pic d'amplitude maximale.
- Amplitude MA : différence entre l'amplitude maximale et l'amplitude moyenne globale du spectre.

- Amplitude MA heuristique : différence entre l'amplitude maximale et l'amplitude moyenne sur la bande de fréquences.
- Produit MA : produit de la fréquence AM et de l'amplitude AM.

LSF (Line Spectral Frequency)

Utilisés pour représenter les coefficients de prédiction linéaires (LPC pour Linear Prediction Coefficiants). Les LSF sont très utilisés en codage de la parole car ils sont plus robustes aux bruits de quantifications que les LPC. On pourra consulter Bäckström & Magi (2006) et Schussler (1976) pour plus de détails.

OBSI (Octave band signal intensity)

Proposé par Essid (2005), ce descripteur est destiné à capturer la structure spectrale des sons instrumentaux. Un banc de 10 filtres triangulaires d'une octave (avec un recouvrement d'une demi-octave) est utilisé pour mesurer la log énergie de chaque bande.

OBSIR (Octave Band Signal Intensities Ratios)

Logarithme des rapports des OBSI entre octaves consécutives proposé par Essid (2005). Utilisé pour mesurer la différence entre des valeurs OBSI de bandes consécutives.

Coefficients de prédiction linéaire

Les coefficients de prédiction linéaire (ou LPC pour Linear Predictor Coefficients) sont très utilisés en codage de la parole. Ils permettent de représenter l'enveloppe spectrale d'un signal de façon compressée (voir Makhoul (1975)).

B.2 Descripteurs Cepstraux

Le *cepstre* se définit comme la *Transformée de Fourier* inverse du logarithme du spectre d'amplitude. Les descripteurs cepstraux suivants ont été extraits :

MFCC (Mel Frequency Cepstral Coefficients)

Permet de représenter l'enveloppe spectrale avec peu de coefficients (Rabiner & Juang (1993)). Le *Mel-cepstre* est basé sur les bandes de fréquences de Mel qui modélisent le système auditif humain et les MFCC sont les coefficients du *Mel-cepstre*. Dans ce travail, nous utilisons les 13 premiers coefficients ainsi que les dérivées temporelles de premier et second ordre (pour la classification d'objets).

Coefficients cepstraux à Q constant

Dans ce descripteur, le calcul du cepste est réalisé en tenant compte des gammes musicales occidentales tempérées (Brown (1991)). Plusieurs résolutions sont considérées : une, la moitié, un tiers et un quart d'octave. Nous utilisons également les dérivées temporelles de premier et second ordre.

B.3 Descripteurs Temporels

Taux de passage par zero (ou ZCR pour Zero Crossing Rate)

Calcul le nombre de fois que le signal change de signe Kedem (1986). Les signaux périodiques ont tendance à avoir un ZCR faible. A l'inverse, les signaux bruités ont tendance à avoir un ZCR fort. Ce descripteur est donc particulièrement utile pour distinguer ces deux types de signaux.

Moments statistiques temporels

Comme pour le spectre, les moments statistiques d'ordre 1 à 4 sont calculés sur les trames du signal. Ainsi, nous obtenons le centroïde temporel, la largeur temporelle, l'asymétrie temporelle et la platitude temporelle en remplaçant les coefficients d'amplitude spectrale par le signal.

Coefficients d'autocorrélation

Les coefficients d'autocorrélation représentent la distribution spectrale dans le domaine temporel. Ce descripteur a déjà été utilisé avec succès dans Brown (1998) pour la classification automatique d'instruments de musique et peut être calculé de la façon suivante pour un signal x :

$$AC(k) = \frac{1}{x_0^2} \sum_{n=0}^{N-k-1} x_n x_{n+k} \tag{B.14}$$

Energie

Nous utilisons également un descripteur d'énergie calculé à partir de la moyenne quadratique des trames du signal :

$$E = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x_n^2} \tag{B.15}$$

Enveloppe d'amplitude

L'enveloppe d'amplitude est obtenue par une approche s'inspirant de celle de Berthomier (1983). Un signal d'analyse y est d'abord calculé sur des fenêtres longues :

$$y(n) = x(n) + i\psi(n), \tag{B.16}$$

où $\psi(n)$ est la transformée de Hilbert du signal $x(n)$. L'enveloppe d'amplitude est par la suite obtenue par :

$$EA(n) = |y(n)| * h(n), \tag{B.17}$$

avec $h(n)$ est une demi-fenêtre de Hanning de 50ms qui permet de réaliser un filtrage passe-bas.

Moments de l'enveloppe temporelle

Les moments statistiques d'ordre 1 à 4 sont calculés à partir de l'enveloppe d'amplitude de la même manière que pour les moments spectraux et temporels.

B.4 Descripteurs Perceptifs

Loudness spécifique

La loudness correspond à la mesure de l'intensité perceptive tel qu'elle est décrite dans Moore et al. (1997). Nous calculons d'abbord la loudness spécifique qui utilise les bandes de fréquences de l'échelle de Bark (Zwicker (1977)) :

$$L(bf) = E(bf)^{0.23}, \tag{B.18}$$

avec $E(bf)$ est l'énergie du signal sur la bande de fréquence bf.

Les coefficients utilisés pour la description sont ceux de la loudness spécifique relative L_r définie comme le rapport de la loudness spécifique sur la loudness totale L_T :

$$L_r(bf) = \frac{L(bf)}{L_T}, \tag{B.19}$$

avec $L_T = \sum_{k \in bf} L(k)$. La normalisation par L_T permet d'être indépendant des conditions d'enregistrement qui peuvent varier de manière importante. De plus, nous utilisons également les dérivées temporelles de premier et de second ordre.

Acuité perceptive

L'acuité perceptive est la version perceptive du centroïde spectral. Ce descripteur, introduit par Peeters (2004), est calculé à partir de la loudness spécifique L :

$$AP = 0.11 \frac{\sum_{bf} bf.g(bf).L(bf)}{L_T}, \tag{B.20}$$

avec $g(bf)$ définie comme il suit :

$$g(bf) = \begin{cases} 1 \text{ si } bf < 15 \\ 0.066 \exp\left(0.171 bf\right) \text{ si } bf \geq 15 \end{cases} \tag{B.21}$$

Etalement perceptif

Mesure l'écart entre la loudness spécifique maximale et la loudness totale. Ce descripteur est proposé par Peeters (2004) et s'obtient comme il suit :

$$EP = \left(\frac{L_T - \max_{bf} L(bf)}{L_T} \right)^2 \tag{B.22}$$

Annexe C

Apprentissage supervisé

Dans cette annexe, nous présentons une introduction à l'apprentissage supervisé "statique" qui constitue un des fondements des méthodes utilisées dans nos travaux.

C.1 Principes

L'apprentissage supervisé est une forme d'apprentissage automatique qui permet, à partir d'échantillons étiquetés par un expert, de prédire les étiquettes de classes de nouveaux échantillons. Dans notre cas, l'information de segmentation et de regroupement des segments apportée par l'initialisation est exploitée par l'utilisateur "expert" qui sélectionne les segments caractéristiques des classes qu'il vise (section 4.3). Cette action de sélectionner des segments distincts pour chaque classe, constitue l'étiquetage expert qui permet d'appliquer les méthodes d'apprentissage supervisé. Dans notre problème, nous avons Q classes $\{C_q\}_{1\leq q\leq Q}$ et nous disposons d'échantillons (en petit nombre) pour chaque classe. En apprentissage, on oppose souvent les méthodes *génératives* aux méthodes *discriminatives*.

Dans les méthodes génératives, on cherche à estimer une densité de probabilité a posteriori $P(C_q|x)$ en utilisant les échantillons connus. On peut obtenir la *densité de probabilité conditionnelle* $p(x|C_q)$ décrivant la distribution des échantillons x de la classe C_q ainsi que la probabilité a priori $P(C_q)$ de chaque classe en utilisant les échantillons connus. On déduit l'appartenance d'un échantillon x à une classe C_{q_0} par la *règle de décision bayesienne* :

$$q_0 = \arg \max_{1\leq q\leq Q} P(C_q|x) \tag{C.1}$$

Le principe du *maximum a postériori* qui régit cette décision garantit une erreur minimale. En appliquant la *formule de Bayes* :

$$P(C_q|x) = \arg \max_{1\leq q\leq Q} \frac{P(C_q)p(x|C_q)}{p(x)}, \tag{C.2}$$

on peut exprimer q_0 en fonction de la *densité de probabilité conditionnelle* :

$$q_0 = \arg \max_{1\leq q\leq Q} P(C_q)p(x|C_q). \tag{C.3}$$

Dans la majorité des cas, l'hypothèse d'équiprobabilité des classes C_q est retenue et permet de simplifier l'équation :

$$q_0 = \arg \max_{1 \leq q \leq Q} p(x|C_q) \tag{C.4}$$

Dans la section suivante, nous présentons les *Machines à Vecteurs Supports* qui sont représentatives de l'approche discriminative.

C.2 Machines à Vecteurs Supports

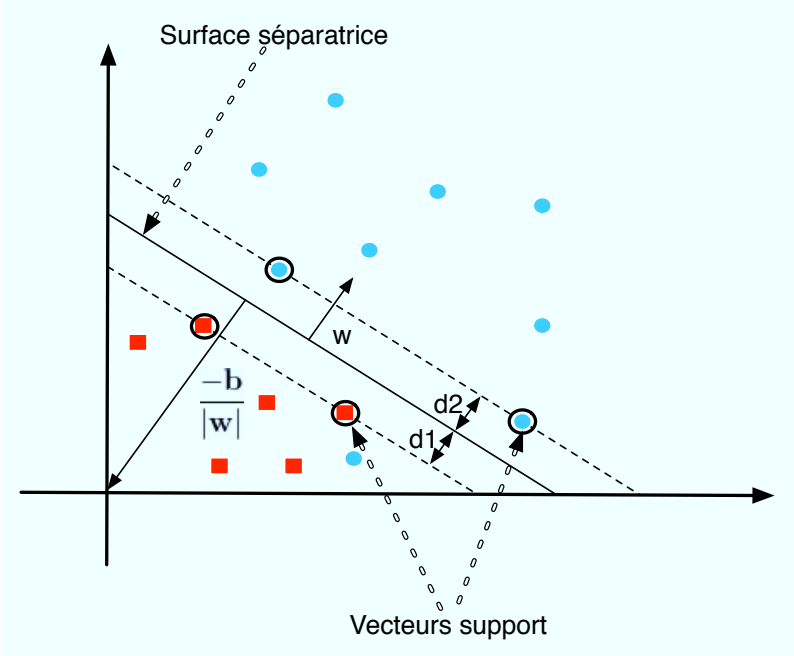

Fig. C.1 – Un cas simple de SVM pour des données presque séparables linéairement

Les *Machines à Vecteurs Supports* (SVM) sont basées sur le principe de la maximisation de la marge introduit par Vapnik et Lerner. Il existe plusieurs types d'implémentation des SVMs dans la littérature. Dans ce travail, nous utilisons la version *C-SVC* (C-Support Vector Classification) présentée dans Boser et al. (1992), Cortes & Vapnik (1995). Les SVMs permettent de choisir une surface séparatrice entre les classes en respectant le principe de *minimisation du risque structurel* ce qui revient à maximiser la marge. La figure C.1 illustre les différentes variables dans un cas simple. Les vecteurs "entourés" sont des *vecteurs supports* : on désigne ainsi les vecteurs les plus proches de l'hyperplan séparateur. Les distances $d1$ et $d2$ sont égales, cette distance est la *marge* des SVM. Nous avons L vecteurs d'entrainement x_i à D attributs auxquels nous associons une étiquette $y_i = \pm 1$. L'hyperplan séparateur peut être décrit par :

$$w^T x + b = 0, \tag{C.5}$$

où w est la normale à l'hyperplan et $\frac{b}{||w||}$ est la distance orthogonale de l'hyperplan à l'origine. En observant la figure C.1, nous déduisons que nos données d'entrainement peuvent être décrites par les équations C.6.

$$\begin{cases} w^T x_i + b \geq +1 \text{ si } y_i = +1 \\ w^T x_i + b \leq +1 \text{ si } y_i = -1 \end{cases} \tag{C.6}$$

Les deux équations peuvent être combinées comme il suit :

$$y_i(w^T x_i + b) - 1 \geq 0 \ \forall i \tag{C.7}$$

Pour gérer le cas ou les données d'entrainement ne sont pas complètement séparables linéairement, on introduit une *variable d'écart* positive ξ_i avec $i = 1...L$. Ainsi l'équation précédente devient :

$$y_i(w^T x_i + b) - 1 + \xi_i \geq 0 \text{ avec } \xi_i \geq 0 \ \forall i \tag{C.8}$$

Finalement, nous cherchons à résoudre le problème d'optimisation suivant :

$$\begin{cases} \min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{L} \xi_i \\ \text{sous les contraintes } y_i(w^T x_i + b) - 1 + \xi_i \geq 0, \ \xi_i \geq 0 \ \forall i \end{cases} \tag{C.9}$$

Avec $C > 0$ est le *facteur d'erreur* qui permet de contrôler le compromis entre le nombre d'exemples mal classés et la largeur de la marge. Le lecteur pourra se référer à Fletcher (2008) qui explique de façon très didactique les détails de calcul pour la résolution du problème d'optimisation des SVM.

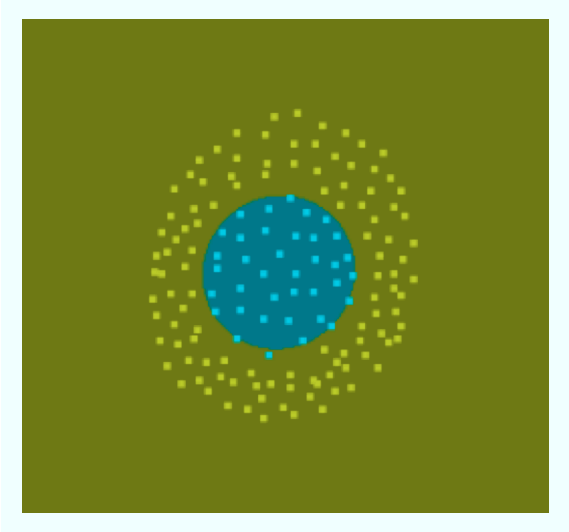

FIG. C.2 – Exemple de données non linéairement séparables avec la surface de décision estimée par un algorithme SVM

Les fondements théoriques que nous venons de présenter s'appliquent uniquement aux données linéairement séparables dans l'espace de description. Or, ce cas idéal est rarement rencontré dans la nature où les distributions peuvent avoir des formes complexes dans l'espace multidimensionnel comme le montre la figure C.2. Pour cette raison, une fonction noyau est utilisée afin d'obtenir des surfaces de décision non linéaires. Le principe d'une fonction noyau est de transformer les données exprimées dans l'espace des attributs à D dimensions dans un espace de dimension plus grande voir infinie. En procédant ainsi, il est

possible de trouver une séparatrice linéaire dans le nouvel espace. Les noyaux rencontrés le plus souvent dans la littérature sont les suivants :

- Le noyau *linéaire* : $k(x, y) = x.y$
- Le noyau *polynômial* de degré δ : $k(x, y) = (x.y)^{\delta}$
- Le noyau *radial exponentiel* : $k(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}}$

Dans ce travail, nous utilisons le noyau *radial exponentiel* (encore appelé noyau gaussien) qui possède des bonnes propriétés de généralisation.

La sortie des SVM est binaire mais il existe des méthodes permettant d'estimer des probabilités a postériori. Cette estimation constitue une information plus complète qu'une étiquette binaire : elle donne un degré de confiance pour l'appartenance à une classe. Pour estimer la probabilité, l'hypothèse de départ est que plus un exemple est éloigné de la surface de séparation, plus l'estimation d'appartenance à la classe considérée est fiable. Dans Platt (1999), l'auteur propose d'utiliser une forme sigmoïdale pour modéliser la probabilité de la classe positive en partant de l'hypothèse empirique que les densités de probabilités conditionnelles sont exponentielles dans la marge. En supposant que f est la fonction de décision, la probabilité conditionnelle d'appartenance à la classe positive s'exprime ainsi :

$$P(y = 1|f(x)) = \frac{1}{1 + \exp(Af(x) + B)} \tag{C.10}$$

On pourra consulter Platt (1999) pour plus de détails.

C.3 Fusion des décisions de plusieurs classifieurs binaires

Ces méthodes servent à prendre une décision pour un problème multiclasse ($Q > 2$) en fusionnant les sorties de classifieurs bi-classes ($Q = 2$). L'étape de fusion est très utile dans le cas des SVM qui sont, par essence, des classifieurs bi-classes. Dans la littérature, on oppose souvent deux stratégies d'apprentissage différentes qui conduisent à la prise de décision finale : l'approche *Un Contre Un* (ou OVO pour *One Versus One*) et l'approche *Un Contre Tous* (ou OVA pour *One Versus All*).

Stratégie OVO

Cette première stratégie décompose un problème multiclasse en un ensemble de sous-problèmes bi-classes. Ainsi, des classifieurs sont construits pour tous les couples possibles de classes distinctes : pour un problème à Q classes, on dénombre $C_2^Q = \frac{Q(Q-1)}{2}$ classifieurs construits. Pour un nouvel échantillon x, on réalise le test de classification de x avec les C_2^Q classifieurs bi-classes construits avant de recombiner les sorties des classifieurs pour la prise de décision finale. Il existe plusieurs approches de recombinaisons de classifieurs. Certaines méthodes utilisent des sorties de classifieurs de type *hard output*. Dans ce cas, les sorties sont binaires : l'échantillon est affecté à une des deux classes du problème bi-classe sans aucune précision sur le degré de confiance de la prédiction. Une méthode courante de fusion avec des sorties *hard output* consiste à effectuer un vote majoritaire : l'échantillon sera affecté à la classe qui récolte le plus grand nombre de voix. Une autre méthode bien

connue proposée par Hastie & Tibshirani (1998) consiste à estimer les probabilités a posteriori $P(C_q|x)$ pour chaque classe. La prise de décision est par la suite réalisée en appliquant la *règle de décision bayesienne* (équation C.1). La stratégie OVO permet d'obtenir de bons résultats mais en contrepartie elle peut également devenir un peu lourde en raison de sa complexité et du nombre de calculs intermédiaires qu'elle implique.

Stratégie OVA

La stratégie OVA est plus simple que la précédente. Dans cette approche, Q classifieurs sont construits, soit un classifieur par classe. Pour un classifieur q donné (avec $1 \leq q \leq Q$), on considère deux types d'échantillons : les échantillons qui appartiennent à la classe q nommés *échantillons positifs* et ceux qui ne lui appartiennent pas nommés *échantillons négatifs*. Pour chaque classifieur, dans l'approche discriminative, on cherche à trouver la surface qui sépare les *échantillons positifs* des *échantillons négatifs*. Comme nous l'avons vu, il est possible de probabiliser la sortie d'un classifieur SVM. Ainsi, on peut obtenir une probabilité d'appartenance P_q pour une classe q donnée. La décision finale, dans un contexte *multiclasse* classique, sera prise en cherchant la classe qui maximise la probabilité d'appartenance P_q :

$$q_0 = \arg \max_{1 \leq q \leq Q} P_q(x) \tag{C.11}$$

L'approche OVA est moins utilisée dans la littérature que la stratégie OVO. Cependant, une comparaison avec d'autres approches plus complexe effectuée dans Rifkin & Klautau (2004) montre que l'approche OVA permet d'obtenir de bons résultats malgré sa simplicité.

On peut noter qu'il existe des méthodes qui tentent de reformuler les SVM en une approche multi-classe (voir Rifkin & Klautau (2004) pour un aperçu), mais les moyens mis en oeuvre sont la plupart du temps gourmands en temps de calcul.

Table des figures

Index

Bibliographie

Alonso, M., Richard, G. & David, B. (2005), 'Extracting note onsets from musical recordings.', *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference* .

Bäckström, T. & Magi, C. (2006), 'Properties of line spectrum pair polynomials : a review', *Signal Process.* **86**(11), 3286–3298.

Bartsch, M. A. & Wakefield, G. H. (2001), To catch a chorus : using chroma-based representations for audio thumbnailing, *in* 'Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the', IEEE, pp. 15–18.

Bartsch, M. A. & Wakefield, G. H. (2005), 'Audio thumbnailing of popular music using chroma-based representations', *IEEE Transactions on Multimedia* **7**(1), 96– 104.

Berthomier, C. (1983), 'Instantaneous frequency and energy distribution of a signal', *Signal Processing* **5**(1), 31–45.

Boser, B., Guyon, I. M. & Vapnik, V. (1992), 'A training algorithm for optimal margin classifiers', *Proceedings of the fifth annual workshop on Computational learning theory COLT 92* p. 144–152.

Bossis, B. (2006), 'The analysis of electroacoustic music : From sources to invariants', *Organised Sound* **11**(02), 101–112.

Brown, J. (1998), Musical instrument identification using autocorrelation coefficients, *in* 'Symposium on Musical Acoustics'.

Brown, J. C. (1991), 'Calculation of a constant q spectral transform', *Acoustical Society of America Journal* **89**, 425–434.

Busoni, F. (1911), *Sketch Of A New Esthetic Of Music*, Schirmer, New York.

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

Cettolo, M. & Vescovi, M. (2003), Efficient audio segmentation algorithms based on the BIC, *in* '2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)', Vol. 6, IEEE, pp. VI– 537–40 vol.6.

Chai, W. (2003), Structural analysis of musical signals via pattern matching, *in* '2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)', Vol. 5, IEEE, pp. V– 549–52 vol.5.

Chai, W. (2005), Automated Analysis of Musical Structure, PhD thesis, Massachusetts Institute of Technology.

Chai, W. & Vercoe, B. (2003), Structural analysis of musical signals for indexing and thumbnailing, *in* 'JCDL '03 : Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries', IEEE Computer Society, Washington, DC, USA, p. 27–34.

Chang, C. & Lin, C. (2011), 'LIBSVM : a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology* **2**(3), 27 :1–27 :27. Software available at http ://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chen, G., Wang, T. & Herrera, P. (2008), A novel music retrieval system with relevance feedback, *in* 'Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control', IEEE Computer Society, Washington, DC, USA, p. 158–.

Cohn, D. A., Ghahramani, Z. & Jordan, M. I. (1996), 'Active learning with statistical models'. Journal of Artificial Intelligence Research, Vol 4, (1996), 129-145.

Cooper, M. (2002), 'Automatic music summarization via similarity analysis', *Proc. IRCAM* .

Cooper, M. & Foote, J. (2003), Summarizing popular music via structural similarity analysis, *in* 'Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.', IEEE, pp. 127– 130.

Cortes, C. & Vapnik, V. (1995), 'Support-Vector networks', *Machine Learning* **20**(3), 273–297.

Couprie, P. (2004), 'Graphical representation : An analytical and publication tool for electroacoustic music', *Organised Sound* **9**(01), 109–113.

Couprie, P. (2006), '(Re) presenting electroacoustic music', *Organised Sound* **11**(02), 119–123.

Couprie, P. (2008), IAnalyse : un logiciel d'aide à l'analyse musicale, *in* 'Journée de l'Informatique Musicale'.

Crucianu, M., Ferecatu, M. & Boujemaa, N. (2004), Relevance feedback for image retrieval : a short survey, *in* 'State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction including Datamodels and Languages (DELOS2 Report)'.

Desainte-Catherine, M. & Marchand, S. (1999), Structured additive synthesis : Towards a model of sound timbre and electroacoustic music forms, *in* 'Proceedings of the International Computer Music Conference'.

Desobry, F., Davy, M. & Doncarli, C. (2005), 'An online kernel change detection algorithm', *IEEE Transactions on Signal Processing* **53**(8), 2961– 2974.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), *Pattern classification*, Wiley.

Eggink, J. & Brown, G. J. (2003), Application of missing feature theory to the recognition of musical instruments in polyphonic audio, *in* 'Proc. of International Conference on Music Information Retrieval'.

Eronen, A. (2001), Automatic Musical Instrument Recognition, Mémoire de master, Tampere University of Technology.

Essid, S. (2005), Classification automatique des signaux audio-frequences : reconnaissance des instruments de musique, PhD thesis, UPMC.

Essid, S., Richard, G. & David, B. (2006), 'Instrument recognition in polyphonic music based on automatic taxonomies', *IEEE Transactions on Audio, Speech, and Language Processing* **14**(1), 68– 80.

Fastl, H. & Zwicker, E. (2007), *Psychoacoustics : facts and models*, Springer.

Fletcher, T. (2008), Support Vector Machines Explained, Tutorial paper, University College London.

Foote, J. (2000), Automatic audio segmentation using a measure of audio novelty, *in* 'Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on', Vol. 1, p. 452–455 vol.1.

Formosa, M., Fremiot, T., Delalande, F., Gobin, P., Malbosc, P., Mandelbrojt, J. & Pedler, E. (1996), *Les Unités sémiotiques temporelles : éléments nouveaux d'analyse musicale*, Laboratoire musique et informatique de Marseille.

Gayou, E. (2006), 'Analysing and transcribing electroacoustic music : The experience of the portraits polychromes of GRM', *Organised Sound* **11**(02), 125–129.

Geslin, Y. & Lefevre, A. (2004), Sound and musical representation : the acousmographe software, *in* 'International Computer Music Conference', Miami, USA.

Gillet, O. & Richard, G. (2004), Automatic transcription of drum loops, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)', Vol. 4, IEEE, pp. iv–269– iv–272 vol.4.

Godsmark, D. & Brown, G. J. (1999), 'A blackboard architecture for computational auditory scene analysis', *Speech Commun.* **27**(3-4), 351–366.

Goeau, H. (2009), Structuration de collections d'images par apprentissage actif crédibiliste, PhD thesis, Université Joseph Fourier de Grenoble.

Goeau, H., Buisson, O. & Viaud, M. L. (2008), Image collection structuring based on evidential active learner, *in* 'International Workshop on Content-Based Multimedia Indexing, 2008. CBMI 2008', IEEE, pp. 388–395.

Goto, M. (2003), SmartMusicKIOSK : music listening station with chorus-search function, *in* 'UIST '03 : Proceedings of the 16th annual ACM symposium on User interface software and technology', ACM, New York, NY, USA, p. 31–40.

Goto, M., Hashiguchi, H., Nishimura, T. & Oka, R. (2002), RWC music database : Popular, classical, and jazz music databases, *in* 'Proc. of International Conference on Music Information Retrieval', p. pp.287–288.

Gulluni, S., Buisson, O., Essid, S. & Richard, G. (2009), Interactive segmentation of Electro-Acoustic music, *in* '2nd International Workshop on Machine Learning and Music'.

Gulluni, S., Essid, S., Buisson, O. & Richard, G. (2011*a*), Interactive classification of sound objects for polyphonic Electro-Acoustic music annotation, *in* 'Audio Engineering Society Conference : 42nd International'.

Gulluni, S., Essid, S., Buisson, O. & Richard, G. (2011*b*), An interactive system for electro-acoustic music analysis, *in* 'Proc. of International Conference on Music Information Retrieval'.

Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A. & Cappe, O. (2009), A regularized kernel-based approach to unsupervised audio segmentation, *in* 'Acoustics, Speech, and Signal Processing, IEEE International Conference on', Vol. 0, IEEE Computer Society, Los Alamitos, CA, USA, pp. 1665–1668.

Hastie, T. & Tibshirani, R. (1998), Classification by pairwise coupling, *in* 'Proceedings of the 1997 conference on Advances in neural information processing systems 10', NIPS '97, MIT Press, p. 507–513.

Hennequin, R., David, B. & Badeau, R. (2011), Score informed audio source separation using a parametric model of non-negative spectrogram, *in* '2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 45–48.

Hist, D. (2004), An analytical methodology for acousmatic music, *in* 'Proc. of International Conference on Music Information Retrieval'.

Hist, D. (2005), Developing an interactive study score for the analysis of electro-acoustic music, *in* 'Australasian Computer Music Conference'.

Hoashi, K., Matsumoto, K. & Inoue, N. (2003), Personalization of user profiles for content-based music retrieval based on relevance feedback, *in* 'Proceedings of the eleventh ACM international conference on Multimedia', ACM, New York, NY, USA, p. 110–119.

Hong, P., Tian, Q. & Huang, T. S. (2000), Incorporate support vector machines to content-based image retrieval with relevance feedback, *in* 'International Conference on Image Processing', Vol. 3, IEEE, pp. 750–753 vol.3.

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

Joshi, A. J., Porikli, F. & Papanikolopoulos, N. (2009), 'Multi-class active learning for image classification', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .

Kane, B. (2007), 'L'Objet sonore maintenant : Pierre schaeffer, sound objects and the phenomenological reduction', *Organised Sound* **12**(01), 15–24.

Kedem, B. (1986), 'Spectral analysis and discrimination by zero-crossings', *Proceedings of the IEEE* **74**(11), 1477– 1493.

Kinoshita, T., Sakai, S. & Tanaka, H. (1999), Musical sound source identification based on frequency component adaptation, *in* 'Proc. IJCAI Worshop on CASA'.

Kitahara, T., Goto, M., Komatani, K., Ogata, T. & Okuno, H. G. (2007), 'Instrument identification in polyphonic music : Feature weighting to minimize influence of sound overlaps', *EURASIP Journal on Advances in Signal Processing* **2007**, 1–16.

Kurtag, G., Di Santo, J., Desainte-Catherine, M. & Guillem, P. (2007), Pédagogie de l'électroacoustique du geste musical à la composition assistée par ordinateur, *in* 'Proceedings of the Journées de l'Informatique Musicale (Jim07)'.

Levy, M., Sandier, M. & Casey, M. (2006), Extraction of High-Level musical structure from audio data and its application to thumbnail generation, *in* 'Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 5, p. V.

Li, X., Wang, L. & Sung, E. (2004), Multilabel SVM active learning for image classification, *in* '2004 International Conference on Image Processing, 2004. ICIP '04', Vol. 4, IEEE, pp. 2207– 2210 Vol. 4.

List, N. & Simon, H. U. (2005), General polynomial time decomposition algorithms, *in* P. Auer & R. Meir, eds, 'Learning Theory', Vol. 3559, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 308–322.

Little, D. & Pardo, B. (2008), Learning musical instruments from mixtures of audio with weak labels, *in* 'Proc. of International Conference on Music Information Retrieval'.

Logan, B. & Chu, S. (2000), Music summarization using key phrases, *in* '2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings', Vol. 2, IEEE, pp. II749–II752 vol.2.

Lu, L., Wang, M. & Zhang, H. (2004), Repeating pattern discovery and structure analysis from acoustic music data, *in* 'Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval', MIR '04, ACM, New York, NY, USA, p. 275–282.

Lukashevich, H., Abeßer, J., Dittmar, C. & Grossmann, H. (2009), From Multi-Labeling to Multi-Domain-Labeling : a novel Two-Dimensional approach to music genre classification, *in* 'Proc. of International Conference on Music Information Retrieval'.

Maddage, N. C. (2006), 'Automatic structure detection for popular music', *IEEE Multimedia* **13**(1), 65– 77.

Maddage, N. C., Xu, C., Kankanhalli, M. S. & Shao, X. (2004), Content-based music structure analysis with applications to music semantics understanding, *in* 'Proceedings of the 12th annual ACM international conference on Multimedia', MULTIMEDIA '04, ACM, New York, NY, USA, p. 112–119.

Makhoul, J. (1975), 'Linear prediction : A tutorial review', *Proceedings of the IEEE* **63**(4), 561–580.

Mandel, M., Poliner, G. & Ellis, D. (2006), 'Support vector machine active learning for music retrieval', *Multimedia Systems* **12**(1), 3–13.

Manning, P. D. (2004), *Electronic and computer music.*, Oxford University Press, New York.

Martin, K. D. (1999), 'Sound-Source recognition : A theory and computational model', *PhD thesis, MIT* .

Mathieu, B., Essid, S., Fillon, T., Prado, J. & Richard, G. (2010), 'YAAFE, an easy to use and efficient audio feature extraction software', *Proc. of International Conference on Music Information Retrieval* .

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. & Krimphoff, J. (1995), 'Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes', *Psychological Research* **58**, 177–192.

Molino, J. (2009), *Le singe musicien : essais de sémiologie et d'anthropologie de la musique*, Actes Sud.

Moore, B., Glasberg, B. & Baer, T. (1997), 'A model for the prediction of thresholds, loudness, and partial loudness', *J. Audio Eng. Soc* **45**(4), 224–240.

Ning, J., Zhang, L., Zhang, D. & Wu, C. (2010), 'Interactive image segmentation by maximal similarity based region merging', *Pattern Recogn.* **43**(2), 445–456.

Nucibella, F., Porcelluzzi, S. & Zattra, L. (2005), Computer music analysis via a multidisciplinary approach, *in* 'Sound and Music Computing'.

Park, T. H., Li, Z. & Wu, W. (2009), EASY does it : The Electro-Acoustic music analysis toolbox, *in* 'Proc. of International Conference on Music Information Retrieval'.

Peeters, G. (2004), A large set of audio features for sound description (similarity and classification) in the CUIDADO project, Tech. rep., IRCAM.

Peeters, G., Burthe, A. L. & Rodet, X. (2002), Toward automatic music audio summary generation from signal analysis, *in* 'Proc. of International Conference on Music Information Retrieval', p. 94–100.

Peeters, G. & Deruty, E. (2008), 'Automatic morphological description of sounds', *The Journal of the Acoustical Society of America* **123**, 3801.

Platt, J. (1999), Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, *in* 'Advances in large margin classifiers', p. 61–74.

Price, B., Morse, B. & Cohen, S. (2009), LIVEcut : learning-based interactive video segmentation by evaluation of multiple propagated cues., *in* 'Proceedings of the IEEE International Conference on Computer Vision (ICCV)'.

Puig, V., Guédy, F., Fingerhut, M., Serrière, F., Bresson, J. & Zeller, O. (2005), Musique lab 2 : A three level approach for music education at school, *in* 'Proceedings of the International Computer Music Conference', Spain.

Qi, G. J., Hua, X., Rui, Y., Tang, J. & Zhang, H. (2009), 'Two-Dimensional multilabel active learning with an efficient online adaptation model for image classification', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(10), 1880–1897.

Rabiner, L. & Juang, B. (1993), *Fundamentals of Speech Recognition*, Prentice Hall.

Reynolds, D., Kenny, P. & Castaldo, F. (2009), A study of new approaches to speaker diarization, *in* 'proc. of INTERSPEECH-2009', pp. 1047–1050.

Ricard, J. & Herrera, P. (2004), Morphological sound description : Computational model and usability evaluation, *in* 'Audio Engineering Society Convention 116'.

Rifkin, R. & Klautau, A. (2004), 'In defense of One-Vs-All classification', *The Journal Machine Learning Research* **5**, 101–141.

Rijsbergen, C. (1979), *Information retrieval*, Butterwoth-Heinmann, 2nd edition, London.

Rocchio, J. & Salton, G. (1971), Relevance feedback in information retrieval, *in* 'The SMART Retrieval System : Experiments in Automatic Document Processing', Prentice-Hall, Englewood Cliffs NJ, pp. 313–323.

Russolo, L. (1913), *The art of noise*, Something Else Press.

Salton, G. (1968), *Automatic Information Organization and Retrieval*, McGraw-Hill.

Savage, J. & Challis, M. (2002), 'Electroacoustic composition : Practical models of composition with new technologies', *Journal of the Sonic Arts Network* .

Schaeffer, P. (1952), *A la recherche d'une musique concrète*, Éditions du Seuil, Paris.

Schaeffer, P. (1966), *Traité des objets musicaux*, Éditions du Seuil.

Scheirer, E. & Slaney, M. (1997), Construction and evaluation of a robust multifeature speech/music discriminator, *in* ', 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97', Vol. 2, IEEE, pp. 1331–1334 vol.2.

Schussler, H. (1976), 'A stability theorem for discrete systems', *Acoustics, Speech and Signal Processing, IEEE Transactions* **24**(1), 87–89.

Sedes, A., Courribet, B. & Thiébaut, J. (2004), Visualization of sound as a control interface, *in* 'Proc. of the 7th Int. Conference on Digital Audio Effects (DAFX)', Naples, Italy.

Settles, B. (2010), Active learning literature survey, Technical report, University of Wisconsin–Madison.

Shan, M., Chiang, M. & Kuo, F. (2008), 'Relevance feedback for category search in music retrieval based on semantic concept learning', *Multimedia Tools Appl.* **39**, 243–262.

Simoni, M., Rozell, C., Meek, C. & Wakefield, G. (2000), A theoretical framework for electro-acoustic music., *in* 'International Computer Music Conference'.

Singh, M., Curran, E. & Cunningham, P. (2009), Active learning for Multi-Label image annotation, Technical report, University College of Dublin.

Smets, P. (2005), 'Decision making in the TBM : the necessity of the pignistic transformation', *International Journal of Approximate Reasoning* **38**(2), 133–147.

Tong, S. & Chang, E. (2001), Support vector machine active learning for image retrieval, *in* 'Proceedings of the ninth ACM international conference on Multimedia', ACM, New York, NY, USA, p. 107–118.

Tranter, S. E. & Reynolds, D. A. (2006), 'An overview of automatic speaker diarization systems', *IEEE Transactions on Audio, Speech, and Language Processing* **14**(5), 1557–1565.

Trohidis, K., Tsoumakas, G., Kalliris, G. & Vlahavas, I. (2008), Multilabel classification of music into emotions, *in* 'Proc. of International Conference on Music Information Retrieval', Philadephia, PA, USA.

Tsoumakas, G. & Katakis, I. (2007), 'Multi-label classification : An overview', *INT J DATA WAREHOUSING AND MINING* **2007**, 1—13.

Tsoumakas, G. & Vlahavas, I. (2007), Random k-Labelsets : an ensemble method for multilabel classification, *in* J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič & A. Skowron, eds, 'Machine Learning : ECML 2007', Vol. 4701, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 406–417.

Van Steelant, D., De Baets, B., De Meyer, H., Leman, M., Martens, J. P., Clarisse, L. & Lesaffre, M. (2002), 'Discovering structure and repetition in musical audio', *IN PROCEEDINGS OF EUROFUSE WORKSHOP* .

Viterbi, A. (1967), 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Transactions on Information Theory* **13**(2), 260–269.

Wieczorkowska, A., Synak, P. & Raś, Z. W. (2006), Multi-Label classification of emotions in music, *in* M. A. Kłopotek, S. T. Wierzchoń & K. Trojanowski, eds, 'Intelligent Information Processing and Web Mining', Vol. 35, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 307–315.

Wu, Y., Kozintsev, I., Bouguet, J.-y. & Dulong, C. (2006), Sampling strategies for active learning in personal photo retrieval, *in* 'Multimedia and Expo, IEEE International Conference on', Vol. 0, IEEE Computer Society, Los Alamitos, CA, USA, pp. 529–532.

Zhou, X. S. & Huang, T. S. (2003), 'Relevance feedback in image retrieval : A comprehensive review', *Multimedia Systems* **8**(6), 536–544.

Zwicker, E. (1977), 'Procedure for calculating loudness of temporally variable sounds', *The Journal of the Acoustical Society of America* **62**, 675.