AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur : ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite de ce travail expose à des poursuites pénales.

Contact : portail-publi@ut-capitole.fr

LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n°92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992
http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 1 Capitole (UT1 Capitole)*

Présentée et soutenue le *10 juillet 2014* **par :**

MAXIME LE COZ

Spectre de rythme et sources multiples

Au cœur des contenus ethnomusicologiques et sonores

JURY

P. JOLY	Président du Jury	IRIT
C. BARRAS	Rapporteur	LIMSI
G. PEETERS	Rapporteur	IRCAM
M. DESAINTE-CATHERINE	Examinateur	LABRI
G. PELLERIN	Examinateur	Parisson
R. ANDRÉ-OBRECHT	Directeur	IRIT
J. PINQUIER	Co-Directeur	IRIT

École doctorale et spécialité :

MITT : Image, Information, Hypermedia

Unité de Recherche :

Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur(s) de Thèse :

Régine André-Obrecht et *Julien Pinquier*

Rapporteurs :

Geoffroy Peeters et *Claude Barras*

Remerciements

Je tiens à débuter cette thèse par des remerciements à ceux qui ont permis, de près ou de loin, à ce que ce manuscrit, mais surtout tous les travaux qui y sont décrits existent.

À commencer par mes deux directeurs : Régine André-Obrecht et Julien Pinquier qui m'ont beaucoup appris et qui ont su me faire confiance dès le Master. Merci à Régine de m'avoir poussé à croire en mes idées et leurs applications. Merci à Julien pour son sérieux autant que ses blagues qui m'ont beaucoup fait avancer... dans la bonne humeur ! C'est grâce à eux que je suis maintenant fier du travail accompli et des nombreuses perspectives qui s'ouvrent pour continuer à faire évoluer mes idées. > oui, il me la faut pr que je puisse vous délivrer l'attestation de reussite. merci Virginie MANGION Merci aux membres de mon jury : Madame Myriam Desainte-Catherine et Messieurs Geoffroy Peeters, Claude Barras, Philippe Joly et Guillaume Pellerin pour avoir pris le temps de juger de mon travail et pour leurs réflexions très intéressantes qui m'ont permis d'envisager mes travaux sous de nouveaux angles.

Merci à tous les membres de l'équipe SAMoVA pour leurs conseils et leurs sympathie et en particulier à ceux qui ont partagé mon bureau : Benjamin Bigot, Hélène Lachambre et Patrice Guyot pour les nombreuses discutions et débats !

Merci encore aux lanceurs de frisbee qui m'ont permis de m'évader et de beaucoup apprendre dans d'autres domaines : TomTom, Terry, Yves-Mat', CedCed, Marie-Prune, Margot, Raphi et tant d'autres !

Merci à Qiong-Yao pour son amitié, sa bonne humeur et ses sourires.

Merci à mes parents qui m'ont toujours supporté et fait confiance.

Et enfin, un très grand merci à celle qui m'a porté et supporté. sans qui cette période aurait été beaucoup plus difficile : Merci beaucoup *Mouts*...

Table des matières

Chapitre 1

Introduction

1.1 Contexte

Avec l'évolution de l'ère du numérique, de plus en plus de données audio, parole comme musique sont générées. En plus des médias traditionnels de diffusion que sont la radio et la télévision, nous sommes en présence d'une explosion des contenus web qui rassemblent des émissions de radio ou TV ainsi que les enregistrements musicaux professionnels comme amateurs. L'indexation est un élément crucial pour l'utilisation des données.

Un flux numérique sans indexation est comparable à un livre sans table des matières ; ses informations sont certes préservées, mais elles ne sont pas réellement accessibles et quiconque en cherche une particulière devra parcourir l'intégralité de l'enregistrement à sa recherche. Cette contrainte est à mettre en perspective avec l'importance de l'information à extraire. Le fait que la recherche de souvenirs pour des particuliers soit fastidieuse peut être perçu comme un problème mineur, mais lorsqu'il s'agit de recherche dans des archives historiques, la problématique d'accès à la mémoire d'une société entière revêt une autre importance.

Dans cette démarche de conservation de la mémoire, nous pouvons, par exemple, penser aux archives de Radio France qui regroupent toutes les émissions radiophoniques du groupe depuis les débuts des années 1920. Dans ce même esprit, nous pouvons citer également les archives de l'**Institut National Audiovisuel**, composées de plus de **5 millions d'heures** d'enregistrements télévisuels et radiophoniques. À ces sources grand public, il faut ajouter les corpus de nombreuses études réalisées dans les laboratoires de sciences humaines. Parmi ceux-ci nous pouvons mentionner le corpus du projet ANR DIADEMS [1] qui vise à regrouper la version numérisée de l'ensemble des données ethnomusicologiques du Laboratoire d'Ethnologie et de Sociologie Comparative (**LESC**). Cette dynamique de

1. `http://diadems.telemeta.org/`

numérisation des données d'intérêt historique et scientifique est devenue un défi national avec la mise en place de la **T**rès **G**rande **I**nfrastructu**R**e Huma-Num [2] (Accès unifié aux documents numériques des sciences humaines et sociales). Cet équipement vise à regrouper les documents des domaines de recherche des sciences humaines et sociales mais insiste également sur l'accès à ceux-ci, notamment via l'indexation. Fournir une navigation à de telles archives c'est permettre un accès au public et aux chercheurs, à un pan entier de l'Histoire.

Le besoin d'outils automatiques pour parvenir à une indexation et une structuration de données de plus en plus volumineuses, offre au monde de la recherche de nombreux défis à résoudre, autant sur le plan théorique qu'applicatif. Il s'agit de comprendre avant de faire apprendre à la machine.

C'est ainsi que dans le domaine de l'indexation automatique, différentes approches de la reconnaissance des formes ou du traitement du signal ont été largement mises à contribution. La tendance a été de proposer, pour chaque caractéristique recherchée dans le contenu, une méthode spécifique. Qu'il s'agisse de localiser ou d'identifier une classe parmi plusieurs, l'information retournée est plus ou moins précise en fonction des *a priori* et de la robustesse de la méthode choisie. L'intérêt de la construction d'un système complet d'indexation est, par l'ordonnancement de méthodes spécialisées, de retourner une hiérarchie d'information de plus en plus précise en corrélant les informations des méthodes entre elles et de fournir ainsi une information de plus en plus porteuse de sens. Cette volonté de créer une chaîne d'outils de plus en plus adaptés à un problème renforce également la volonté de créer des outils génériques et paramétrables qui intègrent une information contextuelle obtenues par les étapes précédentes et adaptent leur mise en œuvre.

Cette stratégie de recherche est celle de l'équipe SAMoVA (**S**tructuration, **A**nalyse et **Mo**délisation de documents **V**idéo et **A**udio). Spécialisée dans l'analyse de documents audio et vidéo, l'équipe a développé des outils d'indexation à différents niveaux conceptuels. Historiquement dédiés à l'analyse des contenus de parole, les travaux se sont étendus à l'analyse de la voix chantée ; l'étude de la polyphonie a ouvert la porte vers l'analyse des contenus musicaux. Ces travaux plus récents illustrent la volonté de l'équipe de proposer des outils se spécialisant progressivement et reposant sur les résultats d'autres travaux comme la détection robuste Parole/Musique/Bruit ; cette démarche permet de tenir compte du type de contenu pour réaliser ou non certains traitements ou les adapter. Une revue détaillée de l'enchaînement des différentes méthodes sera présentée dans le chapitre 1.

La discrimination des zones de parole et de musique, conduit à des zones « homogènes », mais néanmoins, pour nombre d'entre elles, encore difficiles à exploiter.

2. `http://www.huma-num.fr/`

S'il est clair qu'une zone diagnostiquée à la fois « parole » et « musique » implique des défis en termes d'exploitation, des zones d'apparence moins ambiguë sont à la source de défis semblables :

- Sans prétendre être exhaustifs, nous pensons en musique aux zones de polyphonies où plusieurs sources jouent simultanément et au cas extrême qu'est le chœur à l'unisson où les différents chanteurs ont la volonté de produire le même son en même temps. Ce problème se rapproche en parole du problème de la détection de zones de parole superposée, où plusieurs locuteurs parlent en même temps. Dans ces deux problèmes sur des contenus pourtant différents nous pouvons trouver un point commun : il s'agit de détecter la présence simultanée de plusieurs sources harmoniques. Nous allons utiliser cette propriété sur les deux types de contenus pour localiser ces zones d'intérêt.
- La détection de zones de paroles particulières où la voix est plus posée et le rythme maîtrisé présente également un intérêt. Ces zones intermédiaires entre voix parlée et chantée témoignent d'un contenu particulier, mais sont aussi sources de difficultés pour des traitements de type transcription et il convient de les localiser. Leur caractérisation repose principalement sur le rythme et un parallèle peut être établi avec l'analyse du tempo en musique. En parole comme en musique l'analyse de la régularité des attaques de mots ou de notes joue en rôle important dans l'estimation de l'existence ou non d'un style et de sa caractérisation, ce que nous nous sommes proposés d'approfondir.

De manière générale, nous nous sommes efforcés de proposer des approches théoriques génériques : applicables à la fois sur la parole et la musique, afin d'extraire au mieux des informations du signal qui nous ont semblé porteuses de sens.

1.2 Problématique

D'un point de vue scientifique, les deux problèmes soulevés au paragraphe précédent nous amènent à étudier deux sujets avec deux objectifs applicatifs associés.

Le premier consiste à la **détection de zones de coexistence de sources harmoniques**.

L'utilité d'une telle détection peut être diverse en fonction du contexte et du type de contenu. En parole, de nombreuses techniques de transcription automatique de la parole trébuchent en cas de parole superposée. Les localiser précisément offre donc l'occasion d'envisager des parades spécifiques afin d'en améliorer les performances. Autre intérêt : à l'occasion de débats animés, les zones de changements de locuteurs sont souvent l'occasion de recouvrement de parole, un locuteur n'hésitant pas par exemple à couper la parole de son interlocuteur pour l'interpeller ou

le contester. Cette information peut préciser la nature de la détection d'un changement de locuteurs et elle peut s'avérer utile pour la structuration d'un document.

Dans un contexte musical, un tel système peut également servir la structuration en fournissant un découpage immédiat des zones de *solo* par rapport aux zones accompagnées. Cette structuration est parfois non applicable à la musique pop occidentale, dans la mesure où les solos purs (un seul instrument monophonique ou chanteur) sont peu courants, néanmoins, elle est pertinente dans de nombreux autres styles de musique.

La seconde consiste à **quantifier le rythme (présence et mesure)**. En contexte de parole comme de musique, cette information permet une classification du contenu. En musique, si la valeur du tempo peut servir d'indice pour la classification en un genre musical, cette information est essentielle pour la définition des longueurs des éléments de base constituant le morceau. Cette estimation peut alors constituer le fondement de beaucoup d'approches de structuration. Nous montrons qu'il existe également différents niveaux de parole caractérisés par la présence plus ou moins marquée de rythme.

1.2.1 Localisation de sources harmoniques simultanés

La question est la suivante : *« Est-il possible de localiser des zones temporelles où plusieurs sources harmoniques sont présentes* ***en même temps*** *? »* De nombreuses techniques de transcription automatique de la parole trébuchent en cas de parole superposée. Les localiser précisément offre donc l'occasion d'envisager des parades spécifiques afin d'en améliorer les performances.

Différentes heuristiques propres à la musique existent pour l'analyse des sources en présence. L'utilisation d'une échelle prédéfinie de fréquences correspondant aux notes pour la recherche de la présence des sources, ou la contrainte de l'interaction entre sources selon des règles musicologiques pourraient par exemple être utilisées. Nous avons cependant choisi de ne pas les utiliser car, si elles permettent de restreindre l'intervalle de recherche des phénomènes et donc d'apporter un gain de performances, elles le font au prix d'une perte de la généricité entre parole et musique. Nous avons donc choisi de ne pas fonder notre approche sur des règles issues de la musicologie et de garder une approche générique tout en assurant une robustesse maximale au type de contenu.

Plusieurs défis se posent pour la résolution de notre problématique

Le fait de ne pas utiliser d'hypothèses de localisation ou de relation entre sons impliquent que les sources peuvent interagir de n'importe quelle manière. De plus, nous nous restreignons aux sources harmoniques, de ce fait, elles génèrent un motif fréquentiel composé d'une fréquence principale et d'harmoniques associées à ses multiples entiers. Les harmoniques des différentes sources peuvent se superposer

de façon extrêmement complexe et par exemple se recouvrir à différents ordres, rendant la détection plus difficile.

Un autre défi est cette fois propre à l'analyse de la parole. En effet, si la musique est composés de phénomènes harmoniques relativement longs et stables, en parole seulement les 2/3 des phonèmes sont voisés, sons créant des phénomènes harmoniques. La parole étant constituée d'alternance de phonèmes (voisés ou non), la signature d'une source sur ces phénomènes harmoniques est extrêmement hachée. Du fait de ce hachage, les zones où deux phonèmes voisés de deux locuteurs différents se recouvrent sont d'autant plus rares et courtes. Ceci implique un très fort besoin de précision pour être capable de détecter des phénomènes présents sur des durées de l'ordre du quart de secondes. De plus, dans la majorité des contenus, les locuteurs sont conscients du besoin d'intelligibilité de l'enregistrement et essaient de minimiser ces zones de recouvrement et ne parlent simultanément que très peu de temps, réduisant encore la proportion de parole superposée sur tout l'enregistrement.

Pour relever ces différents défis et garder la généricité et une indépendance vis-à-vis du contenu, nous proposons une solution qui passe par l'utilisation de suivi de fréquences. Cette approche permet d'identifier une source dans des zones plus faciles à analyser et en suivre la présence à travers des zones où leur détection est beaucoup plus complexe.

1.2.2 Analyse rythmique

La seconde problématique à laquelle nous nous attaquons peut elle aussi être résumée par la question : *Existe-t-il une structure temporelle régulière du contenu ? Si oui, quelle est-elle ?*

L'utilité de cette information permet, en contexte musical, d'extraire une information de tempo reflétant la vitesse de réalisation du morceau. Cette information peut par ailleurs être ensuite utilisée en tant que paramètre pour de la similarité entre morceaux. De plus, cette information est essentielle pour la définition des longueurs des éléments de base constituant le morceau, et ainsi constitue le fondement de beaucoup d'approches de structuration. La problématique d'extraction du tempo comporte néanmoins un écueil majeur bien connu, celui des erreurs d'estimation par des *tempi* multiples ou sous-multiples de la valeur *correcte*. La notion de tempo « correct » est difficile à définir car sa définition peut induire des critères propres à chaque auditeur. Une définition des différents concepts liés au rythme et une discussion sur leur possibilité d'être retrouvée objectivement par des méthodes automatiques constitue un résultat important en soi.

En revanche, en parole, la notion de rythme est beaucoup moins utilisée en tant que telle. Si la notion de *prosodie* est importante, elle regroupe en général d'autres informations que la simple régularité rythmique puisque, variations de la

fréquence fondamentale, tons et accents font également partie de la *prosodie*.

Si la prosodie est utilisée en parole pour la recherche de structures syntaxiques en transcription automatique ou pour la détection d'émotions, nous présentons ici un objectif différent et original : mettre en évidence des stades intermédiaires entre des contenus de voix « parlée » et de voix « chantée ». Il existe, en effet, un large panel de caractérisation de la voix entre ces deux concepts : « réciter, scander ou encore déclamer » sont autant de catégories intermédiaires dont il est intéressant de relever l'existence. Par l'analyse rythmique et la découverte ou non de constante rythmique dans l'énoncé, nous cherchons à identifier les zones susceptibles de contenir un de ces types *intermédiaires* de contenu vocal.

Notre approche propose l'analyse d'une segmentation du signal afin de mettre en évidence une régularité dans la structure du signal et de ses instants d'accentuation énergétique.

Notre démarche impose de réaliser des méthodes théoriques généralistes, fonctionnant à la fois sur des contenus de parole et de musique. Ce fil conducteur impose des contraintes supplémentaires et la diversité des contenus ne permet pas de faire de nombreuses hypothèses simplificatrices liées à des connaissances *a priori* sur l'un ou l'autre des types de contenu. Cette contrainte forte permet d'explorer leurs propriétés communes afin d'extraire des phénomènes similaires mais qui, suivant le contexte, amènent à l'extraction d'informations différemment exploitables.

1.3 Organisation du manuscrit

Cet exposé des problématiques conduit à un manuscrit articulé autour de quatre chapitres.

Le premier chapitre présente le système d'indexation existant de l'équipe SAMoVA. D'une part, nous y présentons l'enchaînement de chacune des méthodes. D'autre part, nous y insérons nos deux nouvelles approches. Ce chapitre décrit également les différentes modifications et ajouts effectués sur ces outils afin de les intégrer au sein d'un système d'indexation complet et robuste.

Le deuxième chapitre présente tout d'abord une contextualisation de notre travail au sein des recherches existantes dans le domaine de la détection de sources harmoniques multiples. Puis nous détaillons les deux principales méthodes développées : le système de détection de chœur et le système de détection de zones de superpositions de sources harmoniques. Les notions permettant le suivi de fréquences y sont explicitées et le positionnement de cette approche par rapport aux méthodes précédentes y est discuté. Puis, nous commentons les résultats des expériences de faisabilité et validons notre approche théorique. L'implémentation de deux approches issues de l'état de l'art est également présentée et adaptée à notre contexte

d'étude. Une comparaison, puis une fusion des différentes approches sont présentées afin de proposer un système plus performant.

Le troisième chapitre, après une présentation des méthodes et concepts propres au domaine, s'attache à la description étape par étape de notre système d'analyse rythmique. Le découpage en zones homogènes ainsi que la technique d'analyse fréquentielle de ruptures y sont présentés. Ce chapitre propose ensuite deux applications possibles de la méthode : l'une en parole, l'autre en musique qui montrent que l'information rythmique extraite parvient à décrire différents types d'information. La validation de ces propositions par l'expérience ainsi que la discussion des résultats conclut ce chapitre.

Le dernier chapitre propose des expérimentations réalisées sur des données du corpus du projet ANR DIADEMS. Ce corpus est présenté avec ses particularités rendant le défi élevé. Les données de ce corpus sont en effet très hétérogènes au niveau du contenu puisqu'il est composé d'enregistrements musicaux, d'interviews et de contes, et ce pour de nombreuses ethnies différentes de par le monde. Mais la difficulté de ce corpus réside aussi dans la qualité des enregistrements puisque, outre le fait que les enregistrements sont effectués sur le terrain dans des conditions rarement optimales, les enregistrements couvrent une période allant de 1900 à nos jours, avec des supports d'enregistrements de qualité diverse et ayant subi de manière inégale le passage du temps et de la numérisation. Dans ce chapitre le défi applicatif est posé comme validation du défi théorique.

Chapitre 2

Système d'indexation

2.1 Introduction

L'intégralité du travail de recherche de ce doctorat s'est effectuée dans le cadre du projet *DIADEMS*. Ce projet, déjà présenté dans l'introduction, vise à fournir des technologies à des fins d'indexation automatique de données ethnomusicologiques extrêmement variées. Cette variété est présente à tous les niveaux : qu'il s'agisse de contenu (type de paroles, langues, type de musiques...), de contexte ou de support d'enregistrement (qualité, dégradations...)

Dans de telles conditions, il semble illusoire de créer une méthode de type classification suffisamment générique pour pouvoir s'adapter à toutes ces variations tout en restant robuste. Nous nous sommes orientés vers la création d'un système basé sur un enchaînement de méthodes permettant de préciser progressivement le contexte et le contenu du document analysé. Cette organisation hiérarchique permet de disposer, à chaque étape de traitement, des indices de confiance pour non seulement estimer la qualité de l'indexation en cours mais aussi pour tenir compte du score de la décision courante afin de paramétrer de manière plus adéquate la suite du processus.

La figure 2.1 illustre le système initial, composé de deux étapes : la segmentation Parole/Musique/Bruit et la classification Monophonie/ Polyphonie. Dans ce chapitre nous présentons les différentes méthodes utilisées dans chaque étape ainsi que leur intérêt dans l'analyse. La dernière section introduit l'ajout de notre méthode de suivi de fréquences par rapport à la contextualisation issue des étapes précédentes afin d'extraire différentes caractéristiques.

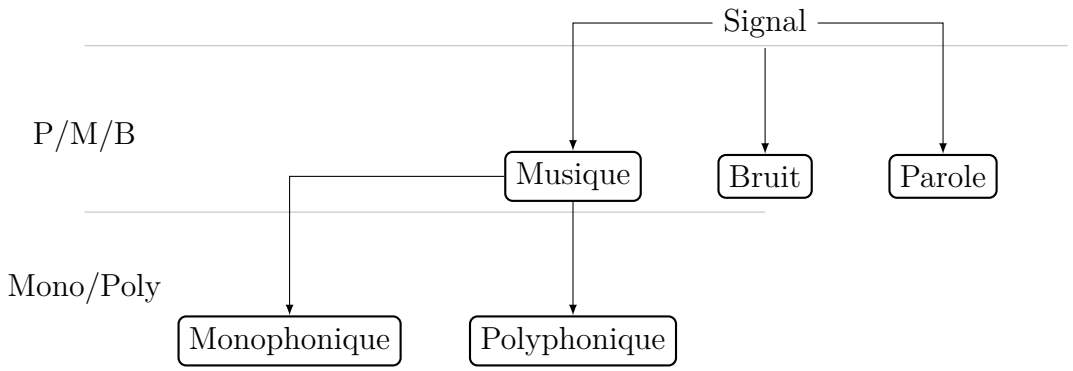


FIGURE 2.1 – Les deux niveaux de traitement du système existant.

2.2 Parole/Musique/Bruit

L'objectif de la première étape est la segmentation du signal en zones primaires : Parole, Musique et Bruit.

Cette étape est réalisée par les travaux de Pinquier en 2004 [1]. Elle implique deux systèmes distincts de détection de Parole ou de Musique.

La conjugaison des sorties des deux outils permet ensuite de réaliser la détections d'un ensemble de classes de manière indépendante. Elle conduit à retrouver à chaque instant l'une des classes suivantes :

- Musique et **non** Parole,
- **non** Musique et Parole,
- Musique et Parole,
- **non** Musique et **non** Parole : cette configuration correspondant à ce que nous appelons par la suite et abus de language : *Bruit*.

Les deux classifieurs se basent sur deux analyses statistiques différentes du signal. Voici leurs descriptions.

2.2.1 Détection de la *Parole*

Les deux paramètres statistiques sur lesquels se fondent la détection de parole sont la **modulation de l'énergie à 4 Hz** et la **modulation de l'entropie** du signal. Ces deux paramètres ont été choisis car chacun reflète une caractéristique du signal de parole. Les deux approches sont ensuite combinées pour détecter les zones de parole (figure 2.2).

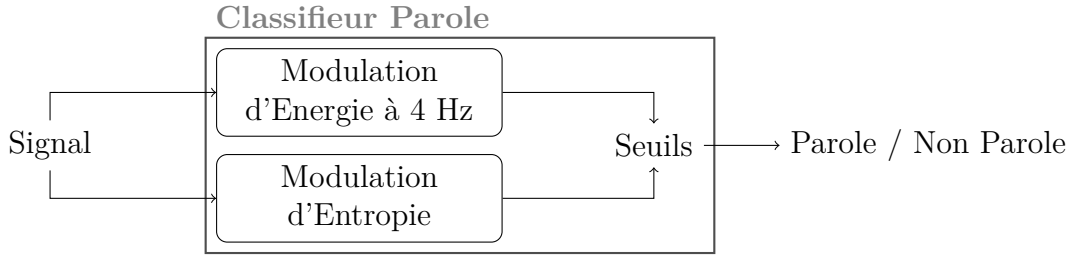


FIGURE 2.2 – Diagramme de flux pour la détection de parole

Modulation d'énergie à 4 Hertz

Ce paramètre vise à estimer la variation de l'énergie dans une bande de fréquence autour de 4 Hertz. Cette fréquence est choisie car elle correspond à une caractéristique de la production de parole, à savoir le débit syllabique. Une forte énergie dans cette bande de fréquence tend donc à exprimer la présence d'un débit proche de ce débit syllabique et il s'agit d'un indice fort de la présence de parole. Le diagramme de flux des différentes étapes de calcul de la fonction E_{4Hz} est présenté sur la figure 2.3.

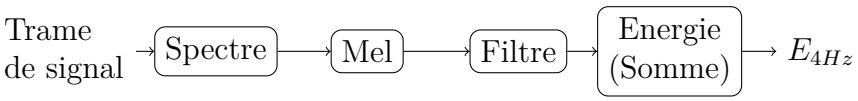


FIGURE 2.3 – Diagramme de flux des différentes étapes de calcul de l'énergie à 4 Hertz notée E_{4Hz}.

L'unité de calcul est une trame de signal de 16 ms. L'énergie de cette trame t, notée $E_{4Hz}(t)$ est extraite. Pour ce faire, nous utilisons 40 bandes de fréquence réparties suivant l'échelle Mel pour représenter le spectre d'énergie au travers de 40 coefficients spectraux. Un filtre à réponse impulsionnelle finie d'ordre 100, est paramétré pour laisser passer les fréquences autour de 4 Hertz.

Cette fréquence a été choisie car elle correspond au débit syllabique souvent observé en parole. Une forte énergie dans cette bande de fréquence tend donc à exprimer la présence d'un débit proche de ce débit syllabique et il s'agit d'un indice fort de la présence de parole. Ce filtre très souple laisse en réalité passer beaucoup plus que le voisinage immédiat de 4 Hertz. L'énergie est finalement calculée comme la somme du spectre filtré (figure 2.4).

La modulation $M_{4Hz}(t)$ est ensuite calculée à partir de la suite E_{4Hz}. Elle correspond à la variance du log de l'énergie à 4 Hertz, calculée sur une fenêtre de **2 secondes** du signal, centré sur chaque trame. Cette durée permet d'obtenir un nombre significatif de points correspondant à la durée d'une phrase.

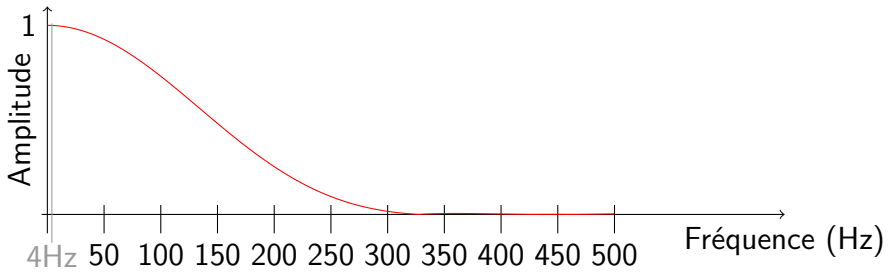


FIGURE 2.4 – Zoom sur la réponse du filtre utilisé pour le calcul de l'énergie autour de 4 Hertz.

La figure 2.5 illustre le comportement de ce descripteur sur un signal test composé à moitié de parole puis de musique.

La valeur de ce descripteur est ensuite comparée à un seuil th_{4Hz} afin d'extraire les segments correspondant à la parole.

Modulation d'entropie

Autre paramètre visant à la détection des zones de parole, l'entropie de Shannon du signal est utilisée afin de différencier la parole de la musique, au travers de son organisation temporelle. L'entropie de Shannon en base b, notée H_b, est une mesure statistique permettant d'exprimer la prédictibilité d'une série. Cette notion est intimement liée à celle de désordre de la série.

Sa définition mathématique est la suivante :

$$H_b = -\sum_{i=1}^{n} P_i log_b P_i \tag{2.1}$$

avec n le nombre de symboles x_i possibles pris par la série ; chaque x_i apparaît avec une probabilité P_i. D'un point de vue pratique, nous utilisons le logarithme népérien (base e), et chaque P_i est estimé à partir de la série.

Le comportement de cette information est très différent en fonction des contenus. Comme nous pouvons le voir sur la figure 2.5, si le signal de parole est très chaotique, celui de musique est nettement plus ordonnée et constant. Le signal de musique possède donc une valeur d'entropie de Shannon beaucoup plus faible que celui de la parole et sert à discriminer les deux parties.

Nous analysons des trames t de 16 ms afin de créer la fonction d'entropie notée $H(t)$.

La modulation est ensuite calculée ; elle correspond à la variance de l'entropie, calculée sur une fenêtre de **2 secondes** du signal, centrée sur chaque trame.

La valeur de ce descripteur est également comparée à son seuil th_{ent} pour extraire les segments correspondant à la parole.

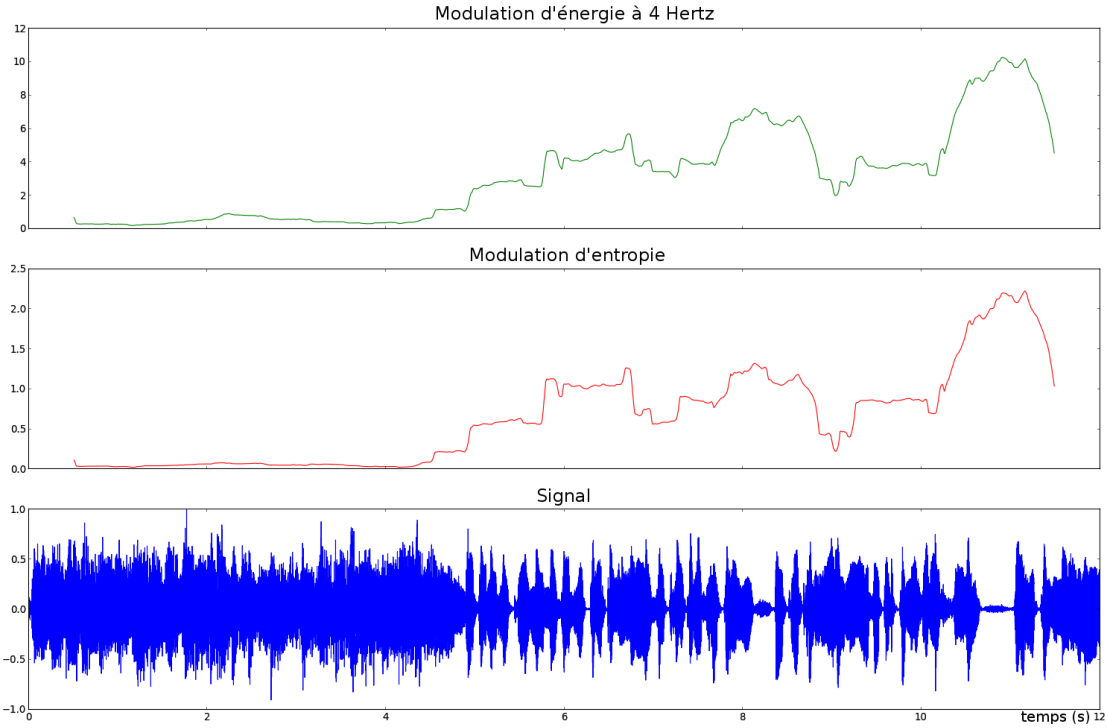


FIGURE 2.5 – Modulation d'énergie à 4 Hertz et modulation d'entropie sur un signal de 12 secondes composée à moitié de musique puis de parole.

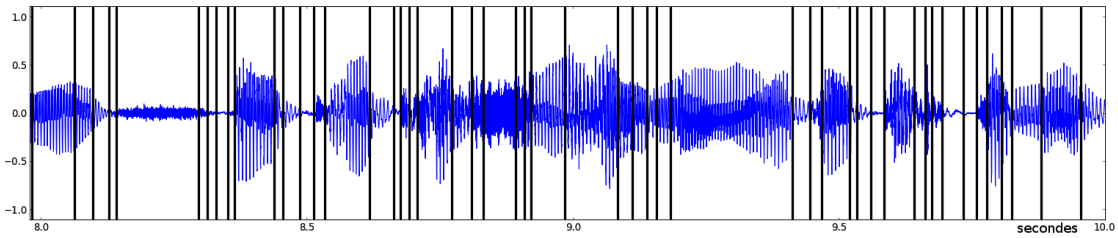
2.2.2 Classifieur Musique

La classification en musique repose sur deux descripteurs. Ces deux paramètres sont liés à l'analyse d'une segmentation en zones quasi stationnaires, chaque zone étant modélisée par un modèle autorégressif gaussien. L'algorithme appelé **Segmentation par divergence Forward-Backward** ou **SFB**, proposée par André-Obrecht [2], est détaillé dans le chapitre suivant consacré à l'analyse du rythme.

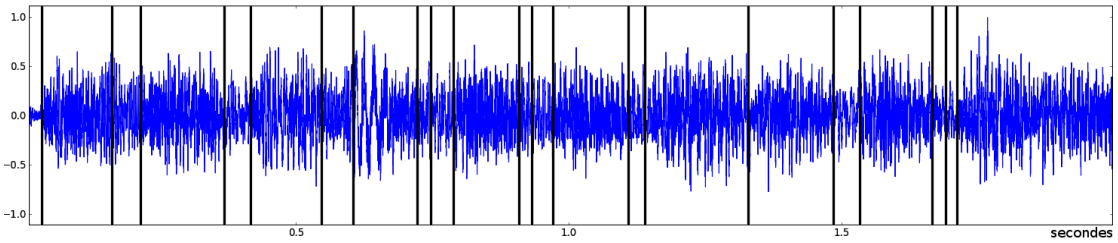
La **SFB** a initialement été conçue pour l'analyse de la parole. Son objectif est de découper le signal en zones sub-phonétiques, permettant d'identifier les différentes phases de réalisation d'un phonème. Appliquée à la musique, cette segmentation s'approche d'une séparation des différentes phases de réalisation d'une notes (cf section 3). Or, les phénomènes liés aux notes ont tendance à être plus longs que ceux liés aux phonèmes.

La figure 2.6 présente les résultats de la **SFB** sur 2 secondes de signal de parole **(a)** et de musique **(b)**. Sur cette comparaison, nous constatons que la segmentation est beaucoup plus dense sur le signal de parole : en parole, les segments sont plus nombreux et beaucoup plus courts qu'en musique.

L'étude des segments générés par cette méthode est utilisée de deux manières



(a) Segmentation sur un signal de parole.



(b) Segmentation sur un signal de musique

FIGURE 2.6 – Segments obtenus par l'algorithme **SFB** sur 2 secondes d'un signal de parole et de musique. La segmentation (traits noirs verticaux) est beaucoup plus hachée sur l'exemple de parole que sur celui de la musique : les segments y sont plus nombreux et beaucoup plus courts.

différentes : nous analysons la **longueur des segments** les plus longs par seconde ainsi que le **nombre de segments** présents dans une seconde.

Longueur des Segments

L'un des paramètres de détection de musique utilise cette différence en analysant la longueur des N_{seg} plus grands segments présents.

A chaque instant t, la valeur du paramètre $L_{seg}(t)$ est définie comme la moyenne des longueurs des segments sur l'intervalle $[t-0.5s, ..., t+0.5s]$. Certains segments peuvent se trouver de part et d'autre d'une limite de l'intervalle, ces segments sont considérés dans leur intégralité pour le calcul de la moyenne.

Cette valeur est ensuite comparée à un seuil de décision th_{LSeg} afin de décider si la seconde analysée contient de la musique ou non.

Nombre de Segments

Le second paramètre extrait, pour la classification en musique d'une zone d'une seconde, est le nombre de segments. Noté $Nb_{seg}(t)$, il représente le nombre de segments présents dans l'intervalle $[t-0.5s, ..., t+0.5s]$. Ici aussi, les segments situés à la marge de l'intervalle d'analyse sont comptabilisés. Il est considéré comme

information complémentaire, même si cette information est corrélée à la longueur des segments.

Ce descripteur est ensuite comparé à un seuil noté th_{NbSeg}.

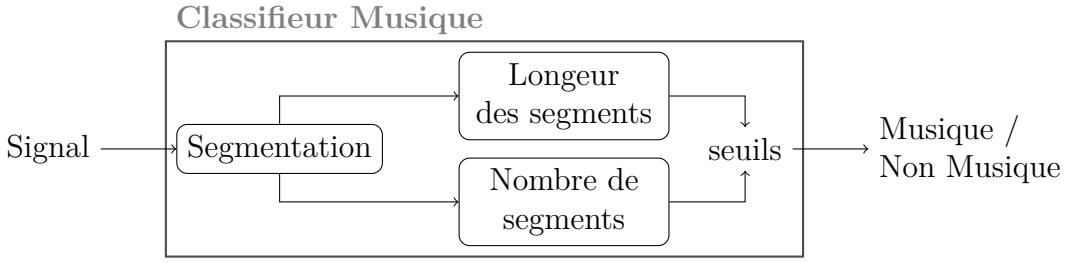


FIGURE 2.7 – Diagramme de flux pour la détection de musique

2.2.3 Robustesse et indices de confiances

Si ces quatre paramètres ont fait leurs preuves indépendamment sur des contenus produits et enregistrés dans de bonnes conditions, nous pouvons nous interroger sur leur comportement dans l'analyse de documents moins préparés comme ceux du corpus DIADEMS.

Comme il semble difficile d'envisager un système pouvant fonctionner parfaitement sur la très grande variabilité du corpus, nous avons choisi de ne plus utiliser une classification binaire basée sur un seuil, mais de retourner un indice de confiance de chaque descripteur sur le type du contenu.

Pour créer cet indice de confiance, nous avons utilisé la même approche pour les quatre descripteurs. Si la valeur du paramètre p est comparée au seuil s pour la classification i , l'indice de confiance I^i_{conf} est calculé selon la formule suivante :

$$I^i_{conf} = \frac{p - s}{s} \tag{2.2}$$

Comme les paramètres p sont toujours définis sur $\mathbf{R}^+$, la valeur minimale de l'indice de confiance est -1. En revanche, cet indice n'est pas borné et peut changer d'échelle d'un paramètre à l'autre. Pour résoudre se problème, nous utilisons la formule bornée suivante :

$$I^i_{conf} = \begin{cases} 1 & si\ p > 2 * s \\ \frac{p-s}{s} & sinon \end{cases} \tag{2.3}$$

La valeur de I_{conf} est donc définie sur l'intervalle $[-1, 1]$, 1 indiquant une confiance forte en la présence de la classe et -1 indiquant une confiance forte en son absence (et donc une confiance forte en la présence de la non-classe). Cette modification du système existant est présenté sur la figure 2.8.

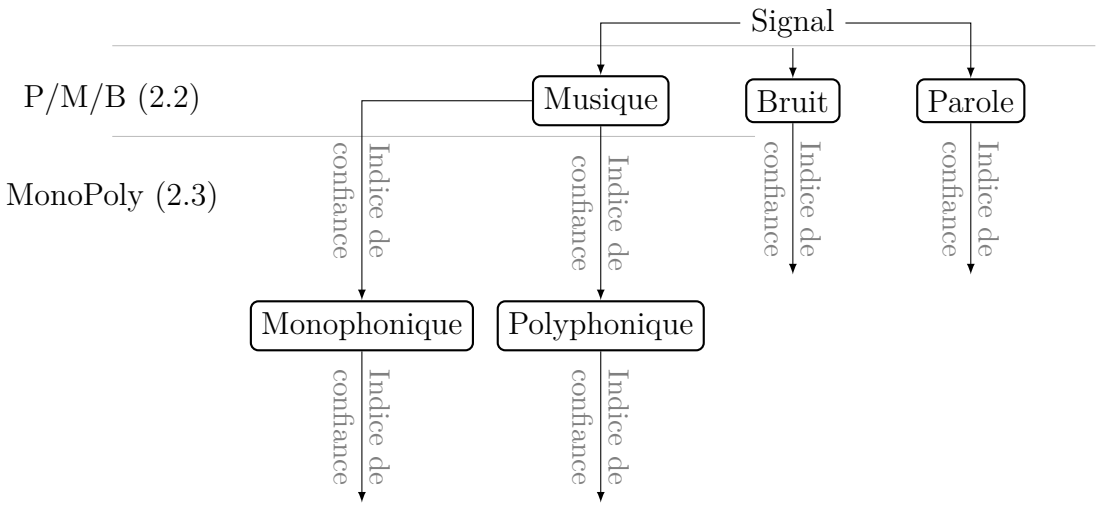


FIGURE 2.8 – Modification du système existant par l'ajout d'un système d'indice de confiance

2.2.4 Fusions et décisions

Cette définition des indices de confiance permet de mettre plus facilement en place des stratégies de fusion des décisions. Une simple moyenne pondérée entre les indices associés à chacun des descripteurs c_1 et c_2 permet de créer une valeur de décision v_d.

$$v_d = \alpha I_{conf}^{c_1} + \beta I_{conf}^{c_2} \tag{2.4}$$

Le signe de v_d permet simplement de savoir si la classe est présente et avec quel degré de certitude.

2.2.5 Conclusion

Le premier niveau d'analyse Parole/Musique/Bruit propose une séparation des composantes principales du signal. Ce traitement s'appuie sur une analyse statistique du signal de différents paramètres représentatifs de la stabilité du signal ou de la présence forte de variation autour de fréquence liée au débit de la parole. Afin de rendre la décision plus souple, un système d'indice de confiance a été mis en place. Celui-ci autorise des remises en cause des résultats.

2.3 Monophonie/Polyphonie

Afin d'affiner la connaissance de la classe musique, une méthode de classification monophonie/polyphonie, a été proposée par Lachambre [3]. Cette méthode s'appuie sur la paramétrisation du signal par une valeur d'harmonicité du signal. Celle-ci est l'une des valeurs intermédiaires calculées par l'algorithme d'extraction de pitch monophonique *YIN* [4] et plus précisément par la fonction *CMNDF*.

Les moyenne et variance de ce paramètre sont comparées à des modèles de musique monophonique et polyphonique. La figure 2.9 présente un diagramme de flux récapitulatif des traitements effectués lors de la détection monophonie/polyphonie.

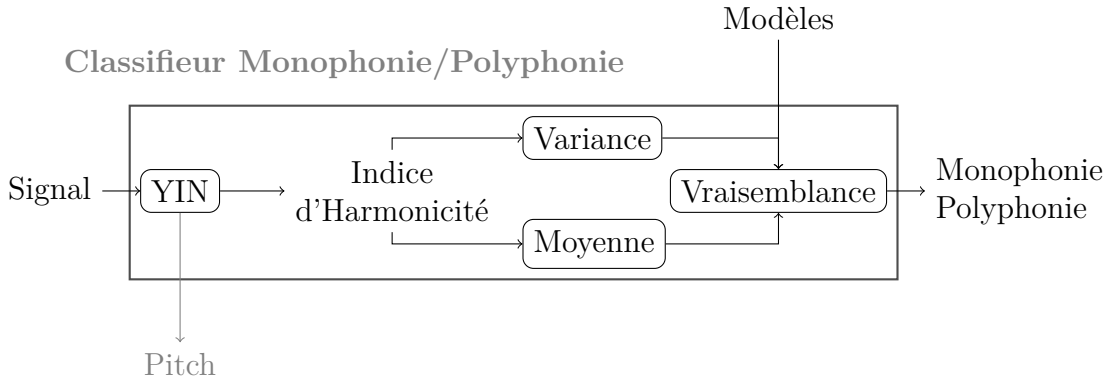


FIGURE 2.9 – Diagramme de Flux de la détection Monophonie/Polyphonie.

Contrairement à la détection Parole/Musique/Bruit qui impliquait deux systèmes différents de classification, il s'agit ici d'une classification binaire Monophonie/Polyphonie.

2.3.1 Cumulative Mean Normalized Difference Function

La *CMNDF* pour **C**umulative **M**ean **N**ormalized **D**ifference **F**unction, est une fonction dérivée de la fonction d'auto-corrélation. Dans un signal parfaitement harmonique et de fréquence f_0, la fonction de différence $d_t(\tau)$, définie plus bas, est nulle pour $\tau = \frac{1}{f_0}$.

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2 \tag{2.5}$$

avec W la taille de la fenêtre d'analyse et x_j le jème échantillon du signal dans la fenêtre d'analyse. Cette fonction atteint des valeurs minimales pour les valeurs de τ liées à la fréquence fondamentale ($f0$) du signal, même si elles n'atteignent pas 0 en raison de l'imperfection de la périodicité. Cette fonction doit être quasi

nulle pour une valeur de τ, cependant, une forte résonance causée par le premier formant (F1) peut conduire à des valeurs très faibles, voire plus faibles que pour $\frac{1}{f_0}$.

Pour résoudre ce problème une autre fonction est définie : la *CMNDF*. Cette fonction est définie à partir des valeurs de la fonction de différence d_t :

$$CMNDF_t(\tau) = \begin{cases} 1 & si\ \tau = 0 \\ d_t(\tau)/\frac{1}{\tau}\sum_{k=1}^{\tau} d_t(k) & sinon \end{cases} \tag{2.6}$$

Chaque valeur à la période τ est pondérée par la moyenne de la fonction de différence obtenue sur toutes ses valeurs calculées sur les périodes inférieures.

Cette fonction permet ainsi de ne pas avoir à utiliser un *a priori* sur la limite basse de la recherche de f_0 puisque la première valeur est artificiellement fixée à 1 et que les faibles périodes sont rehaussées par la normalisation.

Le premier pic de la *CMNDF* de valeur inférieure à 0,1 est sélectionné comme étant le pic correspondant à la fréquence fondamentale $\widehat{f_0}$. La valeur correspondante : $CMNDF_t(\widehat{f_0})$ est utilisée comme indice de l'harmonicité du signal, noté I_{harmo}. La courbe de la fonction $CMNDF$ est présenté sur la figure 2.10.

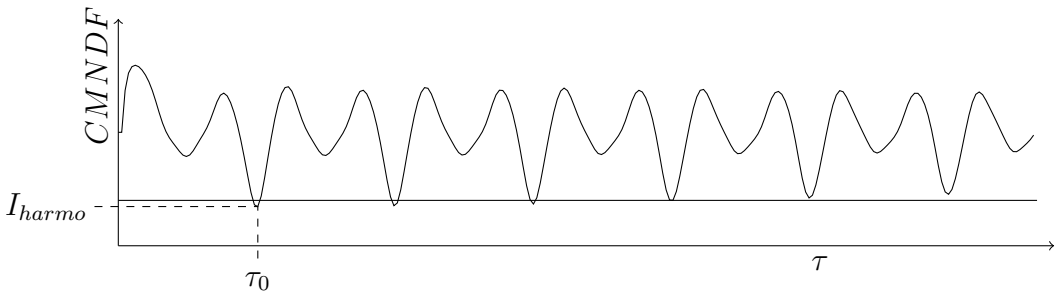


FIGURE 2.10 – CMNDF sur une trame de musique polyphonique.

L'utilisation de la valeur de la $CMNDF$ comme indice de confiance en l'harmonicité tient au fait que plus le signal sera pur et monophonique, plus cette valeur sera proche de 0. De plus, comme seul les pics de valeur inférieure à 0,1 seront sélectionnés, nous connaissons donc les deux *extrema* possibles pour la valeur de l'indice, lui donnant une valeur normalisée.

2.3.2 Classification

Du fait que cette analyse soit effectuée au sein d'une composante musicale, il est justifié de la rechercher sur des segments d'au moins une seconde ; c'est pourquoi une décision est envisagée après analyse individuelle de segments de 1 seconde. La

valeur de $CMNDF$ est calculée sur des segments de 16 ms dans chaque seconde. Les moyenne et variance notées respectivement m et v de la série des $CMNDF$ sont calculées.

Les valeurs m et v extraites sont comparées par maximum de vraisemblance. Deux modèles probabilistes sont appris pour les classes monophonique et polyphonique sur des corpus de musique propre. Les lois utilisées sont des lois de Weibull bivariées. Ces lois ont été choisies pour leur meilleure adéquation aux histogrammes observés des données que d'autres lois statistiques plus couramment utilisée comme les mélanges de lois gaussiennes. Un autre avantage est que l'algorithme d'estimation de ces lois reste simple.

Les figures 2.11 et 2.12 représentent respectivement les densité de probabilité des modèles de contenus monophoniques et polyphoniques sur l'espace des paramètres moyenne et variance des valeurs de $CMNDF$.

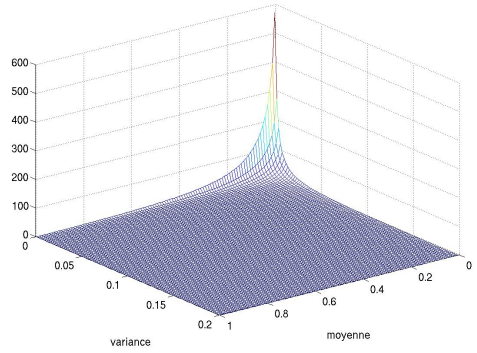


FIGURE 2.11 – Modèle probabiliste pour les contenus monophoniques.

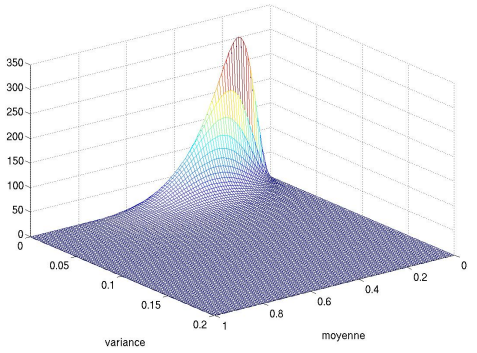


FIGURE 2.12 – Modèle probabiliste pour les contenus polyphoniques.

2.4 Prétraitement pour le suivi de fréquences

Les systèmes de classification présentés précédemment permettent de sélectionner et contextualiser des zones d'analyse avant d'appliquer des traitements ultérieurs tels que notre méthode de détection de zones multi-sources (présentée dans le chapitre suivant). Les méthodes sont mises en œuvre sans nouvel apprentissage et comme proposées par le passé ; les informations ainsi recueillies ne possèdent pas toujours exactement toutes les propriétés assurant le bon déroulement de la prochaine étape.

Afin de permettre de meilleurs résultats, certaines étapes de prétraitement sont ajoutées. Ces étapes, utilisées surtout pour la partie parole, permettent une

correction des zones d'analyse afin de tenter d'éviter à la fois, les fausses alarmes et les détections manquées.

2.4.1 Énergie

Comme nous l'avons présenté dans la partie consacrée au classifieur Parole/ Musique/Bruit (section 2.2), l'un des paramètres utilisés pour la détection de la parole doit rendre compte d'un débit proche ou non du débit syllabique. Or, dans les zones de parole superposées, cette information de débit syllabique risque d'être fortement perturbée. Il en résulte que des zones de parole superposées ne vont pas être considérées comme étant de la parole, et donc risquent de ne pas être analysées.

Pour résoudre ce problème nous avons choisi de tirer partie de la brièveté des phénomènes de parole superposées de manière générale. Nous effectuons un lissage afin de déceler de courts segments de bruits encadrés par des segments de parole. La longueur maximale des segments ainsi récupérables est fixée grâce à un seuil L_{Bruit}. Néanmoins, ces segments peuvent aussi se révéler être réellement du bruit ou du silence auquel-cas ils ne doivent pas être ajoutés aux zones d'intérêt car ils risqueraient de devenir source de fausses alarmes.

Afin d'assurer une correction pertinente, nous comparons l'énergie du segment de bruit par rapport à ses voisins immédiats de parole. Si l'énergie du bruit est supérieure ou égale la moitié de l'énergie des segments de parole, alors la fusion des trois segments en un seul segment de parole est effectuée.

Les différentes étapes de ce processus sont illustrées par la figure 2.13.Le court segment de bruit (noté D)est fusionné avec les deux segments de parole l'encadrant (segments C et D).

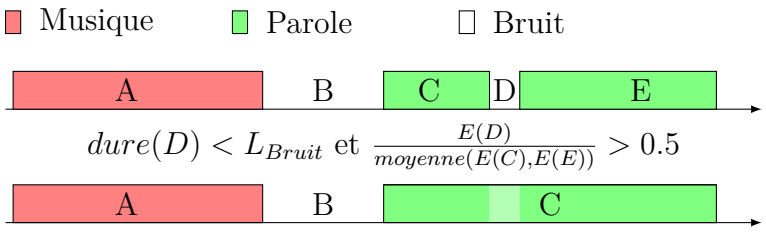


FIGURE 2.13 – Principe de lissage des courts segments de bruits.

Cette correction permet de récupérer des zones de parole superposée. En effet, n'étant pas considérées comme n'étant de la parole, certaines zones n'auraient même pas subi d'analyse. L'intérêt de cet ajout sur l'analyse des contenus de parole sera décri plus amplement dans le chapitre 4 traitant des expérimentations.

2.4.2 Harmonicité des trames

Un autre phénomène peut générer des erreurs : la recherche de sources harmoniques dans des trames n'en contenant pas. Notre méthode est en effet basée sur l'hypothèse de présence d'au moins une source et considère que le pic d'amplitude maximum du spectre appartient à la source la plus prédominante. En cas d'absence prolongée de sources harmoniques, le suivi que nous proposons ne peut plus être exploité correctement. Il est donc important d'éviter au maximum l'analyse sur des zones de bruit, même très courtes.

Pour l'instant, la solution que nous préconisons est l'utilisation du critère d'harmonicité utilisé par le système d'estimation de la section 2.3.

2.5 Conclusion

Nous proposons un système hiérarchique d'indexation primaire permettant de caractériser les contenus sonores sur différents niveaux de précision. Nous nous appuyons sur des travaux précédemment réalisés au sein de l'équipe *SAMoVA*. L'enchaînement de ces méthodes nous permet d'extraire les informations suivantes : Parole/Musique/Bruit et Monophonie/ Polyphonique

Ces travaux ont fait la preuve de leur efficacité dans des conditions traditionnelles d'enregistrement (qualité studio). Néanmoins, dans le cadre du projet DIADEMS, ils devront être utilisés sur des enregistrements très variés en terme de contenu et de qualité. Nous avons ajouté une plus grande flexibilité à la décision sur l'identification des zones de Parole/Musique/Bruit et nous avons choisi de ne plus considérer la seule décision binaire mais de lui adjoindre un système d'indice de confiance permettant la remise en cause éventuelle de la décision en cas de confiance faible.

Les différentes informations extraites servent ensuite à sélectionner les zones analysées qui seront soumises à notre méthode de détection de sources. Seules les zones détectées comme parole ou musique et contenant des sources harmoniques seront étudiées. De plus, cette analyse est paramétrée en fonction du type de contenu et tient compte des propriétés particulières, notamment propres aux contenus de musique ou de parole. La proposition de nouveau système par l'ajout de nouvelles méthodes est présenté dans la figure 2.14.

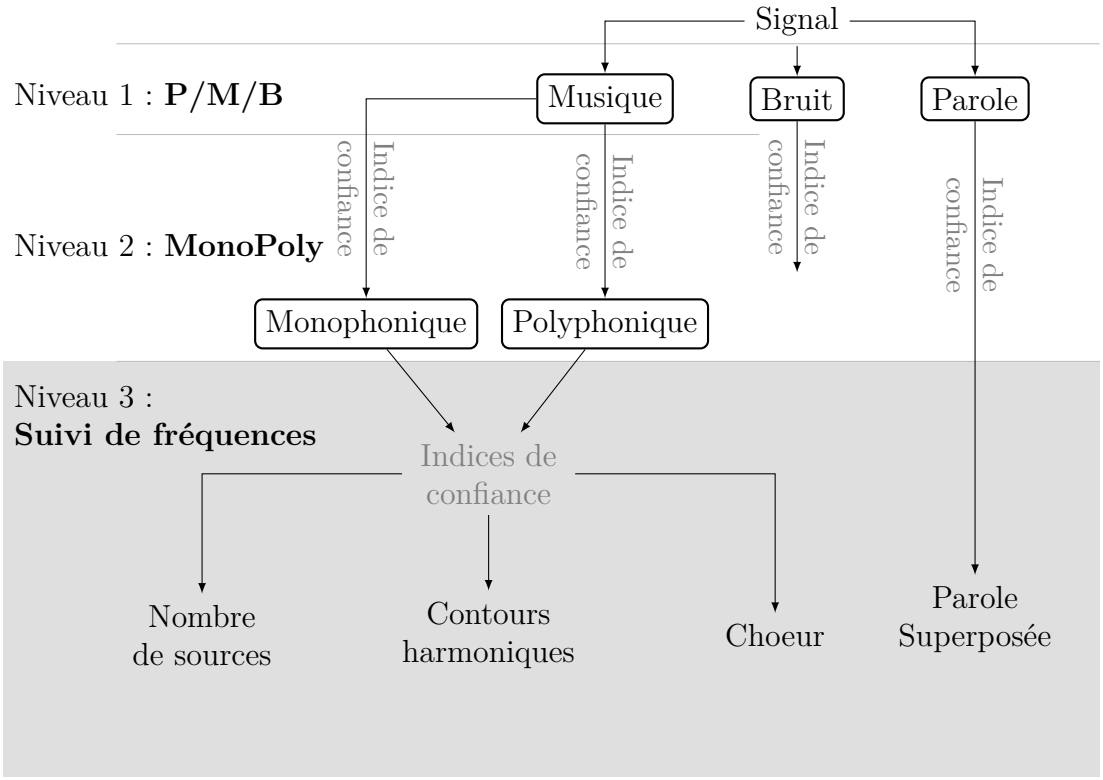


FIGURE 2.14 – Les trois niveaux de traitement du nouveau système proposé

Chapitre 3

Analyse du Rythme

3.1 Introduction

L'information de rythme dans les contenus sonores est très précieuse pour la description de contenus. Qu'elle serve de base à la découverte d'une structure ou qu'elle puisse être en tant que telle caractéristique, son identification et sa caractérisation est souvent précieuse.

Dans cette partie nous présentons la méthode que nous avons développée pour l'analyse rythmique des contenus de parole et de musique. Cette approche identique pour l'analyse de ces deux types de contenus vise à la recherche de régularités dans les positions de forte énergie.

Ce chapitre présente dans un premier temps une revue de quelques méthodes utilisées dans la littérature pour l'extraction d'information rythmique dans les contenus de musique et de parole.

Dans un second temps nous détaillons les différentes étapes de notre approche :

- présentation de la méthode de détection de ruptures dans le signal,
- description de la technique d'analyse fréquentielle développée,
- présentation du système de pondération des frontières porteuses d'information rythmique,

Dans la troisième et dernière partie, nous décrivons deux applications possibles de notre méthode sur des contenus de parole et de musique que nous expérimentons. Leurs résultats sont discutés afin de valider notre approche théorique.

3.2 Revue de l'analyse rythmique

3.2.1 En musique

Le problème de l'estimation du tempo en musique est un problème abordé depuis longtemps ; le tempo est l'une des caractéristique primaires d'un morceau de musique et c'est pourquoi il est très vite apparu comme fondamental dans l'élaboration de méthodes comme la classification en genres ou la similarité. L'information du tempo est également primordiale pour toutes les tâches de structuration du morceau de musique.

Les pulsations définissent les instants possibles de changements dans la structure d'un morceau. Elles correspondent à des accentuations données à la musique de manière cyclique, garantissant que chaque unité dure le même temps, la même période. Le **tempo** est la fréquence correspondant à cette période. Trouver le tempo, c'est non seulement avoir un indice de la vitesse du morceau, mais aussi une information cruciale pour pouvoir localiser les pulsations et donc les temps : l'unité de la structure musicale.

Cette notion de tempo est essentiellement musicologique et définie par le compositeur. En ce sens elle offre un coté subjectif et est par conséquent difficilement extractible depuis le signal audio. Afin de comprendre ce problème, deux concepts différents liés au tempo sont mis en avant par la communauté musicologue : le *tactus* et le *tatum*.

Le **tactus** est défini comme « *le rythme auquel les personnes écoutant la musique tapent dans leur main* » [5]. Cette notion, basée sur la réaction des auditeurs, est subjective puisqu'au sein d'un même groupe nous pouvons trouver des personnes frappant à des vitesses différentes. Ces différences sont néanmoins généralement liées comme étant des multiples entiers. Ces différences peuvent s'expliquer par des différences de culture musicale entre les membres de l'auditoire. De part ce caractère subjectif, cette mesure ne semble donc pas totalement adaptée à l'indexation et il est nécessaire d'introduire un concept plus objectif pour décliner une recherche automatique.

Dans ces mêmes travaux [5], la notion de **tatum** est définie de manière plus objective puisqu'il s'agit de « *l'intervalle coïncidant le plus souvent avec les débuts de note* ». C'est cette valeur, beaucoup plus mesurable que nous utiliserons par la suite pour calculer le tempo. Ce tempo est calculé comme la fréquence correspondant à cet intervalle, exprimé en **bpm** (battement par minute). L'estimation de la valeur du tempo est primordiale pour construire des informations structurelles de plus haut niveau. Elle permet, entre autres, d'obtenir une idée de la taille des segments minimaux composant le morceaux analysé. En raison de son importance,

des campagnes d'évaluations internationales comme celle de MIRex [1] ou celle interne au projet ANR QUAERO [2] ont proposé une tâche d'estimation de tempo. En 2004, par exemple de nombreux systèmes différents furent proposés à l'évaluation MIRex. Cette évaluation a donné lieu à une synthèse comparant les performances et caractéristiques des différents systèmes [6].

La plupart des méthodes d'estimation de tempo se basent sur un détecteur d'**onset**. L'onset est l'instant correspondant au début d'une note. Différentes façons d'extraire la position des onsets sont proposées à travers la littérature.

Par exemple, l'un des algorithmes de Dixon [7] utilise un calcul de l'énergie afin d'extraire les positions probables des onsets. Une analyse des intervalles inter-onsets est ensuite effectuée afin de regrouper ceux pouvant correspondre à une même valeur de tempo et ainsi extraire les hypothèses de *tempi* les plus représentées. La décision finale est ensuite effectuée entre ces *tempi* les plus probables en utilisant un suivi de temps basé sur chacune des hypothèses. L'hypothèse générant les temps les plus en phase avec les instants de saillance du signal est considérée comme correcte. Cette saillance est calculée en utilisant une combinaison de durée de note, d'amplitude et de pitch.

Une autre approche utilisée par Alonso [8] repose sur une représentation spectrale filtrée du signal afin de mettre en évidence des changements dans la dynamique de cette représentation fréquentielle. Ces instants de variations sont ensuite considérés comme des onsets probables et une fonction de probabilité de présence d'onsets est générée. Cette fonction est ensuite analysée selon deux méthodes (l'auto-corrélation et le produit spectral) afin de déterminer la périodicité la plus présente.

Afin de prendre en compte différents types de changements, les changements mélodiques étant plus difficiles à déceler que les changements percussifs, l'algorithme de Klapuri [9] propose une extraction de la dérivée de l'énergie sur 36 sous-bandes de fréquences. Les valeurs dans ces différentes bandes sont ensuite combinées dans 4 bandes représentant les accentuations. Un banc de filtre (type peignes) est ensuite utilisé sur ces 4 bandes d'accentuation afin de déterminer des paramètres qui servent, grâce à une approche statistique à l'estimation conjointe des valeurs du *tatum*, du *tactus* et de la mesure.

D'autres méthodes, comme celles de Uhle [10] ou Dixon [11], impliquent le calcul de l'enveloppe du signal sur différentes bandes de fréquences. La fonction d'auto-corrélation de cette représentation est ensuite calculée afin de mettre en valeur la présence d'un rythme particulier.

Des approches comme celle de Tzanetakis [12] utilisent une représentation temps-fréquence différente de celle de Fourier classiquement utilisée en musique

1. `http://www.music-ir.org/mirex/wiki/MIREX_HOME`
2. `http://www.quaero.org/`

et proposent une analyse basée sur une transformée en ondelettes [13]. La suite du processus est plus classique puisqu'elle consiste aussi en une extraction de l'enveloppe sur 5 bandes de fréquences. Ces bandes correspondent à des octaves. La fonction d'auto-corrélation est calculée sur chacune de ces bandes pour mettre en valeur certaines valeurs candidates de tempo. Trois combinaisons différentes sont proposées pour prendre une décision à partir des résultats de l'auto-corrélation.

Autre méthode d'analyse du tempo, celle de Peeters [14], offre la possibilité de localiser des changements de rythme au sein d'un même morceau. Bien que peu courants dans les enregistrements de musique pop, ces changements de rythmes peuvent très bien se produire dans différents autres types de musique moins stéréotypées. Pour son analyse, cet algorithme utilise une méthode de ré-assignement de spectrogramme proposée par [15] qui permet une nette amélioration des précisions temporelles comme fréquentielles du spectrogramme en réaffectant les coordonnées du spectrogramme afin de mieux représenter la distribution de l'énergie.

La détection d'*onsets* s'effectue sur ce spectrogramme réaffecté, ce qui permet une localisation plus précise dans le plan temps-fréquence. Le spectrogramme est ensuite filtré selon différentes méthodes et séparé en différentes bandes de fréquences afin d'optimiser la détection de tous les types de notes jouées et de se rapprocher du système perceptif humain. Un signal de détection d'*onset* est finalement produit en sommant sur les fréquences.

Afin de localiser les récurrences temporelles locales au sein de ce signal de détection d'onset une analyse est effectuée par trame. Sur chaque trame, une combinaison de transformée de Fourier et de fonction d'auto-corrélation est calculée afin de réduire les erreurs d'estimation. Enfin, plusieurs *tempi* candidats sont estimés en lien avec la valeur trouvée (double, moitié ...). La série de tempo correspondant le mieux aux observations au cours du temps est finalement calculée à l'aide de l'algorithme d'alignement de Viterbi [16].

Il en résulte que le schéma général de la recherche de tempo est donc relativement semblable parmi les différentes méthodes présentées (figure 3.1). Toutes ces analyses se basent sur une recherche d'accentuation dans le signal en faisant l'hypothèse que ces accentuations correspondent aux onsets. De nombreuses méthodes effectuent cette recherche dans différentes sous-bandes de fréquences afin d'être le plus sensible possible aux différents instruments, et limiter l'effet de masquage que pourrait avoir un instrument sur un autre. Une analyse fréquentielle de cette représentation de l'accentuation est ensuite effectuée, la plupart du temps par des techniques d'auto-corrélation, afin de déterminer la période la plus présente.

Chaque méthode apporte ses spécificités soit en terme de recherche des accentuations, soit en terme d'analyse fréquentielle. Toutes se basent néanmoins sur le fait que la période la plus présente dans le positionnement des accentuations correspond au tempo. Elles se réfèrent donc toutes implicitement à la valeur du

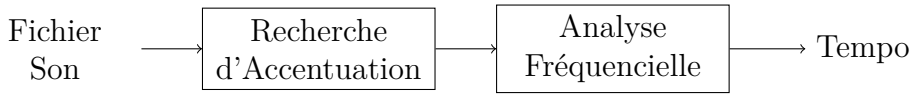


FIGURE 3.1 – Schéma général des méthodes d'analyse du tempo.

tatum.

3.2.2 En parole

Dans le domaine du traitement de la parole, l'étude du rythme est intrinsèquement liée à la prosodie. Le concept de **prosodie** regroupe différentes choses, et définir la prosodie n'est pas chose aisée. Il existe dans la littérature des définitions du concept de prosodie qui regroupent plusieurs réalités différentes en fonction de l'objectif. Le terme de prosodie a longtemps oscillé entre les deux concepts de métrique et d'intonation en fonction des traditions, littéraires ou linguistiques. Ces deux concepts de métriques et d'intonation sont considérés comme étroitement liés dans l'établissement de ce qu'est la prosodie.

Nous retiendrons pour notre étude la définition de la prosodie liée aux sciences du langage proposée par Di Cristo [17] :

« *Appréhendée en tant que discipline des sciences du langage, la prosodie (que l'on pourrait alors qualifier de prosodologie) est couramment définie comme le champ d'étude d'un ensemble de phénomènes tels que l'accent, le rythme, les tons, l'intonation, la quantité, les pauses et le tempo, qui constituent ce qu'il est convenu d'appeler les éléments prosologiques [...]* »

En parole, **rythme** et **tempo** sont donc des éléments pertinents de description des contenus. Un autre paramètre évoqués dans cette définition de la prosodie est l'**accent**. Ce terme d'accent revêt ici aussi plusieurs définitions. Il peut être lié à une prononciation particulière et constante d'une langue (accent anglais ou toulousain par exemple). L'autre sens qui ici nous intéresse est dès lors qu'il dénote une marque particulière apportée à un instant précis, « *dont la réalisation s'accompagne d'une augmentation de la force articulatoire, de l'intensité, de la durée et de certaines modifications du spectre acoustique de la voyelle de la syllabe accentuée* » [17].

Nous pouvons donc trouver dans ces accents, porteurs d'une augmentation forte de l'énergie du signal, le pendant des **onsets** de la musique au sens ou ils seront également porteurs du rythme du message. Cette similarité de concepts entre les deux domaines (musical et verbal) renforce notre conviction d'utiliser une approche commune permettant d'extraire une information commune tout en étant porteuse d'un sens différent.

La notion de prosodie revêt également une information mélodique dans le sens

où la définition renvoie également à des notions de variations de fréquence fondamentale. Bien que nous ne traitons pas ce phénomène dans notre analyse, d'autres systèmes proposés dans la littérature, et spécifiquement adaptés au domaine de l'analyse de la parole, se servent de cette information.

Afin de contextualiser notre travail, voici quelques méthodes utilisant une estimation de la prosodie selon différentes approches et ce, dans différents buts.

Le système présenté dans l'article de Bartkova [18] vise à extraire une structure prosodique. Cette structure prosodique est définie comme étant une composition hiérarchique de zones accentuées de la parole. Ces zones étant accentuées différemment, leur enchaînement est révélateur d'une structure. L'hypothèse est faite que la structure prosodique ainsi détectée est directement liée à la structure syntaxique de la phrase. Cette analyse de la parole propose donc de renforcer les informations issues d'un système de transcription de la parole afin d'y ajouter une structuration syntaxique améliorant sensiblement sa compréhensibilité. A partir d'une transcription, les durées et énergies moyennes des voyelles situées en fin de mots sont analysées et normalisées par les durées et énergies moyennes des autres voyelles précédentes . Ce qui permet de normaliser les tailles et énergies par rapport au contexte. La pente d'évolution de la fréquence fondamentale ainsi que le delta de f_0 entre le début et la fin avec les voyelles non terminales sont également utilisées comme descripteur phonétique. Différents marqueurs de frontières prosodiques sont estimés à partir de la variation prosodique. De cette segmentation est ensuite extraite, par différentes règles, la structure syntaxique qui l'a générée.

D'autres méthodes d'analyse recherchent via la prosodie une estimation de l'état du locuteur.

Dans des études telles que celles de [19] ou celles citées dans la revue de [20], il est montré que, grâce à des paramètres prosodiques, l'interaction homme-machine peut être améliorée. En effet, si la prosodie peut être utilisée pour une meilleure compréhension des messages, en tant que reflet des émotions des locuteurs, elle offre également de nombreuses informations *para-linguistiques*. L'objectif de ces méthodes est ainsi, par exemple, de pouvoir estimer si le locuteur a parfaitement saisi une question ou s'il a besoin d'éclaircissements en détectant le doute dans sa voix. Afin d'atteindre ces différentes informations de *feedback*, de nombreux paramètres sont extraits comme le pitch, des informations timbrales ou encore différentes valeurs statistiques issues de l'énergie. Ces paramètres tendent à réagir différemment en fonction de l'état émotionnel du locuteur pouvant agir directement sur les organes de production de la parole. Signe qu'un intérêt grandissant est apporté à ce genre de méthodes au sein de la communauté du traitement de la parole, un effort de normalisation des corpora et des métriques utilisées a été fourni [21]. Cet effort permet à ces méthodes de se comparer plus facilement et ainsi d'optimiser les efforts de recherche de toute la communauté.

Même si cette autre piste est moins explorée, les paramètres prosodiques peuvent également être utilisés pour l'identification de la langue [22]. Les paramètres prosodiques reflétant à la fois les règles syntaxiques et les différentes accentuations propres à une langue, ils peuvent se révéler discriminants dans la différentiation de certaines langues. La prosodie peut d'ailleurs se révéler primordiale dans la compréhension d'une langue et sa mauvaise utilisation peut devenir un réel problème. Dans un but d'utilisation de la prosodie pour l'apprentissage des langues, certaines études ont entamé un travail d'annotation de la prosodie. L'étude de Nakamura [23] propose par exemple une annotation du japonais en fonction du sexe et du contexte afin d'améliorer l'apprentissage de cette langue.

En résumé, si la définition du terme de prosodie reste à préciser, les informations de rythme, de hauteur ou de longueur d'unité de signal peuvent être utilisées pour l'extraction de nombreuses informations du signal de parole. De la syntaxe aux informations para-linguistiques, elles servent à préciser l'information et à donner de la structure et un supplément de sens au simple contenu énoncé.

Nous avons vu que les aspects rythmiques sont actuellement utilisés pour l'extraction d'informations pertinentes en parole comme en musique. À notre connaissance, il n'existe pas de méthode servant à extraire une information sur ces deux types de contenu. La méthode que nous présentons maintenant vise à utiliser une approche générique en montrant que l'information à extraire se retrouve de façon identique dans ces deux types de contenu.

3.3 Définition d'un nouveau descripteur rythmique : le Spectre de Rythme

Lors de la recherche des composantes primaires, nous utilisons des descripteurs qui intègrent une certaine dimension temporelle du signal : la modulation de l'énergie à 4 Hz pour traduire le rythme syllabique de la parole, le nombre de segments stationnaires par unité de temps pour discriminer la parole (nombre de phonèmes par unité de temps) de la musique (nombre de notes par unité de temps). Il est clair qu'une segmentation en unités stables au sens traitement de signal donne par conséquent des indications sur le rythme et par conséquent le contenu du document audio. Nous avons naturellement exploré cette piste et proposé de décrire une zone sonore par l'analyse des position des frontières. Cette approche conduit à une caractérisation plus fine des contenus audio que ce soit en musique ou en parole. Dans cette section nous détaillons les différentes étapes de notre système d'analyse du rythme. Cette présentation se fait en cinq parties. Après un rappel de la méthode de segmentation du signal qui est au centre de notre étude (1ère partie), nous analysons la sémantique des segments obtenus que ce soit en con-

texte de musique ou de parole (2ème partie). La troisième partie présente l'analyse fréquentielle de la répartition des segments que nous avons développée pour l'estimation du rythme. La quatrième partie propose une amélioration de cette analyse par un système de pondération des frontières, pour définir le *Spectre de Rythme*.

3.3.1 Segmentation en zones homogènes

La méthode que nous avons choisie pour l'extraction des informations temporelles de base est la méthode de divergence forward-backward [2]. Cette méthode est conçue pour la segmentation de la parole en zones homogènes sub-phonétiques.

Elle s'appuie sur une description du signal y_n comme suite de zones quasi-stationnaires ; chacune est caractérisée par un modèle auto-régressif gaussien :

$$\begin{cases} y_n = \sum a_i y_{n-i} + e_n \\ var(e_n) = \sigma_n^2 \end{cases} \tag{3.1}$$

avec y_n le signal à traiter et e_n un bruit blanc gaussien.

Chaque zone quasi-stationnaire peut donc être modélisée par le vecteur $M(A, \sigma)$ suivant :

$$M(A, \sigma) = (a_1, \ldots, a_p, \sigma) \tag{3.2}$$

Le problème consiste à détecter des changements dans le modèle et repérer les instants de ces changements. Pour cela, à chaque échantillon, la distance de Kullback-Leiber est calculée entre un modèle long terme M_1, estimé depuis la dernière rupture jusqu'à un instant n_t et un modèle court terme M_2 de longueur L estimé entre $n_t - L$ et L est calculé (figure 3.2).

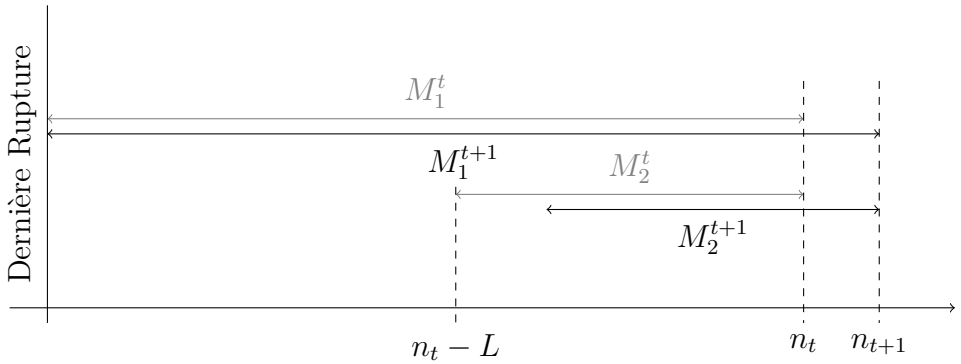


FIGURE 3.2 – Identification des modèles long et court terme (resp. M_1 et M_2) sur 2 pas d'analyse consécutifs.

Les modèles sont initialisés sur une fenêtre de L_{init} échantillons. Cette durée initiale de création du modèle fixe une limite minimale à la durée des segments. Ceci offre aussi l'avantage d'éviter une trop grande sur-segmentation du signal qui perdrait beaucoup de son sens.

La divergence w_n est calculée pour chaque pas entre les deux modèles. Le test est la somme cumulée des divergences :

$$W_n = \sum_{k=1}^{n} (w_k + \delta) \tag{3.3}$$

Un biais δ est introduit pour faciliter la détection de l'instant de rupture. Une rupture est localisée par un changement significatif de pente. Pour ce faire, le maximum courant $W_{n_{max}}$ est comparé à la valeur courante W_n ; un seuil λ est introduit et une rupture est détectée dès lors que

$$W_{n_{max}} - W_n > \lambda \tag{3.4}$$

L'instant de rupture validée est $r_i = n_{max}$ (figure 3.3).

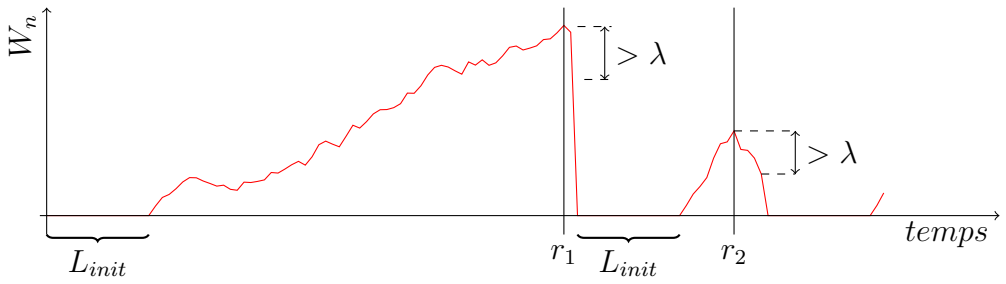


FIGURE 3.3 – Position des ruptures par rapport à la statistique W_n. Chaque frontière est placée sur un maximum local de la divergence. La valeur de λ permet d'autoriser de petites fluctuations dans la pente d'ascension de W_n.

Cette méthode est effectuée dans le sens du signal. Mais, le calcul de la divergence n'étant pas symétrique, certaines ruptures peuvent être oubliées dans cette analyse. Pour éviter ce problème, une seconde passe peut être effectuée en cas de suspicion d'oubli. Cette analyse en retour arrière est utilisée en cas de segment **suffisamment long** :

$$r_i - r_{i-1} > L_{max} \tag{3.5}$$

Deux cas de figures peuvent alors se produire :

– soit aucune frontière n'est détectée, alors l'analyse continue dans le sens du signal, à partir de la frontière r_i.

– soit une frontière est trouvée à un temps r'_i alors cette frontière est validée et l'analyse continue dans le sens du signal à partir de cet instant. La frontière située à r_i est invalidée.

Cette analyse découpe le signal en segments homogènes. Grâce à ce découpage en unités minimales de temps, nous pouvons déduire une information de débit en étudiant la position de ces différentes frontières.

3.3.2 Quelle sémantique pour les segments ?

En fonction du contenu (parole ou musique), la sémantique n'est pas la même et il est intéressant d'en observer les résultats.

En parole

Cette approche a initialement été conçue pour la segmentation des signaux de parole, c'est donc sans surprise que le résultat de la segmentation est une segmentation sub-phonétique. Elle découpe les phonèmes en leurs différentes phases de réalisation : transitions et phases stables. Les frontières correspondent à des variations spectrales : hausse de l'énergie, modification de la structure harmonique ou formantique...

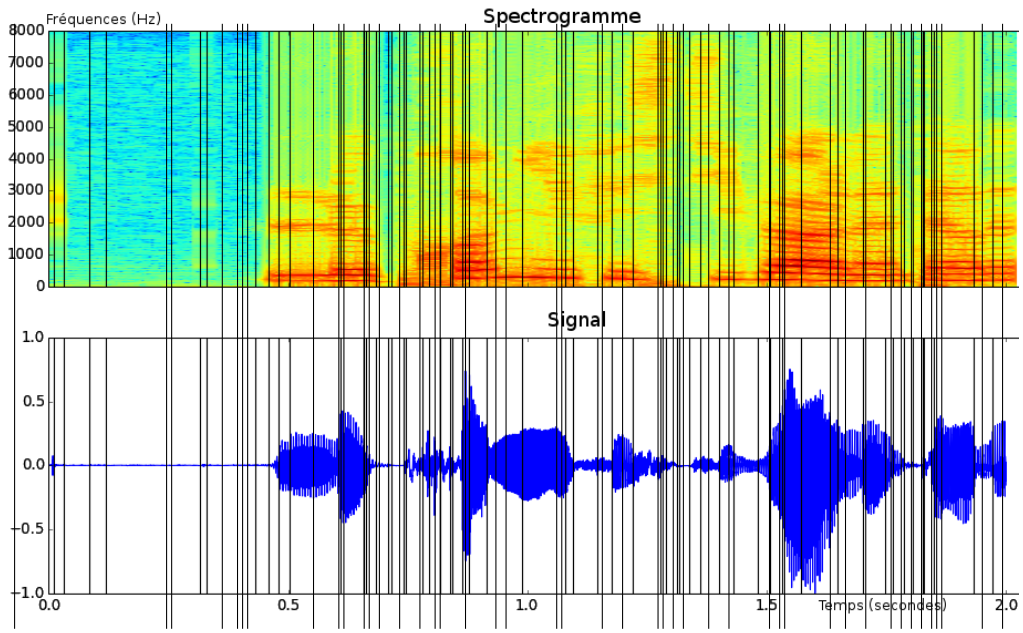


FIGURE 3.4 – Segmentation de 2 secondes de parole lue par une femme. La lectrice prononce « *Le premier festival inter...* » Les différentes phases de la parole sont retrouvés : explosion, variation d'énergie ou de motif harmonique...

Les segments sont relativement courts comme le signal de parole évolue rapidement. La figure 3.4 illustre le résultat d'un découpage par cette méthode sur 2 secondes de signal de parole lue.

En musique

L'utilisation de cette segmentation dans des contenus de musique apporte également des informations intéressantes. De par sa conception et ses propriétés, cette segmentation est capable de découper un signal correspondant à une note en ses différentes phases.

Les phases d'une note peuvent être explicitées par les différentes phases de variation de l'enveloppe du signal :

- l'augmentation rapide de l'énergie du signal due à la production de la note : l'*attack*,
- la petite décroissance de l'énergie vers un régime constant : le *decay*,
- la stabilité de l'énergie pendant sa résonance : le *sustain*,
- la diminution progressive de l'énergie : la *release*.

Il est intéressant de constater que la durée de chacune de ces phases peut beaucoup varier et que cette information est liée au type d'instrument utilisé. La figure 3.5 schématise la variation d'énergie lors de ces différentes phases, à partir de l'*onset* (le début de la note).

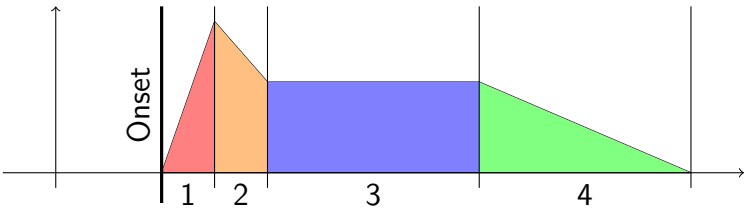


FIGURE 3.5 – Schéma des phases d'évolution de l'enveloppe du signal d'une note. Dans l'ordre les différentes phases sont notées : *attack* (1), *decay* (2), *sustain* (3) et *release* (4).

Cette division du signal de musique est, de manière analogue à la segmentation de parole, composée de segments minimaux. Nous analyserons ensuite la répartition de la position de ces segments afin d'y déceler une structure temporelle.

3.3.3 Analyse fréquentielle

Pour analyser la répartition fréquentielle des frontières extraites, nous créons un signal $b(t)$ reflétant la segmentation. Ce signal est composé de fonctions de Dirac, chaque Dirac étant positionné à l'instant r_i .

L'expression de $b(t)$ est la suivante :

$$b(t) = \sum_{k=1}^{N} \delta(t - r_k) \tag{3.6}$$

avec N le nombre total de ruptures trouvées.

L'analyse fréquentielle de cette somme de fonctions de Dirac est effectuée. Elle permet de mettre en avant la présence de récurrences fréquentielles dans la position des ruptures et de révéler une information de tempo si elle existe.

La transformée de Fourier $B(f)$ de $b(t)$ s'exprime comme ceci :

$$B(f) = \int_{\mathbb{R}} b(t) e^{-2i\pi ft} dt \tag{3.7}$$

$$= \int_{\mathbb{R}} \sum_{k=1}^{N} \delta(t - r_k) e^{-2i\pi ft} dt \tag{3.8}$$

En utilisant les propriétés de la transformée de Fourier d'une fonction de Dirac, à savoir que :

$$\int_{\mathbb{R}} \delta(t - r_k) e^{-2i\pi ft} dt = e^{-2i\pi f r_k} \tag{3.9}$$

l'expression de $B(f)$ se simplifie :

$$B(f) = \sum_{k=1}^{N} e^{-2i\pi f r_k} \tag{3.10}$$

Cette formule offre l'avantage d'être simple et donc rapide à calculer, puisqu'il s'agit d'une simple somme. Elle donne également la complète maîtrise du domaine fréquentiel que nous souhaitons analyser. Cela permet de réaliser des analyses temporelles plus ou moins précises en fonction des besoins.

La figure 3.6 présente le résultat de l'analyse d'un enregistrement de musique Pop annoté manuellement à 100 bpm. Peu de pics sortent réellement de la valeur moyenne à l'exception du plus grand pics situé à 50 bpm. Malgré la dominance de ce pic lié au tempo, ce résultat n'est pas satisfaisant pour une bonne détection car la valeur du tempo ne ressort pas. En effet, des pics non liés à ce tempo ont des amplitudes proches du seul pic d'importance liés à la valeur annotée.

Ce phénomène de non prédominance des valeurs liées au tempo s'explique par le fait que toutes les frontières identifiées sont considérés de la même façon, qu'elles marquent le début ou la fin d'un phénomène. Afin d'améliorer notre système, nous avons mis en place un système de pondération des frontières pour mettre en relief les frontières situées en début de phénomène, plus porteuses de structure. En musique, elles marquent l'arrivée d'une nouvelle note et sont donc plus représentatives de la structure du morceau.

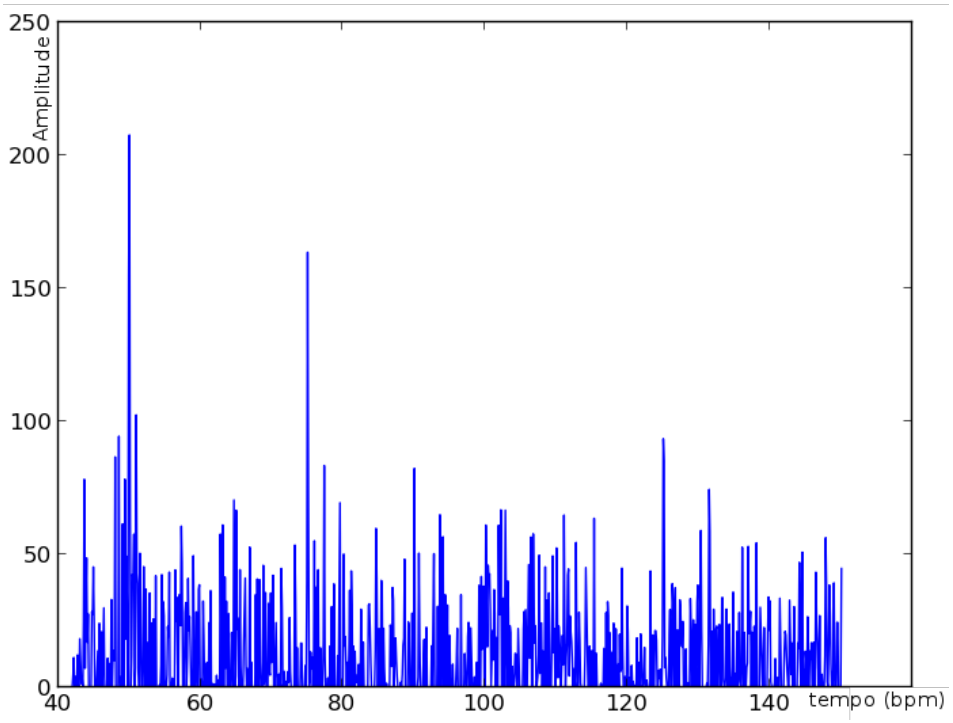


FIGURE 3.6 – Analyse d'un morceau de trois minutes de musique Pop annoté à 100 bpm. Une valeur prédomine à 50 BPM, correspondant à la moitié de l'annotation manuelle en tempo.

3.3.4 Pondération

En parole comme en musique, nous cherchons à mettre en évidence une structure temporelle parmi les positions des ruptures du signal. Afin d'éliminer au maximum le bruit généré par les ruptures détectés dans les phases internes du phonème de la note, nous proposons une approche consistant à pondérer les frontières en fonction de leur importance dans la structure temporelle du signal. Nous partons de l'hypothèse que les frontières les plus importantes sont celles situées en début de syllabes ou de notes. C'est en effet la partie d'une note ou d'une syllabe la plus simple à positionner consciemment afin de créer un effet de rythme.

Ces frontières intéressantes ont pour caractéristiques communes d'être des instants subissant une hausse d'énergie. Nous proposons un système de pondération des frontières basé sur une différence locale d'énergie.

Nous utilisons le calcul de l'énergie RMS définie comme suit :

$$E(t_{debut}, t_{fin}) = \sqrt{\frac{\sum_{i=t_{debut}}^{t_{fin}} y(i)^2}{t_{fin} - t_{debut}}} \tag{3.11}$$

La pondération $Poids(r_t)$ de la frontière située à l'instant r_t se calcule avec la formule suivante :

$$Poids(r_t) = E(r_t, r_t + N_{energ}) - E(r_t - N_{energ}, r_t) \tag{3.12}$$

avec N_{energ} la longueur de la fenêtre utilisée pour le calcul de l'énergie.

Cette pondération est effectuée sur toutes les frontières. Notons que les poids ainsi obtenus peuvent être négatifs. Nous choisissons de conserver ces poids négatifs qui permettent comparativement de rehausser l'importance des frontières positives. Ce principe est illustré par la figure 3.7 où le poids $Poids(r_k)$ est construit par soustraction de l'énergie avant la rupture (en rouge) à celle après (en bleu).

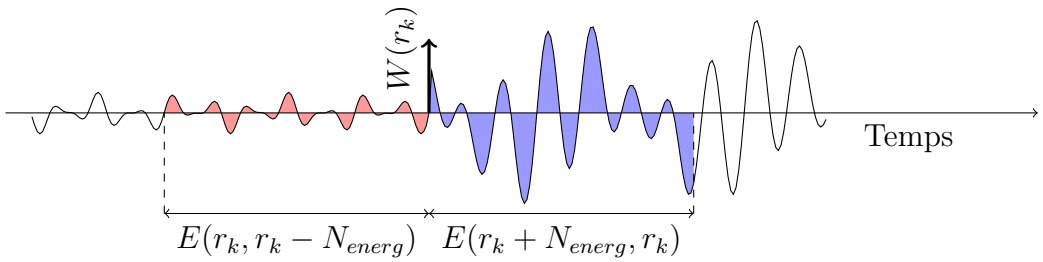


FIGURE 3.7 – Chaque frontière est pondérée par la différence entre l'énergie RMS de N_{energ} échantillons avant et après elle.

Cette pondération est ensuite prise en compte dans le calcul du signal $b(t)$ de l'équation 3.6 qui devient :

$$b_p(t) = \sum_{k=1}^{N} \delta(t - r_k) Poids(r_k) \tag{3.13}$$

La figure 3.8 présente $b_p(t)$ pour la segmentation présentée en figure 3.4.

Les frontières situées en forte augmentation énergétique se trouvent fortement pondérées et mises en valeur. Elles correspondent à des débuts de voyelles et localisent donc une syllabe.

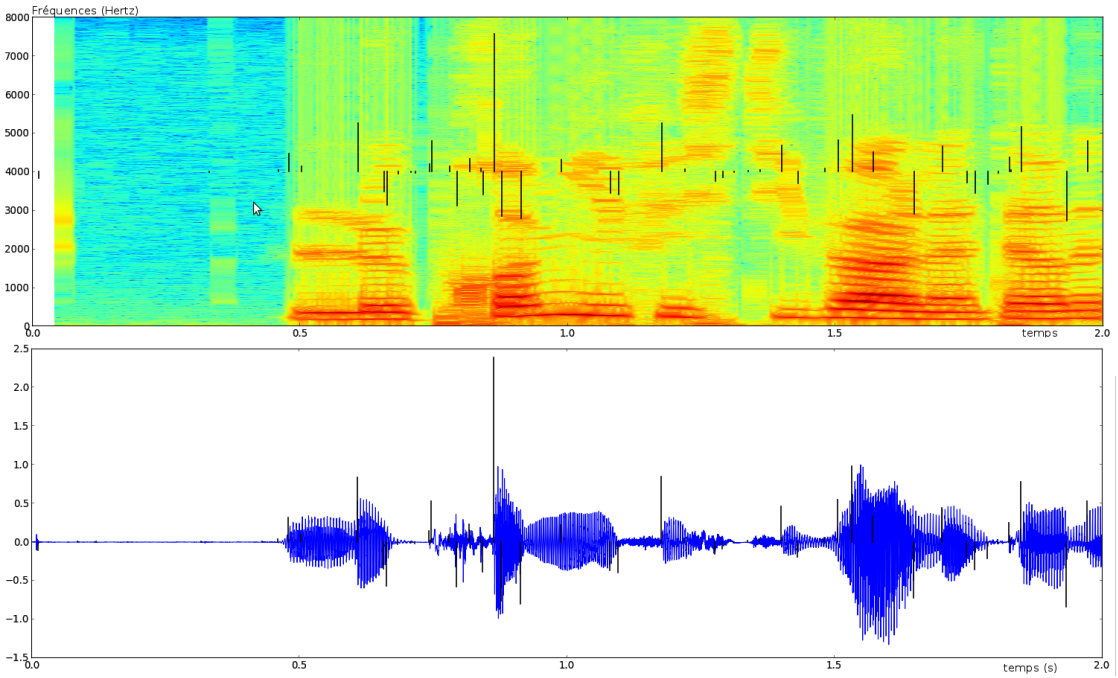


FIGURE 3.8 – Segmentation pondérée sur l'exemple de parole présenté en figure 3.4. Les débuts de voyelle génèrent des frontières aux poids les plus importants alors que les autres frontières ont des poids faibles voire négatifs.

Cet ajout d'une pondération se propage également à l'expression de $B(f)$ dans les équations 3.8 et 3.10 :

$$B_p(f) = \int_{\mathbb{R}} \sum_{k=1}^{N} \delta(t - r_k) e^{-2i\pi ft} Poids(r_k) dt \tag{3.14}$$

$$= \sum_{k=1}^{N} e^{-2i\pi f r_k} Poids(r_k) \tag{3.15}$$

L'ensemble des valeurs $(B_p(f), f = 1, f_{max})$ est le descripteur **Spectre de Rythme** qui nous sert désormais d'information de base pour atteindre une classification de type rythmique de la zone sonore étudiée. Cette formule garde l'avantage d'être très rapide à calculer et assure un contrôle total et facile sur le domaine de définition de la fonction aussi bien sur les bornes d'analyse f_{max} que nous appellerons plage d'analyse, que sur la résolution.

La figure 3.9 illustre clairement le gain produit par la pondération sur le même fichier que celui présenté par la figure 3.6, annoté à 100 bpm. Sur cette nouvelle analyse, seuls quelques pics ressortent clairement avec une énergie au moins 5

fois supérieure à la moyenne. Ces pics sont tous liés à la valeur annoté du tempo puisque les trois plus grands pics correspondent aux valeurs exactes, moitiés et doubles de celui-ci.

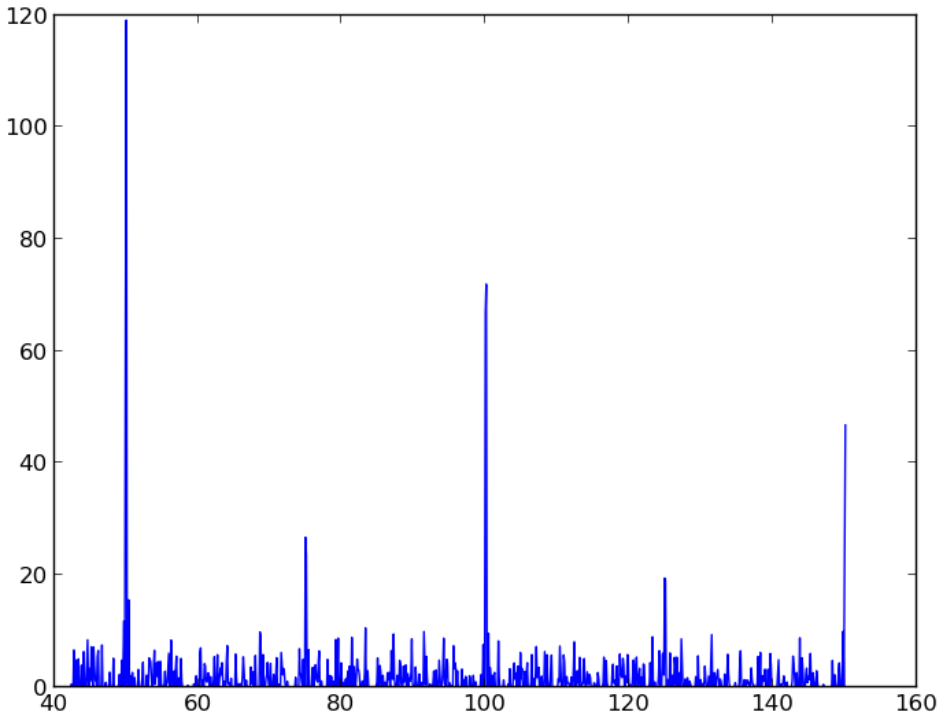


FIGURE 3.9 – Analyse automatique du même morceau que celui de la figure 3.6. Cette fois-ci, les pics principaux sont tous clairement prédominants et liés à la valeur du tempo obtenue manuellement (100 bpm).

Nous allons maintenant présenter deux applications de notre méthode pour deux objectifs différents : l'**estimation de tempo** et l'**exploration du continuum entre voix parlée et voix chantée**. Si ces deux applications sont présentées respectivement sur des contenus de musique et de parole, cette séparation n'est motivée que par l'objectif sur lequel nous nous focalisons : et ces deux applications sont facilement transposables d'un type de contenu à l'autre.

3.4 Traitement de la Musique : Estimation de Tempo

3.4.1 Principe

En musique, le rythme est clairement une information essentielle pour la structure d'un morceau. Notre approche, comme le montre l'exemple donné ci-dessus, semble révéler cette structure temporelle.

Si nous considérons un morceau à tempo fixe, cette analyse peut être utilisée sur l'ensemble de l'enregistrement en précisant son domaine de définition par une information *a priori* sur l'intervalle des *tempi* plausibles. En musique occidentale, cette plage se situe généralement à des fréquences situées entre 4 et 0,5 Hz (30 à 240 pulsations par minute).

L'analyse de fichiers musicaux montre la présence de pics clairs aux valeurs liées au tempo. Ces valeurs ne sont pas toujours les valeurs annotées comme étant le tempo de référence du morceau, mais elles se rapportent généralement au concept de *tatum* (plus mesurable) correspondant à l'intervalle de temps (entre notes) le plus présent au sein du morceau.

Nous proposons une méthode pour le choix du pic correspondant au bon tempo. Deux approches différentes sont utilisées : la méthode inter-frontières et la méthode des peignes.

Ces méthodes cherchent à choisir quelle valeur de tempo est la valeur correcte parmi les pics les plus significatifs du descripteur Spectre de Rythme. Comme lors de chaque analyse, seulement quelques pics possèdent réellement une plus grande amplitude que les autres, nous cherchons à distinguer le pic lié au tempo parmi les **4** pics de plus grande amplitude. Ce nombre a été choisi empiriquement, après avoir constaté que dans le pire des cas de notre analyse, le tempo correct correspond au quatrième pic d'importance. Soit P la liste ordonnée des quatre plus grands pics du spectre avec :

$$B_p(p_i) > B_p(p_{i+1}) \tag{3.16}$$

pour $i = 1, ..., 4$.

Méthode inter-frontières

Pour cette approche, nous reprenons les frontières possédant les poids les plus élevés. Pour cela nous utilisons un seuil $Poids_{min}$ afin d'éliminer les frontières trop faibles. Cette méthode permet de ne garder que les frontières qui sont les plus probablement situées sur les *onsets* des notes. À partir de ces frontières la liste J des intervalles entre frontières est calculée.

Pour chaque fréquence p_i nous calculons la période liée $L_i = \frac{1}{p_i}$ ainsi que les valeurs de $\frac{L_i}{4}, \frac{L_i}{3}, \frac{L_i}{2}, 2L_i$ et $3L_i$. Nous effectuons le comptage de tous les intervalles de J dont la longueur correspond à une des ces 6 valeurs, soit $Num(p_i)$ ce nombre.

La mise en œuvre de ce calcul nécessite d'introduire une tolérance δ. Une longueur d'intervalle correspond à une des 6 valeurs, si sa différence avec cette valeur est inférieure à ce seuil.

Pour finir, nous estimons le tempo correct $\hat{p}$ par le tempo étant le plus en lien avec les intervalles inter-frontières.

$$\hat{p} = \arg\max_{p_i} Num(p_i) \tag{3.17}$$

Méthode Peigne

Pour cette autre approche nous utilisons la convolution entre un Peigne Harmonique de Dirac PH et le Spectre de Rythme. Ce peigne est paramétré par un tempo *tempo* et comporte N_d dents espacées de *tempo*.

Le critère de choix *Amp* que nous utilisons est le suivant :

$$Amp(tempo) = \frac{B_p * PH(tempo)}{7} \tag{3.18}$$

Une tolérance autour de chaque dent est également fixée pour considérer l'énergie présente dans un petit voisinage autour d'elle. Ce voisinage est fixé à $\pm 1BPM$ dans la mesure où la décision est retournée en tant que valeur entière en BPM.

Pour chaque analyse, sept peignes sont utilisés qui correspondent aux sept valeurs définies de la manière suivante : la valeur de p_1 qui correspond au pic de plus grande amplitude, toujours lié à la valeur correcte du tempo et les 6 autres candidats, respectivement : $\frac{p_1}{4}, \frac{p_1}{3}, \frac{p_1}{2}, 2p_1, 3p_1$ et $4p_1$.

Le choix du tempo $\hat{p}$ s'effectue avec la valeur de tempo retournant la valeur de *Amp* la plus élevée.

$$\hat{p} = \arg\max_{tempo} Amp(tempo) \tag{3.19}$$

Fusion

Afin de profiter au mieux des caractéristiques de ces deux méthodes de décision, nous proposons une fusion simple des résultats. Pour cette fusion nous utilisons l'enchaînement des deux méthodes. Les expériences préliminaires que nous avons effectuées montrant que l'utilisation de la méthode des peignes retourne la valeur correcte du tempo au pire des cas en deuxième position, nous utilisons celle-ci comme première passe du système de fusion. Les deux *tempi* avec la plus forte

valeur de *Amp* par la méthode des peignes sont utilisés comme candidats potentiels. La décision entre les deux est ensuite effectuée par la technique inter-frontières. La figure 3.10 présente le diagramme du système de fusion que nous utilisons.

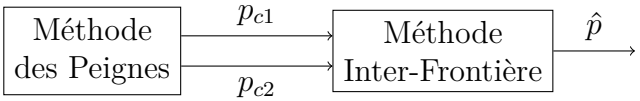


FIGURE 3.10 – Diagramme de flux de la fusion entre les deux approches. Seules les deux meilleurs *tempi* de la méthode des peignes sont utilisés en entrée de la méthode inter-frontières.

3.4.2 Évaluation

L'expérience que nous avons menée est décrite dans [24] et répond à la tâche classique d'estimation du tempo dans des enregistrements musicaux, c'est-à-dire à la détermination d'un unique tempo sur tout l'enregistrement analysé ; le tempo est considéré fixe au sein d'un même enregistrement. Les spécifications d'évaluation que nous avons utilisées correspondent à celles utilisées dans la littérature [6].

Un tempo est jugé correct si il diffère de moins de 4 bpm de la valeur annotée.

Une première métrique est proposée, nommée $Accuracy_1$. Elle s'exprime de la façon suivante :

$$Accuracy_1 = \frac{|tempi\ corrects|}{|pistes\ analyses|} \tag{3.20}$$

Une seconde métrique, plus souple, est proposée afin de considérer comme correctes les valeurs multiples. Une tolérance sur la valeur est également utilisée. Les multiples et ratios estimés sont considérés comme corrects si leur distance à 2, 3, 4, 1/3 et 1/2 est inférieure à 0,03.

Cette seconde métrique $Accuracy_2$ s'exprime alors comme suit :

$$Accuracy_2 = \frac{|tempi\ correct\ ou\ multiples|}{|pistes\ analyses|} \tag{3.21}$$

Corpus

Pour notre expérience nous utilisons la base RWC [25][26], et plus précisément, sa partie annotée en tempo. Il s'agit de la partie consacrée à la musique Pop. Cette base a été conçue afin de fournir une référence pour l'analyse de méthodes d'extraction d'information musicale. À ce titre, il est intéressant de l'utiliser afin de comparer notre approche avec celles d'autres algorithmes de la littérature.

La partie Pop que nous avons utilisée contient 100 pistes de chanson Pop japonaises d'une durée allant de 2 minutes 50 secondes à 6 minutes 7 secondes. Les enregistrements ont été convertis au format WAVE, échantillonnés à 16 kHz, 16 bits par échantillons et mono canal.

Paramétrisation

Pour notre analyse les différents paramètres ont été fixés aux valeurs suivantes :
- $N_{energie}$ la longueur de la fenetre utilisée pour le calcul de l'énergie est fixé à $0,1s$,
- N_d le nombre de dents pour l'analyse par peigne est fixé à 7.

Résultats

Afin de mesurer le gain apporté par chacune des méthodes proposées pour l'estimation du tempo, nous avons choisi d'en afficher les résultats séparément.

Tout d'abord, nous nous contentons de sélectionner *le pic maximal comme valeur du tempo*. Le tableau 3.1 présente la répartition des 100 *tempi* estimés en terme de lien avec le tempo annoté.

TABLE 3.1 – Répartition des 100 *tempi* obtenus par une décision basée uniquement sur la localisation du plus grand pic.

Rapport avec le tempo annoté	1/2	1	2	3	4	Pas de lien
Nombre de tempi	2	3	60	0	32	3

Nous pouvons constater deux choses sur cette répartition :
- les *tempi* ont clairement tendance à être surestimés, dans **92%** des cas, la valeur retournée est un multiple entier du tempo (2 ou 4).
- l'information de tempo peut clairement être extraite par notre approche puisque dans **97%** des cas, la position du plus grand pic est en lien avec le tempo annoté.

Si le score d'$Accuracy_2$ est très bon (97%), il est en revanche extrêmement dommage que le score de décision exacte ($Accuracy_1$ de seulement 3%) soit si bas. Nous comptons sur nos méthodes de décision afin d'augmenter le score d'$Accuracy_1$.

Pour la **méthode inter-frontières**, nous devons considérer uniquement les frontières correspondant à de possibles *onsets*. Dans cette perspectives, seules les frontières ayant au minimum un poids de **10% du poids maximal** ont été retenues. Les résultats obtenus sont présentés dans le tableau 3.2 en terme de nombre de *tempi* évalués en rapport avec le tempo de référence.

TABLE 3.2 – Répartition des *tempi* pour la méthode de décision inter-frontières, basée sur l'analyse des intervalles.

Rapport avec le tempo annoté	1/2	1	2	3	4	Pas de lien
Nombre de *tempi*	7	56	28	0	1	8

Nous remarquons qu'un grand nombre de *tempi* estimés comme double, sont maintenant corrects et que ceux identifiés comme quadruples sont maintenant doubles.

Cette répartition des résultats conduit aux mesures d'*Accuracy* suivantes : $Accuracy_1 = 56\%$ et $Accuracy_2 = 92\%$.

Cette solution augmente nettement la valeur de l'$Accuracy_1$ tout en gardant un score d'$Accuracy_2$ élevé même si il baisse. Néanmoins, beaucoup de *tempi* restent évalués au double de leur valeur.

Pour la *méthode des peignes*, la répartition des tempi est proche de celle de la méthode inter-frontières(tableau 3.3) ; cependant moins de *tempi* estimés ont une valeur plus faible que le tempo annoté. Cette méthode donne donc de meilleurs résultats que la méthode basée sur le calcul des intervalles : $Accuracy_1 = 64\%$ et $Accuracy_2 = 96\%$.

TABLE 3.3 – Répartition des *tempi* par la méthode des peignes.

Rapport avec le tempo annoté	1/2	1	2	3	4	Pas de lien
Nombre de tempi	3	64	29	0	0	4

Il est également à noter que si nous retournons les valeurs des deux *tempi* les plus probables selon cette méthode, dans 98% des cas, le tempo annoté est présent.

C'est cette particularité que nous tentons d'exploiter dans la *méthode de fusion* que nous proposons. La répartition est présentée dans le tableau 3.4.

TABLE 3.4 – Répartition des *tempi* estimés par la fusion des méthodes.

Rapport avec le tempo annoté	1/2	1	2	3	4	Pas de lien
Nombre de tempi	13	78	2	0	0	7

Le nombre de *tempi* correctement estimés augmente largement la valeur de l'$Accuracy_1$ avec un score de 78%. Cette augmentation se fait en revanche au prix d'une petite baisse de l'$Accuracy_2$ avec une valeur de 93%.

Ces chiffres sont révélateurs d'erreurs de décision prises par la méthode inter-frontières et il est raisonnable d'envisager de la perfectionner compte-tenu de la marge de progression en terme d'$Accuracy_1$.

En résumé, la comparaison des résultats présentées dans les tableaux 3.1 à 3.4 est proposée dans la figure 3.11. Dans ce graphique, on distingue clairement la progression d'estimation des *tempi* corrects gagnée sur une forte diminution du nombre d'erreur d'estimation double .

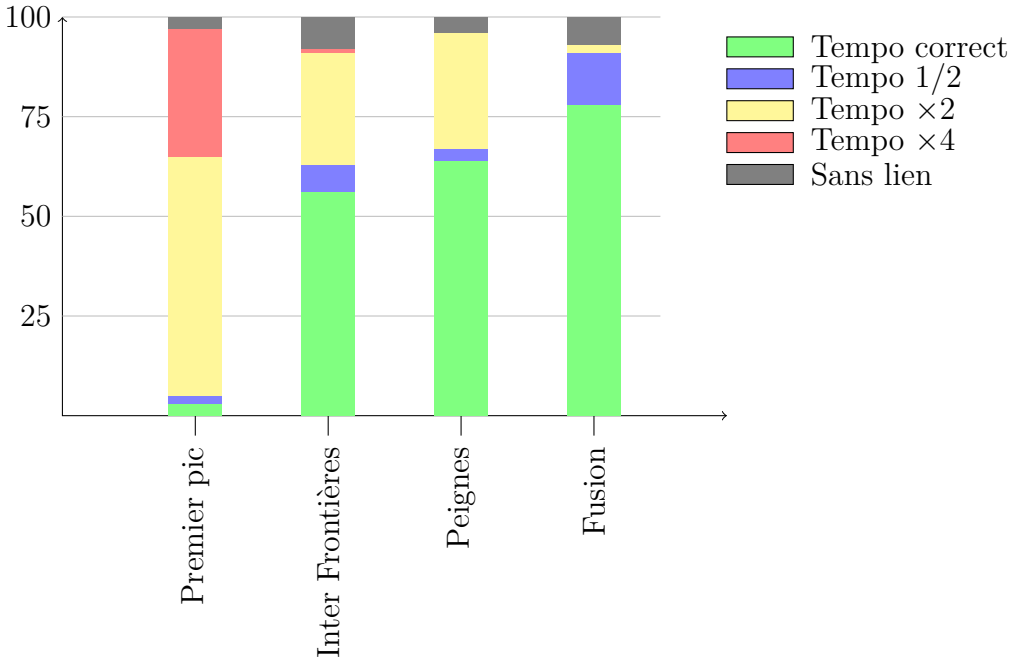


FIGURE 3.11 – Répartition des résultats en nombre d'enregistrements par lien avec le tempo annoté pour les différentes méthodes d'estimation de tempo.

Discussion

Ces résultats sont à mettre en perspective avec les résultats obtenus par d'autres algorithmes sur des contenus similaires. De nombreux algorithmes présentés dans notre contextualisation ont participé à la campagne d'évaluation MIREX 2004 [6], pour laquelle les contenus sont proches des nôtres. Ce corpus contenait des extraits musicaux d'une durée comprise entre 2 et 30 secondes de différents genres musicaux.

Le meilleur algorithme de la campagne est celui de Klapuri [9] et il a obtenu des scores d'$Accuracy_1$ de 67,29% et d'$Accuracy_2$ de 85,01% ; il obtient un maximum d'$Accuracy_2$ à 91,18% sur le sous-ensemble composé uniquement de chansons. Une

fusion par vote des différents algorithmes a également été testée avec des résultats proches : $Accuracy_1$ de 68% et $Accuracy_2$ de 86%. Nos résultats permettent donc une bonne estimation du tempo par rapport à ceux cités. Il est néanmoins à remarquer que notre méthode fonctionnerait difficilement sur de très courts extraits de moins d'une dizaine de secondes. Notre étude nécessite en effet un minimum de frontières de forts poids afin d'être statistiquement significative et plus ce nombre est grand, plus le résultat s'en trouve consolidé.

3.5 De la voix parlée à la voix chantée

Si une analyse globale du tempo sur un morceau de musique occidentale peut avoir du sens car la majorité de ces morceaux possèdent un tempo fixe, il en est tout autrement pour l'analyse de musiques plus exotiques ou des enregistrements de parole. Sur ces types de contenus, nous ne pouvons pas faire d'hypothèse de stabilité rythmique. Il nous faut au contraire pouvoir adapter l'analyse à la possibilité d'identifier des variations de rythme. Nous avons pour ce faire défini un nouvel outil de tempogramme, par analogie avec le spectrogramme classique qui permet de suivre les formants et les harmoniques en parole. Cette technique peut donc s'appliquer à l'identification de zones de rythmes stables dans tout type d'enregistrements.

Une telle analyse peut se révéler utile notamment dans l'analyse de zones de paroles particulières, à mi-chemin entre voix parlée et voix chantée. Le suivi de tempo peut révéler des zones de tempo maintenu dans le flot de parole. Ces zones sont révélatrices d'une volonté du locuteur de donner une structure rythmique à son discours, et elles peuvent donner un fort indice quant à la nature de la voix et sa catégorisation entre voix parlée et voix chantée, par exemple : récitée, scandée, déclamée...

Cet objectif de déceler des catégories intermédiaires entre parole et musique est un objectif qui n'a à notre sens jamais été visé en tant que tel. Il s'agit donc de réaliser de façon exploratoire l'analyse rythmique de segments de parole et d'y rechercher un éventuel motif, signe d'un type de parole particulier. Nous avons procédé à des expériences pour valider notre approche sur des exemples de données de différents types.

3.5.1 Le Tempogramme

Pour réaliser un tempogramme, nous utilisons une approche voisine de l'établissement d'un spectrogramme. Sur l'ensemble de l'enregistrement audio, l'analyse de tempo est réalisée sur une fenêtre glissante $Ftemp$ de longueur N_{Ftemp} et de pas S_{Ftemp}. Nous nommons *tempogramme* l'ensemble des spectres de Rythme (fig-

ure 3.12). Chaque Spectre de rythme correspond à une colonne d'une matrice et chaque colonne correspond à une fenêtre d'analyse ; deux analyses sont espacées d'une durée S_{Ftemp}.

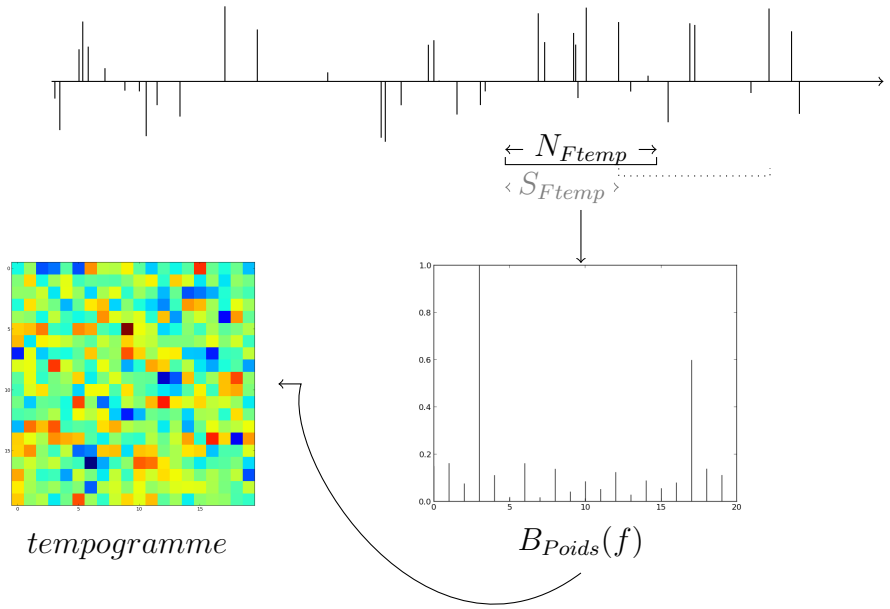


FIGURE 3.12 – Illustration du calcul d'un *tempogramme*.

Il est important, dans cette analyse, que la valeur de N_{Ftemp} soit suffisamment élevée, en effet, il est indispensable que le nombre de frontières à analyser soit statistiquement significatif. Afin d'obtenir une précision plus grande, nous pourrons jouer sur la valeur de S_{Ftemp} afin de permettre un plus grand recouvrement entre les fenêtres d'analyse. Ces paramètres de la fenêtre d'analyse sont liés au type d'application ; à titre indicatif, une durée de 10 secondes est préconisée.

Les phénomènes observés sur le *tempogramme* révèlent la présence ou l'absence de structures temporelles répétitives. Cette représentation permet de suivre l'évolution du rythme au cours du temps et se rapproche de travaux de Peeters [27].

Cette analyse se montre effectivement utile dans l'analyse de zones de paroles particulières, à mi-chemin entre voix parlée et voix chantée, comme nous allons le voir ci-après.

En effet, si ces zones ne présentent pas de stabilité rythmique particulière, nous pourrions conclure qu'il ne s'agit pas de musique mais plutôt de parole. L'analyse par tempogramme peut révéler des zones de tempo maintenu dans le flot de parole. Ces zones étant révélatrices d'une volonté du locuteur de donner

une structure rythmique à son discours, ceci est un fort indice qu'il s'agisse de l'une des catégories intermédiaires entre voix parlée et voix chantée, par exemple : récitée, scandée, déclamée...

Il s'agit donc de réaliser de façon exploratoire l'analyse rythmique de segments de parole et d'y rechercher un éventuel motif, signe d'un type de parole particulier. La partie suivante propose des expériences servant à valider notre approche sur des exemples de données de différents types.

3.5.2 Évaluation

Ce travail mené dans le domaine de la parole, est une expérience purement qualitative. Elle vise à illustrer la possibilité d'utiliser notre approche. Notre objectif est la mise en évidence de catégories intermédiaires entre voix parlée et voix chantée. Si la voix chantée peut être identifiée par des techniques proposées au sein de notre équipe, soit par la détection de musique [1], soit par des méthodes spécifiques [3], les distinctions automatiques entre types de voix parlée n'ont encore jamais fait l'objet d'études à notre connaissance.

Nous n'avons donc pas pu nous baser sur des corpus ou des annotations existantes et il est nécessaire de préciser le contexte expérimental.

Corpus

Pour analyser le comportement de notre approche, nous avons étudié 3 enregistrements, tous échantillonnés à 16 kHz, 16 bits par échantillons et mono canal.

- Le premier enregistrement $f1.wav$ est un extrait de 3 minutes d'un programme TV. Cet enregistrement est issu de la campagne d'évaluation française ETAPE [28]. Différents intervenants discutent et conversent de manière libre, sans lecture de notes. Cet exemple illustre à notre sens, la parole commune, non préparée et peu cadrée.
- Le deuxième enregistrement $f2.wav$, dure 30 secondes et provient du corpus BREF [29]. Il s'agit d'un enregistrement dans lequel une lectrice lit des passages du journal *Le Monde* en mettant le ton. C'est l'exemple représentatif de la parole maîtrisée et préparée.
- Le troisième enregistrement $f3.wav$ est un extrait du corpus DIADEMS. Il s'agit d'un enregistrement ethnomusicologique, d'une longueur de 3 minutes dans lequel un conte africain est enregistré. Ce récit est scandé et une accentuation régulière est mise dans sa prononciation. Il ne s'agit pas de musique, mais c'est clairement un extrait où la mise en rythme est importante.

Ces trois extraits illustrent trois réalités distinctes de la notion de parole, de la plus fluide et spontanée à la plus cadrée. Nous avons choisi ces différents corpus,

car ils nous semblaient de qualité et représentatifs de différents types de parole ; le corpus **BREF** a été crée pour fournir une base de parole lue et le corpus **ETAPE** une base de parole plus spontanée. L'extrait du corpus **DIADEMS** est quand à lui représentatif d'un cas extrême de parole et représente également un cas concret d'application au sein de ce projet. Nous testons notre analyse sur ces trois fichiers et observons les résultats.

Mise en œuvre du tempogramme

L'analyse de ces trois enregistrements a été réalisée avec le même jeu de paramètres. Les valeurs de N_{Ftemp} et S_{Ftemp} sont fixées respectivement à 10 secondes et 1 seconde.

L'analyse est effectuée entre les fréquences de 0,1 Hz et 2,5 Hz sur 1024 fréquences. Dans cette expérience, nous avons choisi de garder la paramétrisation standard de la méthode de segmentation dans un objectif de généricité.

Analyse des Tempogrammes

Le tempogramme de l'enregistrement $f1.wav$ est présentée dans la figure 3.13. Sur celle-ci, aucune structure particulière ne se distingue, seulement un pêle-mêle de zones de fortes et faibles énergies.

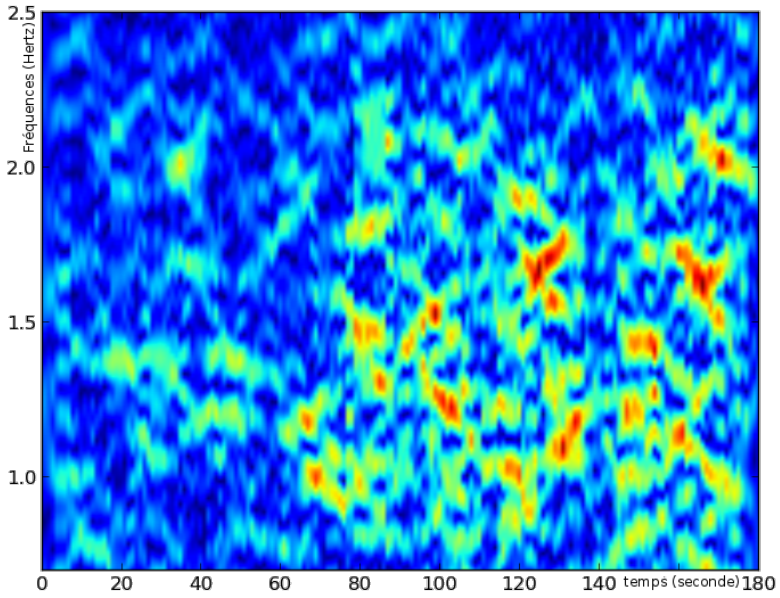


FIGURE 3.13 – Tempogramme de $f1.wav$ (3 minutes). Dans cette analyse de parole non contrainte, aucun motif ne se détache.

Cette absence de structure au sein du *tempogramme* correspond à ce que nous pouvions attendre de l'analyse d'un extrait de parole « *libre* ». L'absence de structure temporelle récurrente dans la production de la parole se retrouve donc sur le tempogramme.

Le tempogramme de l'enregistrement $f2.wav$ est présenté dans la figure 3.14. Sur ce *tempogramme* de parole lue, nous retrouvons également un « méli-mélo » de zones d'amplitudes différentes. En revanche, contrairement à l'enregistrement $f1.wav$, sur la deuxième moitié du signal, une fréquence autour de 1 Hz ressort plus clairement. Les zones de plus faibles énergies dans les plus hautes fréquences semblent également contenir un certain motif harmonique.

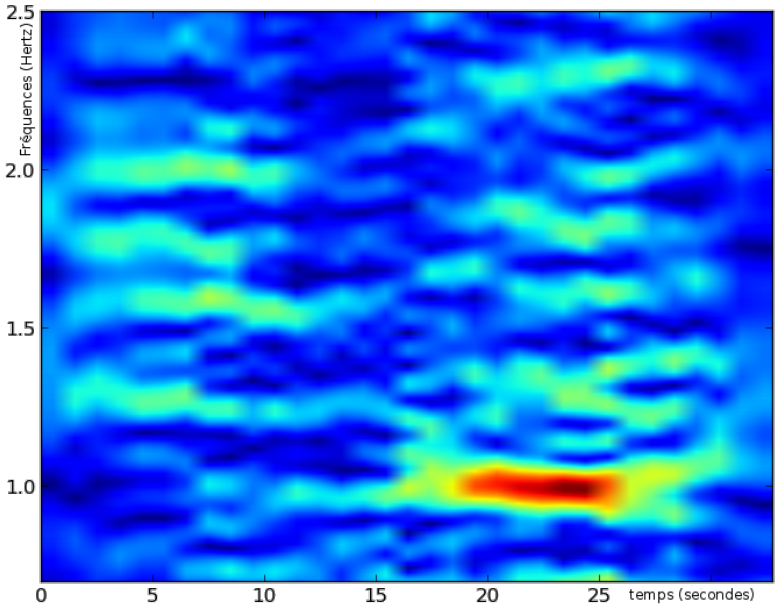


FIGURE 3.14 – Tempogramme du fichier $f2.wav$. Nettement plus court que le précédent (32 secondes), nous y distinguons néanmoins une structure et une fréquence se détachant autour de 1 Hz.

De manière générale, le tempogramme semble plus structuré et reflète, semble-t-il, le caractère plus cadré de la parole lue, d'un ton plus posé et d'un débit plus contrôlé. Il faut néanmoins tenir compte du fait que ce fichier est beaucoup plus court que le précédent et se méfier des effets d'échelle !

Enfin, le tempogramme du troisième enregistrement $f3.wav$ est présenté sur la figure 3.15.

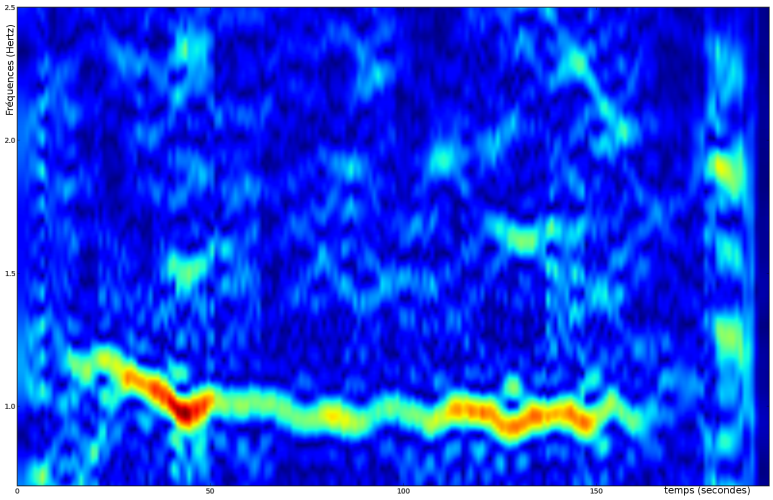


FIGURE 3.15 – Tempogramme du fichier $f3.wav$ (3 minutes). L'image extrêmement structurée temporellement révèle clairement la présence d'une fréquence dominante durant la majorité du fichier.

Sur cet enregistrement, nous distinguons clairement une fréquence prédominante sur la majorité du morceau (autour de 1 Hz). Cette zone correspond clairement à la portion d'enregistrement contenant le conte. Des variations de rythme effectuées au long du récit sont également visibles. Cette analyse montre une claire différence de comportement vis-à-vis du premier exemple de parole $f1.wav$.

Discussion

Cette tâche de caractérisation de catégories intermédiaires entre voix parlée et voix chantée est une tâche aussi originale que difficile. En effet, l'existence même de ces catégories, si elle n'est pas remise en cause est encore très mal définie, même au sein de la communauté des linguistes.

Au sein du consortium du projet DIADEMS, différentes réunions de travail ont d'ailleurs lieu parmi les ethnolinguistes et ethnomusicologistes afin de les définir précisément. Nous pouvons donc difficilement affecter des labels sur ces phénomènes ou en créer des prototypes ou des modèles.

Néanmoins, notre analyse sur différents types de paroles retournent bien des caractéristiques différentes, gradées en fonction de leur structure temporelle plus ou moins marquée. Si nous ne pouvons pour l'instant espérer effectuer une application directe de classification dans différentes catégories, nous espérons que nos premiers résultats pourront être utilisés afin de mieux définir ces catégories.

3.6 Conclusion

Ce chapitre présente notre approche sur l'analyse rythmique de contenus sonores. Elle est basée sur une analyse fréquentielle d'instants de ruptures dans la modélisation du signal. En utilisant l'hypothèse que les ruptures les plus porteuses de sens en terme de structure temporelles sont celles de plus forte croissance d'énergie du signal, nous proposons une pondération de ces frontières et leur analyse fréquentielle au travers du Spectre de Rythme, B_p. Un suivi est obtenu au travers du tempogramme.

Nous illustrons notre travail à travers deux exemples applications : l'une sur la musique et l'autre sur la parole.

La première se concentre sur les contenus musicaux afin d'estimer la valeur du tempo d'un morceau à tempo fixe. Une analyse du Spectre de Rythmes nous permet d'obtenir pour 78% des cas la valeur correcte du tempo, et 95% de valeurs directement liées à l'annotation, sur une base de référence. Ces performances sont légèrement supérieures à différentes méthodes existantes dans la littérature.

La seconde application se focalise sur la parole : le tempogramme permet une caractérisation de différents types de voix intermédiaires entre la voix parlée et la voix chantée. En analysant la présence ou non d'un « *tempo* » de parole marqué, nous pouvons retrouver ces différentes catégories en fonction de leur aspect plus ou moins structuré sur le plan temporel. L'évaluation qualitative sur trois extraits permet de conforter nos espoirs.

Ces exemples ne sont dont pas restrictifs du cadre d'application de ces outils et il est certain que l'utilisation d'un tempogramme peut s'utiliser pour l'analyse de morceaux à tempo variables, voire à la localisation de bruits récurrents comme par exemple des pas.

Chapitre 4

Suivi de fréquences et nombre de sources

4.1 Introduction

L'analyse des zones polyphoniques présentent de nombreux défis scientifiques : l'un d'eux peut être dans certains cas leur simple détection. Nous nous sommes intéressés à cette problématique et plus précisément à la détection de phénomènes mettant en jeu plusieurs sources harmoniques. Nous nous sommes concentrés sur deux types de phénomènes différents : les chœurs à l'unisson et les contenus polyphoniques plus traditionnels. Pour ces deux types de détection, nous nous basons sur une détection et un suivi des fréquences prédominantes du signal. Leur exploitation est néanmoins différente, du point de vue de la stratégie comme de la mise en œuvre : les paramètres dépendent des hypothèses propres à chacune des tâches.

Après une revue de l'état de l'art des techniques traitant des sources harmoniques multiples, nous présenterons l'approche de discrimination entre chœur et solo, avant de traiter la détection de sources multiples.

4.2 État de l'art

La superposition de plusieurs zones harmoniques est abordée de manière très différente selon le contexte « Parole » ou « Musique ». Le survol de l'état de l'art dans ces deux contextes a pour but de donner les tendances et pointer les différences.

4.2.1 Parole superposée

Quels enjeux ?

Dans le contexte de parole, la détection de superpositions entre plusieurs locuteurs a longtemps été un problème ignoré par les méthodes d'indexation automatique. Ceci est principalement dû à l'extrême rareté de ce phénomène dans les contenus analysés jusqu'à récemment : la plupart des méthodes d'analyse de la parole, que ce soit en transcription ou en segmentation-regroupement en locuteurs ne s'intéressaient qu'à des contenus de parole préparée, voire lue, comme des journaux télévisés. Dans de telles situations, souvent mono-locuteur, la présence de locuteurs multiples, est extrêmement contrôlée à des fins d'intelligibilité maximale du message et la présence de contenus de parole superposée est peu ou pas permise.

L'augmentation des performances des systèmes de transcription automatiques sur de tels contenus a poussé les équipes de recherches à s'aventurer sur le terrain de la parole non préparée, spontanée. La présence de zones de superposition entre les différents locuteurs devient beaucoup moins anecdotique, jusqu'à devenir très présente dans des contenus de type débats. Par exemple lors de débats politiques, interrompre d'autres interlocuteurs est une technique à part entière !

Dans ces contextes de superposition, les approches classiques de transcription de la parole, basées sur des modèles statistiques phonétiques appris sur des heures de corpus radiophoniques ou télévisuels propres se trouvent en échec. Sur des enregistrements disposant de nombreuses zones de superposition les performances du système se trouvent sérieusement affectées. Une solution qui consiste à multiplier les apprentissages sur les différents types d'intervention co-occurrente de plusieurs locuteurs, pose un problème d'explosion combinatoire des cas.

Une approche pour améliorer les performances des systèmes de transcription consiste à localiser les zones présentant une co-occurrence de locuteurs afin de les traiter par une méthode spécifique. Cette problématique a donc récemment émergé parmi les problématiques clefs de l'augmentation des performances des systèmes de transcription sur des contenus de parole moins contrôlés. Conséquence de cet intérêt grandissant, une tâche de localisation de parole superposée a été proposée lors de la campagne d'évaluation 2012 du projet ANR *Etape* [1].

Cette tâche représente un grand défi scientifique car contrairement à la musique traditionnelle, la parole est beaucoup plus chaotique en ce sens où des zones voisées alternent très rapidement avec des zones non voisées, avec des énergies pouvant également varier très vite. Dans ce contexte, la contribution des différentes sources au mélange des voix est instable et l'une ou l'autre des sources peut tour à tour se retrouver très prédominante par rapport à l'autre, ce qui impacte le problème de

1. `http://www.afcp-parole.org/etape.html`

la localisation.

Différentes approches ont été proposées pour relever ce défi : nous en présenterons trois types différents. Les méthodes de détection de zones de parole superposée sont pour plusieurs d'entre elles héritières des méthodes classiques de traitement de la parole : certaines, de type traitement de signal, sont basées sur l'extraction d'informations directes sur le signal ; les approches segmentation-regroupement en locuteurs proposent une analyse des segments et de leur répartition pour détecter les zones de superpositions ; d'autres approches impliquent l'apprentissage de paramètres issues de zones étiquetées comme étant des superpositions.

Approche traitement de signal

L'un des problèmes principaux de la détection de sources multiples est le mélange des différentes harmoniques générées par les sources qui peuvent interagir entre elles, conduisant à une estimation erronée du nombre de fréquences fondamentales, et donc de sources. Afin de répondre à ce besoin d'une estimation précise du nombre de sources, Liénard [30][31] propose une méthode d'analyse fondée sur le filtrage du spectre du signal par un banc de peignes conçus spécifiquement afin d'éliminer les problèmes de mauvaise estimation de pitch aussi bien en contenu monopitch que multipitch.

Une fonction de pitch, notée $\Phi(f)$ est définie par le produit scalaire entre le spectre d'amplitude du son analysé $|S(f)|$, centré et un peigne paramétré par sa fréquence caractéristique F_c, noté $P(F_c, f)$. Les différentes valeurs de pitch sont estimées à partir des fortes valeurs dans la fonction Φ. La recherche de fréquence fondamentale se fait dans un intervalle $[F_{0min}, F_{0max}]$.

Les différents peignes proposés sont les suivants :

- Les **P**eignes **U**niformes **I**nfinis (**PUI**). Ces peignes consistent simplement en une série de Dirac situés à des fréquences séparées par des intervalles constants. Ce peigne, le plus simple proposé, fixe l'amplitude de toutes les dents du peigne à 1 et les dents sont définies jusqu'à la fréquence maximale du spectre. Un exemple de ce type de peigne est présenté sur la figure 4.1.
- Les **P**eignes à **D**ents **N**égatives (**PDN**). Ces peignes sont définis grâce à un paramètre supplémentaire correspondant à un ordre. Cet ordre Δ est fixé à un entier naturel supérieur ou égal à 2. Dans ce type de peigne, seule une dent sur *Delta* est fixée à une amplitude de 1 alors que les autres sont fixées à une valeur négative a. Cette valeur de a dépend de l'ordre selon la formule suivante : $a = \frac{1}{1-\Delta}$. Ce type de peigne est conçu afin de limiter la valeur de Φ pour des valeurs de fréquences correspondant à une sur-harmonique, alors qu'elles ont tendance à être relativement élevées avec d'autres types de peignes. Un exemple de ce type de peigne est présenté sur la figure 4.2.

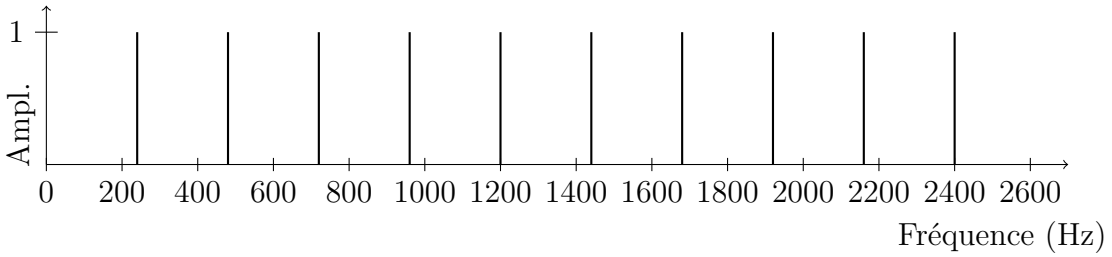


FIGURE 4.1 – Représentation d'un PUI de fréquence caractéristique $F_c = 240$ Hz. Les dents du peigne sont uniformément réparties et leur amplitude est fixée à 1.

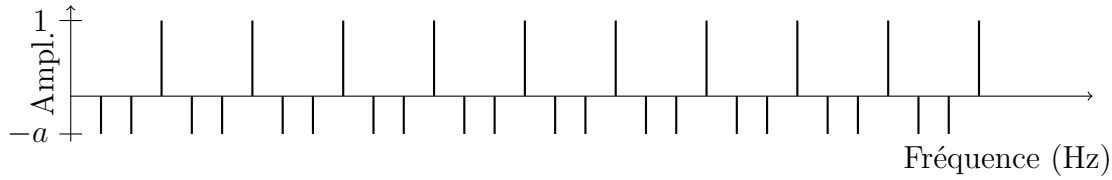


FIGURE 4.2 – Peigne à dents négatives de fréquence caractéristique $F_c = 240$ Hertz et d'ordre $\Delta = 3$. L'amplitude des dents négatives est définie en fonction de l'ordre Δ et permet de limiter l'impact des pics sur-harmoniques en soustrayant l'amplitude des sous-harmoniques de F_c.

– Les **P**eignes à **D**ents **M**anquantes (**PDM**). Ces peignes sont également paramétrés par un ordre Δ. Cet ordre est un nombre premier supérieur à 2. Une dent sur Δ voit son amplitude annulée alors que les amplitudes des autres dents sont fixées à la valeur a. Ce peigne vise à défavoriser les sous-harmoniques de la fréquence fondamentale et vise à être utilisé en conjonction avec le **PDN**. Un exemple de ce type de peigne est présenté sur la figure 4.3.

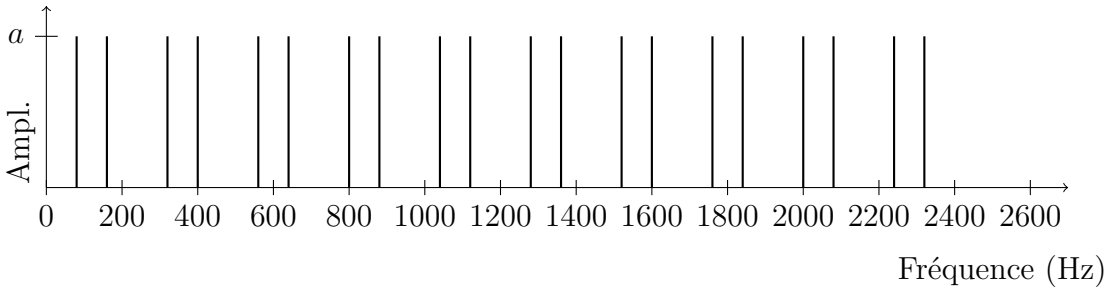


FIGURE 4.3 – Peigne à dents manquantes de fréquence caractéristique $F_c = 80$ Hz et d'ordre $\Delta = 3$. Ce peigne vise à diminuer les erreurs de sous-estimation de la fréquence fondamentale.

Liénard, propose un système composé de l'enchaînement de différents peignes uniformes infinis, à dents manquantes de différents ordres et à dents négatives également de différents ordres. La combinaison de ces peignes permet de faire clairement ressortir les pics correspondant aux fréquences fondamentales, même en contexte multipitch. La figure 4.4 illustre le gain de l'utilisation de cette succession d'analyse (partie *b*) par rapport à l'utilisation d'un simple **PUI** (partie *a*) sur l'analyse d'un spectre harmonique de synthèse comportant deux pics. Le spectre analysé est celui présenté sur la figure 4.5.

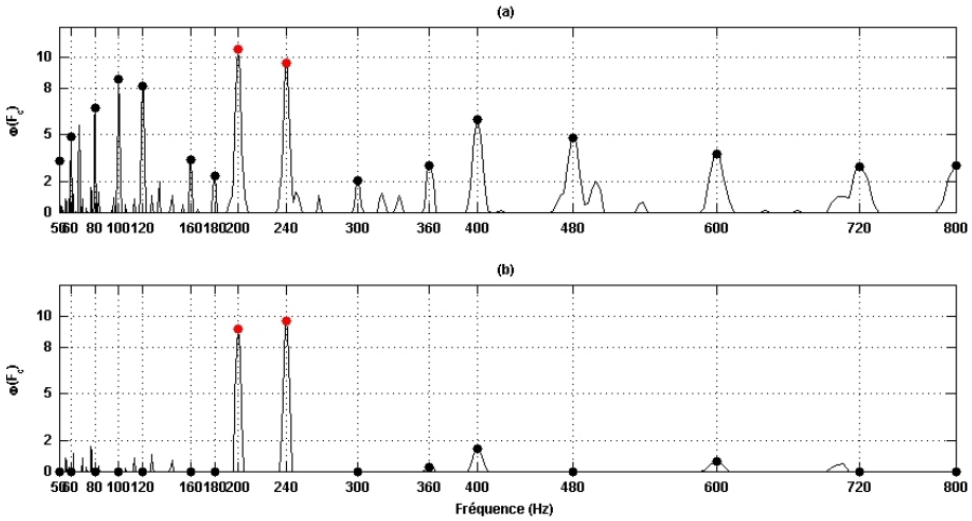


FIGURE 4.4 – Représentation de la fonction de pitch $\Phi(f)$, pour une analyse par un simple **PUI (a)** ; pour une analyse par application successive de trois peignes **PUI**, **PDM**s et **PDN**s **(b)**.

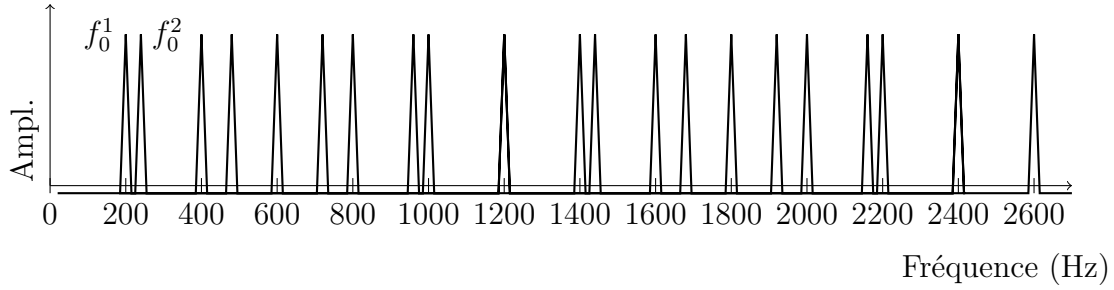


FIGURE 4.5 – Exemple de spectre harmonique de synthèse comportant deux fréquences fondamentales $f_0^1 = 200$ Hz, et $f_0^2 = 240$ Hz.

L'analyse par **PUI** produit de nombreux pics secondaires (en noir). Leur amplitude est certes moindre que celle des deux pics correspondant aux fréquences fondamentale (en rouge), mais elle reste néanmoins comparable, et en l'absence d'information *a priori* sur le nombre de sources, les risques d'erreurs d'estimation restent importants. En revanche, la fonction Φ de l'analyse complète est sans ambiguïté quant à la présence de deux pitchs, qui ressortent clairement.

D'autres approches utilisent une analyse statistique du signal pour la détection de zones de superpositions. On peut citer les travaux de Ben-Harush [32] qui se base sur l'hypothèse que les trames contenant plusieurs locuteurs ont une entropie plus élevée, au sens de l'entropie de Shannon [33]. Partant de ce constat, et sous l'hypothèse de la présence de deux uniques locuteurs, un histogramme des entropies calculées par trames sur l'ensemble de la conversation est réalisé. Par différentes techniques d'approximation, cette distribution est modélisée par quatre lois gaussiennes. Selon sa moyenne, chacune correspond alors par ordre croissant aux classes : *non parole*, *locuteur A*, *locuteur B* et enfin *zones de parole superposée*, d'où la détermination *a posteriori*.

Approches par apprentissage

Une autre stratégie, développée par le LIMSI et Orange Labs [34], est une approche basée sur l'apprentissage d'une représentation des zones de parole superposée. Elle nécessite de disposer d'une base de phrases étiquetées « non superposées » et « superposées ».

Les deux systèmes diffèrent de par la paramétrisation : la méthode LIMSI utilise 12 PLP (**P**erceptual **L**inear **P**redictive, [35]) et la log énergie, calculés par trame ainsi que leurs dérivées premières et secondes. Celle de Orange utilise les MFCC (**M**el **F**requency **C**epstrum **C**oefficients, [36]), en lieu et place des PLP. Les deux paramétrisations permettent de rendre compte de la perception humaine.

L'utilisation de paramètres cepstraux comme les MFCC contient une information timbrale liée aux phonèmes et, à plus grande échelle, à la source ; l'information liée à la fréquence fondamentale est fortement atténuée.

L'analyse PLP est une analyse classiquement utilisée pour produire une représentation spectrale indépendante du locuteur.

Si les paramètres de ces deux méthodes sont proches, leur approche de détection est divergente. Le système du LIMSI utilise une décision trame à trame basée sur le maximum de vraisemblance entre trois modèles correspondant respectivement à de la non-parole, de la parole non superposée et de la parole superposée. À cette vraisemblance s'ajoute une analyse selon l'approche de Liénard exposée plus haut. En effectuant une combinaison linéaire entre la vraisemblance et le nombre de fréquences fondamentales extraites, la décision de parole superposée est ensuite prise.

Le système proposé par Orange utilise lui une approche plus classique du domaine du traitement de la parole. Une décision se fait selon une détection par trois modèles GMM à 256 gaussiennes. Les deux premiers modèles correspondent respectivement à de la parole non superposée prononcée par un homme et par une femme, le troisième modèle correspondant à de la parole superposée. Un HMM à deux classes regroupant les deux modèles HMM de parole non superposée est ensuite construit afin de détecter les zones de parole superposée. Le décodage par l'algorithme de Viterbi est ensuite utilisé sur tous les segments identifiés comme de la parole. Enfin, un lissage est effectué pour éviter la détection de zones trop courtes.

Autre méthode plus originale, celle de Vipperla [37] ne se base pas à proprement parler sur une segmentation en locuteurs mais sur une technique proche de la factorisation en matrice non négative ou *NMF* (**N**on-negative **M**atrix **F**actorization, [38]).

Cette technique consiste à décomposer une représentation matricielle du signal, telle qu'un spectrogramme, comme le produit de deux matrices représentant respectivement les éléments de base de la construction du signal (appelées atomes) et leurs activations temporelles.

Les bases correspondant aux différents locuteurs sont apprises sur des segments de parole non superposée, ces segments étant issus soit d'une annotation soit d'une segmentation-regroupement en locuteur automatique. L'activation des atomes correspondant aux motifs spectraux des différents locuteurs est ensuite calculée selon la méthode NMF. Les activations suffisamment fortes conjointes d'atomes correspondant à différents locuteurs sont finalement utilisées pour détecter les segments de parole superposée.

De récentes approches tentent d'ajouter de nouveaux paramètres afin d'améliorer les performances d'un système simplement basé sur une analyse des MFCC ou des PLP. Citons par exemple la méthode proposée par Yella [39] qui utilise comme paramètre la répartition des zones de silences, pour la détection de ces zones dans des enregistrements de réunions. Cette approche part du principe que la disposition des silences est suffisamment différente entre les zones mono-locuteurs et les zones de prise de parole disputées pour les rendre plus facilement détectables.

Cette problématique de détection de sources harmoniques simultanées se pose aussi depuis longtemps sur des contenus de musique. Le contexte et les propriétés étant différentes, des techniques différentes ont été proposées.

4.2.2 Multipitch en Musique

Quels enjeux ?

L'étude des zones comportant de multiples sources dans les contenus de musique a commencé avec les débuts de l'analyse automatique des contenus musicaux. Les premiers types de musique analysés ont été issus de la musique occidentale et il s'agissait d'enregistrement de bonne qualité et facile d'accès. De nombreuses méthodes ont vu le jour avec des objectifs variables et elles sont utilisées soit en tant que telles, soit en tant que prétraitement à d'autres méthodes. Une revue de différentes approches utilisées pour l'étude de la musique dite « multipitch » est présentée par Klapuri et Davy dans [40].

La majorité des approches vise à une **estimation des différentes fréquences fondamentales présentes** (c'est-à-dire des notes jouées) en fonction du temps. Ces méthodes sont développées dans un but de *transcription* de la musique sans identification des instruments. D'autres techniques au contraire essaient de **localiser les voies des différents instruments** à des fins de *séparation de sources*. Enfin, l'objectif peut consister à la **localisation de zones où plusieurs instruments interviennent à la fois**, des zones de *polyphonie*. Ce sera l'objectif principal de notre propre étude.

Bien sûr, chaque méthode ne poursuit pas un objectif unique, et différentes informations peuvent être estimées conjointement. Néanmoins, les points forts et faibles des différentes méthodes sont choisis en fonction de leur objectif principal.

Quels problèmes ?

Dans le cadre de la transcription musicale, si nous pouvons considérer le problème de l'analyse monopitch comme un problème clos grâce aux performances d'algorithmes basés sur l'auto-corrélation tel le YIN [4], l'analyse des zones « multipitch » propose un défi plus relevé.

De manière générale, la caractérisation de différentes sources se heurte à deux difficultés principales :

- la **corrélation des sources**. Les sources harmoniques interagissant de manière très complexes, il peut parfois se révéler très difficile d'identifier certaines sources. Des effets de masquage entre harmoniques multiples peuvent intervenir [41].
- le **nombre de sources**. Beaucoup de méthodes de transcription multipitch se basent sur un schéma « analyse-soustraction » où chaque source identifiée voit sa signature soustraite au signal (ou à sa représentation fréquentielle) jusqu'à ce que le résidu soit jugé ne plus contenir de source. Une mauvaise quantification du résidu peut facilement conduire à une omission ou un ajout erroné d'une source.

En fonction des objectifs et des techniques employées, nous présentons les différentes approches qui ont été envisagées.

Quelles réponses ?

Dans le domaine de la transcription *multipitch*, un gros effort a été fourni pour le développement des méthodes avec notamment une tâche dédiée dans la campagne d'évaluation MIRex [2] et dans le projet QUAERO [3].

Outre leurs objectifs, différents types d'approches ont été envisagées pour le développement de systèmes d'analyse de signaux multi-source.

Ces différentes approches ont, comme en parole, différentes philosophies qui permettent de traiter des zones contenant plusieurs sources harmoniques. Dans cette section nous proposons une courte revue de différentes approches utilisées en musique.

Approches par apprentissage

L'une des premières approches a consisté à transcrire les parties « mélodie » et « basse » d'un morceau. Il s'agit en quelque sorte d'une simplification du problème en fixant artificiellement le nombre de sources à 2. L'algorithme de Goto [26] s'attaque à cette problématique. L'analyse s'effectue sur deux bandes de fréquences disjointes correspondant à la basse et à la mélodie principale. Le découpage permet d'obtenir dans chaque bande une source clairement prédominante sur l'autre. Un modèle tonal des positions des pics d'amplitude de chaque bande est réalisé et sur chaque trame analysée, la note jouée est détectée par vraisemblance avec les modèles appris. Une contrainte sur la trajectoire des fréquences extraites est ajoutée *a posteriori* pour éviter les trajectoires trop chaotiques et permettre des suivis plus réalistes.

Raczynski [42] utilise aussi un apprentissage préalable. Son approche est basée sur une factorisation en matrice non-négative de la transformation à Q-constant. Cette représentation spectrale est mieux adaptée à l'analyse de la musique en assurant une résolution constante en fonction des octaves. Cette représentation est expliquée plus en détail dans la prochaine partie. Les bases utilisées peuvent être apprises au préalable. L'auteur justifie ce choix par le fait que les bases automatiquement identifiées peuvent varier d'une analyse à l'autre et que rien ne force les bases ainsi estimées à correspondre à des notes. Le fait d'utiliser des motifs harmoniques correspondant à des notes augmente également la sensibilité au bruit et permet une interprétation directe de la matrice d'activation comme transcription.

2. `http://www.music-ir.org/mirex/`
3. `http://www.quaero.org/`

L'approche par apprentissage est l'une des approches les plus anciennes, visant une analyse de corpora spécifiques car, si la technique fonctionne très bien sur des données proches de leur apprentissage, elle ne peut fonctionner correctement sur des types de contenus ou sous des conditions qui lui sont inconnues. Assurer une grande flexibilité à ces méthodes demande une augmentation importante du contenu utilisé en apprentissage, ce qui n'est pas toujours simple, voire possible. L'inconvénient principal de cette méthode est que malgré un très bon taux d'estimation sur du contenu proche de l'apprentissage, l'hétérogénéité des contenus augmente très fortement le volume de donnés nécessaire à l'apprentissage. Ce type d'approche n'est pas adapté à l'étude de corpora caractérisés par leur hétérogénéité, il impliquerait un apprentissage de toutes les familles d'instruments potentiellement présents voire de qualités d'enregistrement différentes.

Afin de lutter contre cet inconvénient, d'autres études ont proposé des méthodes de modélisation avec découverte des modèles. Sakaue [43] propose une approche dérivée de la NMF pour laquelle les bases et la matrice d'activation correspondent à des distributions gaussiennes. Un nombre K d'atomes représentatifs sur le signal est fixé et les K atomes les plus différents sont extraits de la représentation spectrale afin de modéliser les différentes configurations spectrales. Chaque base est affectée *a posteriori* à une note (ou à son absence) pour ensuite obtenir la transcription du morceau analysé. L'utilisation d'une représentation par gaussienne semble réduire la sensibilité aux bruits et autres variations.

Enfin, celle de Yoshii [44] utilise une modélisation plus complexe des sources et une approche bayésienne non paramétrique. Ni le nombre de sources présentes n'est fixé ni le nombre de leurs harmoniques. Cette analyse part de l'hypothèse que le nombre de sources ou d'harmoniques est infini, mais que la plupart sont négligeables et éliminées par l'analyse statistique. L'avantage principal de ces approches en comparaisons avec celles utilisant un apprentissage *a priori* réside dans le fait qu'elles s'adaptent au contenu analysé et ainsi ne nécessitent pas de larges ensembles d'apprentissages pour assurer une bonne robustesse.

Approches de type caractérisation du signal

La majorité des méthodes proposées pour la transcription polyphonique d'enregistrements musicaux se basent sur une analyse statistique du signal ou des paramètres qui en sont extraits. Ces approches, plus proches des techniques issues du traitement du signal sont de loin les plus représentées dans les campagnes d'évaluation récentes comme QUAERO ou MIRex. Nombre d'entre elles intègrent le fait que l'oreille humaine possède une très bonne capacité de discrimination des notes, mêmes jouées simultanément. Nous pouvons discerner deux groupes selon qu'il est fait appel à des heuristiques propres au contexte musical ou à la perception humaine. Ces deux types d'approches sont illustrés ci-dessous.

Représentation acoustique adaptée à la note

La représentation acoustique est basée sur une analyse fréquentielle multi-résolution. Cette analyse est développée afin de contrer le problème de changement de dynamique entres notes de différentes hauteur.

En effet, si nous considérons notamment le cas de la musique occidentale, chaque fréquence des notes de la gamme est définie de la manière suivante. La gamme est composée d'octaves, le passage d'une note à la même note de l'octave supérieure se fait en multipliant la note d'origine par 2. Chaque octave est découpée en 12 notes équitablement réparties, ce qui conduit à la relation entre deux notes n_1 et n_2 consécutives suivante :

$$freq(n_2) = \sqrt[12]{2} * freq(n_1) \tag{4.1}$$

La définition de la gamme comme une suite géométrique conduit à ce que l'écart entre notes consécutives augmente de manière exponentielle avec le rang de la note (voir figure 4.6).

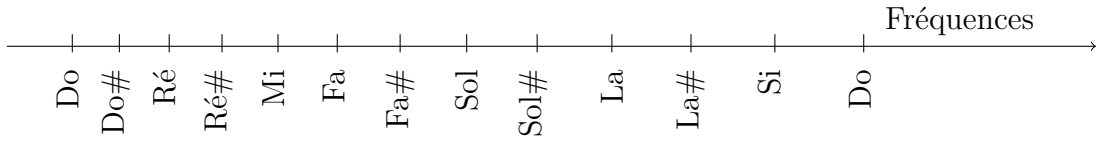


FIGURE 4.6 – Répartition exponentielle des 12 notes de la gamme occidentale sur l'échelle des fréquences.

Or, en utilisant une analyse par spectrogramme classique, chaque bin du spectre correspond à l'énergie d'une plage de fréquence fixe sur toute l'analyse et la dynamique dans les fréquences les plus élevées est sensiblement plus grande que pour les fréquences basses. La détection tend donc à être beaucoup plus précise et fiable sur les notes élevées. Afin de prendre en compte ce problème, plusieurs approches visent à utiliser une analyse à différentes échelles guidées par les connaissances *a priori* des fréquences propres aux notes.

L'une des premières méthodes présentée pour une représentation spectrale guidée par les fréquences des notes est la **transformée en Q-constant** proposée par Brown en 1992 [45].

Dans cette méthode d'analyse fréquentielle, chaque bin est calculé sur une plage de fréquence logarithmique avec une largeur de fenêtre d'analyse fixée pour une même précision autour de chaque note. Les figures 4.7 et 4.8 présentent respectivement le principe d'une analyse spectrale classique avec des fréquences espacées d'un pas fixe, et le principe de la transformée à Q-constant avec des bins placés sur les positions fréquentielles des notes.

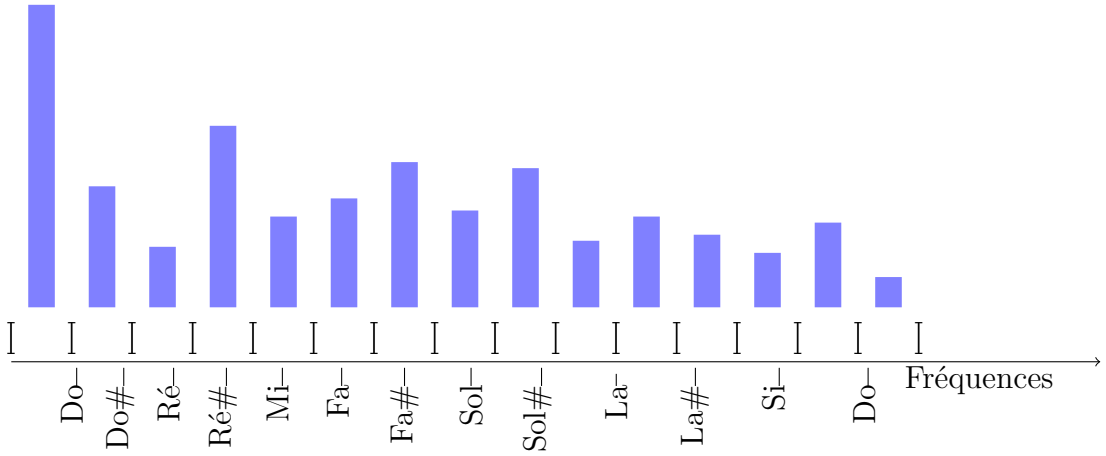


FIGURE 4.7 – Analyse fréquentielle classique par transformée de Fourier discrète. Les fréquences des bins sont définies par un découpage uniforme de la plage $0 - f_e$ et ne correspondent pas au positionnement des notes. Certains bins peuvent alors englober plusieurs notes, ou n'en contenir aucune.

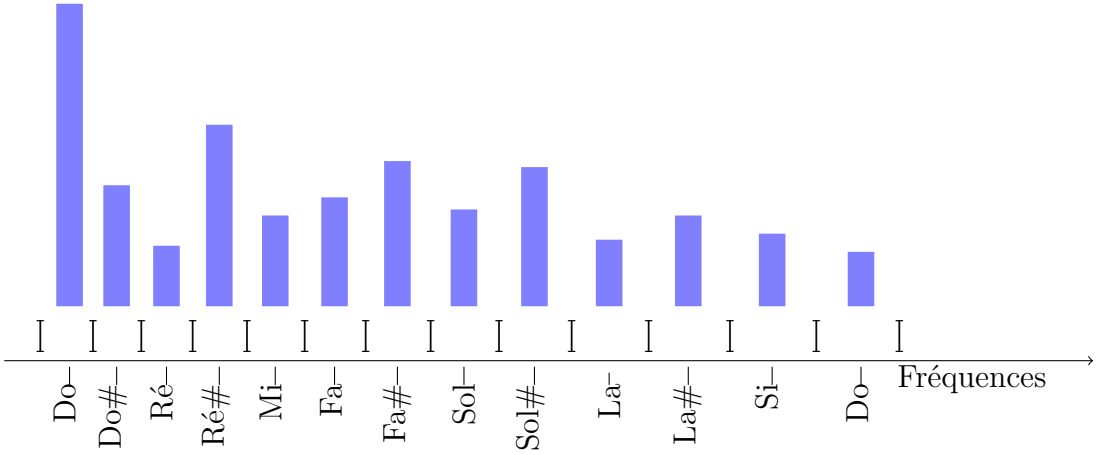


FIGURE 4.8 – L'espacement entre les différentes fréquences centrales des bins est fixé de manière logarithmique et chaque bin est centré sur les fréquences des notes. La définition fréquentielle de chaque note dans le spectre obtenu est rendu linéaire.

D'autres méthodes ont exploité cette idée de non linéarité de l'analyse fréquentielle. La transformée à Q-constant se révélant coûteuse, Dressler [46] propose une adaptation multi-résolution se basant sur une adaptation de l'algorithme de la transformée de Fourier rapide. Cette adaptation offre un bon compromis afin de pouvoir conserver une définition fréquentielle comparable entre les différentes octaves tout en permettant une analyse plus rapide.

Analyse du signal par annulation

Une méthode classique de détection multipitch basée sur une analyse du signal par spectre traditionnel est celle proposée par Klapuri [9] en 2006. Cette méthode, pose les bases d'une approche très utilisée parmi les méthodes de transcription, l'approche estimation-annulation.

Klapuri n'utilise pas de représentation fréquentielle fondée sur la construction musicale comme les travaux présentés dans la partie précédente, mais il utilise un blanchissement spectral (*Spectral whitening*). Le blanchissement spectral est une méthode de filtrage visant à rééquilibrer les amplitudes des différentes fréquences afin de supprimer au maximum l'information de timbre et ainsi s'abstraire des types de sources utilisées. Ce blanchissement est une technique largement utilisée dans les méthodes de transcription musicale et peut être effectuée de différentes manières.

La méthode utilisée par Klapuri est la suivante :

– Un banc de filtre, dont chaque bande a une réponse triangulaire H_b, est généré. Chaque centre de bande c_b est défini par la fonction suivante :

$$centreFreq(b) = 229 \times 10^{(b-1)/21.4} - 1 \tag{4.2}$$

avec $b \in [1, ..., 30]$ le rang du filtre. La réponse triangulaire du filtre b est défini sur $[centreFreq(b-1), ..., centreFreq(b+1)]$. La figure 4.9 illustre les réponses H_b.

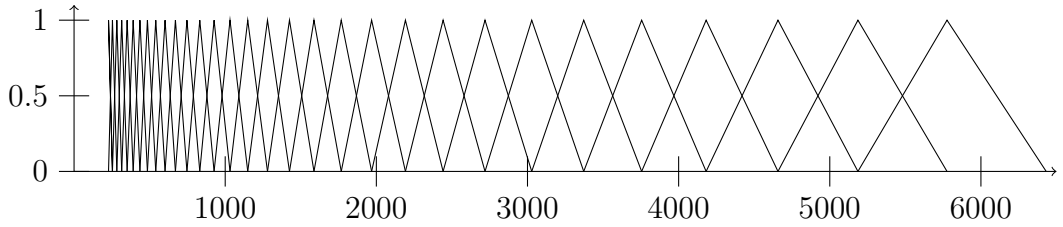


FIGURE 4.9 – Réponses H_b utilisées pour le blanchissement spectral. Celles-ci sont paramétrées par la valeur de $centreFreq$, définie en 4.2.

– L'écart type de l'application de chaque bande σ_b au spectre X est ensuite calculé :

$$\sigma_b = \sqrt{\frac{1}{K} \sum_k H_b(k)|X(k)|^2} \tag{4.3}$$

avec K le nombre de points du spectre X. Cet écart-type permet ensuite la définition d'un coefficient de compression par bande :

$$\gamma_b = \sigma_b^{v-1} \tag{4.4}$$

où v est un coefficient de compression fixé à $0,33$ dans l'article de Klapuri.

– Finalement, une fonction de poids γ est définie par interpolation linéaire entre les points (c_b, σ_b). Cette fonction de poids est appliquée au spectre afin d'obtenir le spectre blanchi :

$$Y(k) = \gamma(k)X(k) \tag{4.5}$$

Sur le spectre ainsi blanchi, la technique de Klapuri consiste à calculer une fonction de saillance $s(\tau)$ mettant en évidence les fréquences appartenant aux sources. Cette fonction de saillance est calculée à partir de $Y(k)$ par l'application d'une autre fonction de poids $g(\tau, m)$:

$$s(\tau) = \sum_{m=1}^{M} g(\tau, m)|Y(f_{\tau,m})| \tag{4.6}$$

avec $f_{\tau,m}$ la fréquence correspondant à la $m^{ième}$ harmonique de la fréquence τ et M le nombre maximum d'harmoniques prises en compte.

L'enjeu principal de la détection est ensuite d'optimiser la fonction de poids en vue d'un taux d'erreur minimal. La fonction retenue après expérience est de la forme suivante :

$$g(\tau, m) = \frac{f_s/\tau + \alpha}{mf_s/\tau + \beta} \tag{4.7}$$

avec f_s la fréquence d'échantillonnage du signal et α et β des paramètres de modération, fixés empiriquement en fonction de la taille de la fenêtre d'analyse, pour modéliser au mieux la fonction de poids observée dans ses expériences. Le nombre de sources présentes étant inconnu, le système utilise une approche classique estimation-annulation.

La figure 4.10 en présente le principe.

L'idée est d'éliminer toutes les composantes significatives présentes dans le spectre. Sur le spectre de saillance une fréquence fondamentale f_0 est détectée en prenant la position de la valeur maximale du spectre de saillance. Ensuite, un spectre des amplitudes correspondant à f_0 et ses harmoniques est défini grâce à la fonction de poids $g(\tau, m)$ présentée précédemment. Ce spectre des saillances liés à f_0 est ensuite soustrait du spectre de saillance globale. Ce processus est réitéré jusqu'à ce que le résidu soit considéré comme ne contenant plus de phénomène harmonique.

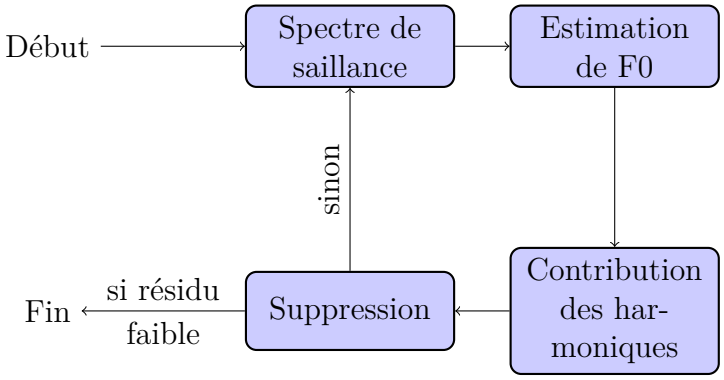


FIGURE 4.10 – Diagramme de flux des approches estimation-annulation.

4.2.3 Conclusion

L'analyse multipitch de contenus en musique est un domaine de recherche actif depuis plusieurs années. C'est pourquoi de nombreuses approches ont été proposées en utilisant des techniques diverses issues des domaines du traitement de signal comme de la reconnaissance de formes. Ces méthodes peuvent s'appuyer sur des heuristiques issues de la musicologie afin de contraindre les systèmes et d'éviter à moindre coût de nombreuses erreurs. L'approche estimation-annulation développée par Klapuri semble la plus proche de notre philosophie. C'est pourquoi nous l'avons implémentée afin d'en étudier les performances et les comparer à nos propres développements.

Dans le domaine du traitement de la parole, l'analyse de plusieurs locuteurs en est à ses débuts. L'objectif de ces méthodes est pour l'instant l'amélioration des techniques de transcription automatique, très mal à l'aise dans les cas de recouvrement de la parole. Les approches sont principalement issues de techniques utilisées traditionnellement dans le traitement de la parole : modélisation GMM ou segmentation-regroupement en locuteur. La technique des peignes développée par Liénard présente une technique purement issue du traitement de signal que nous implémenterons par la suite, également à titre de comparaison avec nos études.

4.3 Distinction chœur-solo

Nous parlons de **chœur à l'unisson** lorsqu'un groupe de chanteurs chante une même mélodie de manière synchronisée. Ce type de production diffère du canon pendant lequel différents sous-groupes de chanteurs créent un décalage temporel sur la mélodie, et de chants en chœurs utilisant des intervalles mélodiques entre différents sous-groupes (tierce, quinte...). Une zone de chœur à l'unisson est de

fait détectée par notre système monophonie/polyphonie comme une zone monophonique.

Le problème de la localisation de chœurs à l'unisson s'est posé clairement dans le cadre de l'analyse des documents ethnomusicologiques proposés par le projet *DIADEMS*. En effet, si ce type de chœur est rare dans la musique occidentale contemporaine, son existence est plus fréquente dans les contenus du projet. Cette information sur la nature du chœur peut être une information structurante déterminante dans certains enregistrements de rituels ou de contes dans lesquels l'assemblée reprend une mélodie initiée par un soliste dirigeant la cérémonie ou le conte. La recherche d'une telle structuration a motivé le développement de notre système de détection de zones de chœurs à l'unisson.

Notre recherche se focalise sur les segments monophoniques et elle est basée sur la détection de divergence dans les harmoniques. En effet, malgré la volonté des différents chanteurs d'entonner la même mélodie, de petits décalages temporels ou fréquentiels apparaissent. Ces décalages très fins sont amplifiés sur les harmoniques comme nous pouvons le voir sur la figure 4.11.

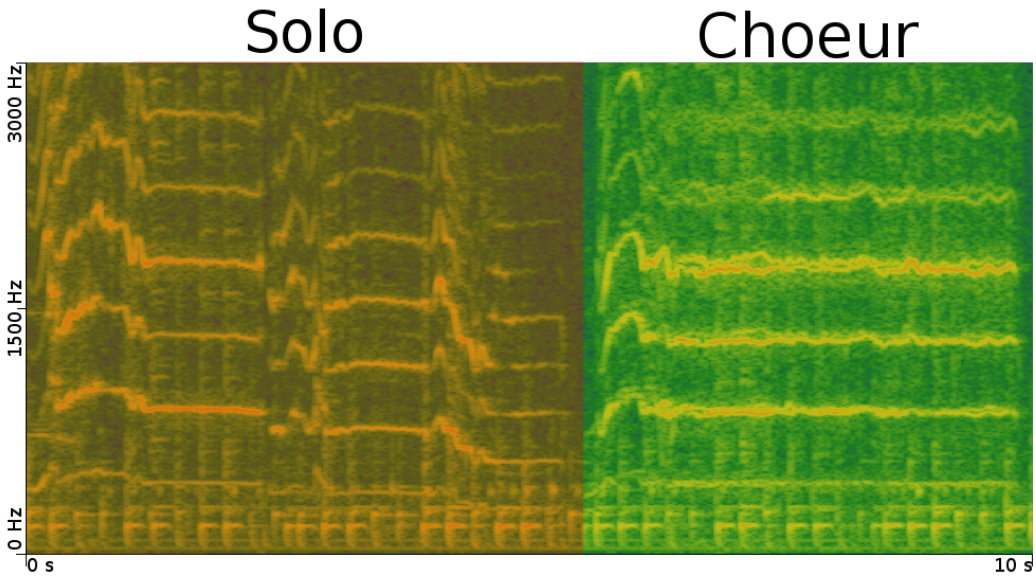


FIGURE 4.11 – Phénomène de coexistence de deux sources harmoniques en musique. L'extrait présenté est une transition entre chant solo et chœur.

Sur cette figure nous pouvons clairement distinguer la différence entre une zone solo composée d'un seul phénomène harmonique et une zone de chœur aux contours moins nets et comportant des divergences.

La stratégie que nous proposons est définie en trois étapes :

- la recherche de zones temps-fréquence pouvant contenir des zones correspondant à du chœur,
- le suivi des fréquences,
- la classification en Chœur/Solo.

Cet enchaînement de différentes étapes est détaillé dans le diagramme de flux présenté dans la figure 4.12.

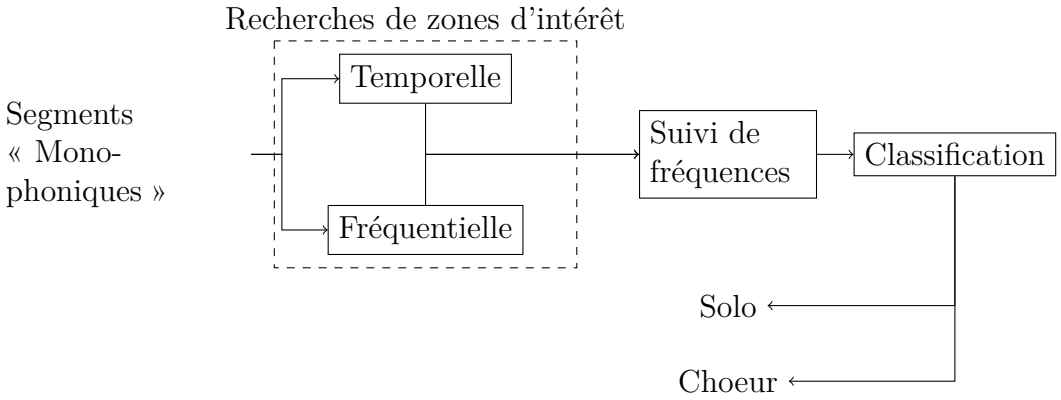


FIGURE 4.12 – Diagramme de flux du système de détection de chœur. Une sélection des zones de recherche est effectuée avant le calcul de suivi afin d'éviter au maximum les fausses alarmes.

4.3.1 Extraction des zones d'intérêt

Cette étape consiste à isoler des zones dans lesquelles le phénomène de divergence entre harmoniques peut se produire. Cette étape de sélection est très importante pour éviter un calcul du suivi de fréquence dans des zones bruitées ou de silence qui serait très coûteux et générateur de nombreuses erreurs de détection. La première restriction consiste à n'effectuer notre analyse que sur les segments retournés comme monophonique par le programme de classification monophonie-polyphonie déjà présenté dans le chapitre 2. Dans ces zones, nous cherchons à confirmer le diagnostic de la monophonie, ou détecter un chœur.

La sélection des zones d'analyse au sein des segments de monophonie s'effectue par une restriction dans les espaces temporel et fréquentiel des zones d'analyse à quelques bandes temps-fréquences qui se révèlent pertinentes.

Localisation temporelle

La localisation temporelle est effectuée en utilisant la segmentation Forward-Backward développée par André-Obrecht [2]. Pour plus de détails, voir la description donnée au chapitre 3. Comme nous l'avons déjà vu, son utilisation sur des contenus musicaux donne des résultats intéressants puisqu'elle tend à segmenter la musique dans les différentes phases de la note.

Sur chaque segment monophonique, nous nous concentrons sur les 2 segments issus de la segmentation **Forward-Backward** les plus longs. Ce segments correspondent généralement à des phases de *Sustain*, c'est-à-dire des phases de notes tenues, phases durant lesquelles la fréquence fondamentale reste stable mais dans lesquelles les divergences entre harmoniques, s'il existe plusieurs chanteurs apparaissent clairement.

Les segments sélectionnés doivent également être suffisamment harmoniques. Pour cela, nous utiliseront la moyenne du critère de confiance de l'algorithme du YIN [4], calculé sur le segment. Cette valeur doit être supérieure à un seuil th_{Harmo}.

Localisation fréquentielle

Une fois les segments isolés temporellement, l'intégralité de leur spectre ne sera pas analysée. Seules quelques bandes de fréquences seront extraites pour la suite de la méthode.

Pour cela, nous appliquons l'algorithme d'extraction de fréquence fondamentale YIN sur chaque trame t du segment considéré pour obtenir la fonction d'estimation de la fréquence fondamentale $f_0(t)$. Les segments étant homogènes et ce afin d'éviter des erreurs d'estimation, un lissage médian est effectué sur les estimations de fréquence fondamentale. De même, une correction des sauts d'octaves est effectuée en se basant sur l'octave majoritaire dans le segment.

La fonction d'estimation de fréquence fondamentale corrigée, que nous continuons d'appeler $f_0(t)$, sert de support pour définir une série de N_{bands} bandes de fréquence dans laquelle l'analyse sera effectuée. Au niveau de la trame t, chaque bande i est centrée autour de

$$i \times f_0(t)$$

, pour correspondre à une valeur approchée de la $i$ème harmonique.

La largeur de chaque bande est également définie en fonction du rang de l'harmonique selon la formule suivante :

$$bandwidth_i(t) = min(f_0, i \times bw \times f_0(t)), \tag{4.8}$$

avec bw un ratio correspondant au pourcentage de $f_0(t)$ à utiliser autour de la valeur de l'harmonique.

Cette augmentation de la largeur de la bande d'analyse en fonction des harmoniques permet de garder la même information spectrale dans chaque bande avec une définition de plus en plus élevée.

Cette restriction de la zone d'analyse à la fois sur les plans fréquentiel et temporel permet de concentrer la suite de l'analyse sur les seules zones pouvant contenir le phénomène d'intérêt. Cette double restriction est symbolisée par les bandes vertes sur la figure 4.13.

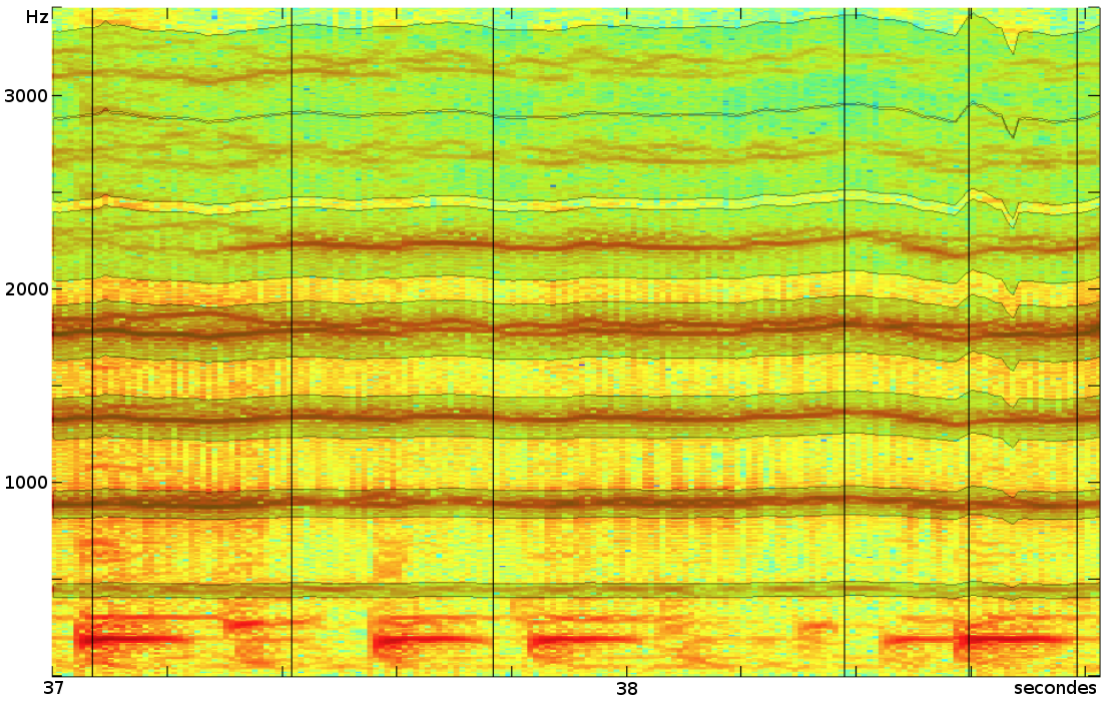


FIGURE 4.13 – Restriction temporelle et fréquentielle. Les bandes vertes symbolisent l'espace fréquentiel dans lequel le suivi va être calculé. Les traits verticaux noirs représentent la segmentation temporelle du signal. Seuls les deux plus grands segments par seconde sont conservés.

4.3.2 Sélection des pics

Malgré la restriction opérée à la fois sur le plan fréquentiel et temporel, nous ne sélectionnons que les principaux pics comme représentatifs.

Afin de ne garder que ces pics, nous utilisons un seuil basé sur l'énergie. Dans chaque trame d'analyse, sur chaque bande de fréquence seuls les pics supérieurs à ce seuil sont sélectionnés.

Ce seuil de sélection est fixé à 80% de l'énergie de la trame sur la bande en question. Tous les pics présents sont sélectionnés.

4.3.3 Suivi de fréquences

Sur les bandes d'analyse, le suivi de fréquences appartenant à une source doit permettre de discerner une divergence propre au phénomène de chœur.

L'étape de suivi de fréquences est réalisée trame à trame en essayant de retrouver dans deux trames consécutives, si des pics spectraux correspondent ou non à l'évolution du même phénomène et ainsi pouvoir localiser ce phénomène dans le plan temps-fréquence. La méthode utilisée est empruntée aux travaux de Tanigushi [47] et conduit à la définition de *segments sinusoïdaux*.

Nous utilisons notre propre implémentation de la méthode proposée. Dans cette méthode, chaque segment sinusoïdal est constitué d'une série de pics spectraux $p^i = (f^i, a^i)$ reliés de la manière suivante...

Par définition, un pic i d'une trame t est défini par $p_t^i = (f_t^i, a_t^i)$ et peut être lié à un pic j de la trame suivante $p_{t+1}^j = (f_{t+1}^j, a_{t+1}^j)$.

Pour décider si deux pics doivent être liés, une distance entre ces pics de trames consécutives a été proposée par Tanigushi. Cette distance est la suivante :

$$dTani_{i,j} = \sqrt{\left(\frac{f_t^i - f_{t+1}^j}{C_f}\right)^2 \times \left(\frac{amp_t^i - amp_{t+1}^j}{C_p}\right)^2} \tag{4.9}$$

Cette distance est ensuite comparée à un seuil th_{Tani} : les pics ne sont liés et n'appartiennent au même segment sinusoïdal que s'ils se trouvent à une distance suffisamment basse.

De manière schématique, nous pouvons considérer que le lien entre deux pics est acceptable si les pics se trouvent dans un voisinage proche de l'espace temps-fréquence. Ce voisinage est une ellipse dont deux diamètres sont de longueur $x.C_f$ et $x.C_p$.

Ce phénomène de suivi de fréquences est illustré sur la figure 4.14 pour deux trames d'analyse consécutives.

Les *segments sinusoïdaux* se tracent dans l'espace temps-fréquence, au sein du nuage de pics candidats. La figure 4.15 illustre ce traçage des segments.

Seuls les pics appartenant à un même segment sinusoïdal sont ensuite conservés. Néanmoins, seuls les segments suffisamment longs sont considérés comme valides. Il arrive, en effet, que deux pics de trames consécutives puissent se trouver fortuitement à une distance suffisamment faible pour être liés, par exemple des pics de bruits ayant malgré tout passé le filtrage du seuil dynamique.

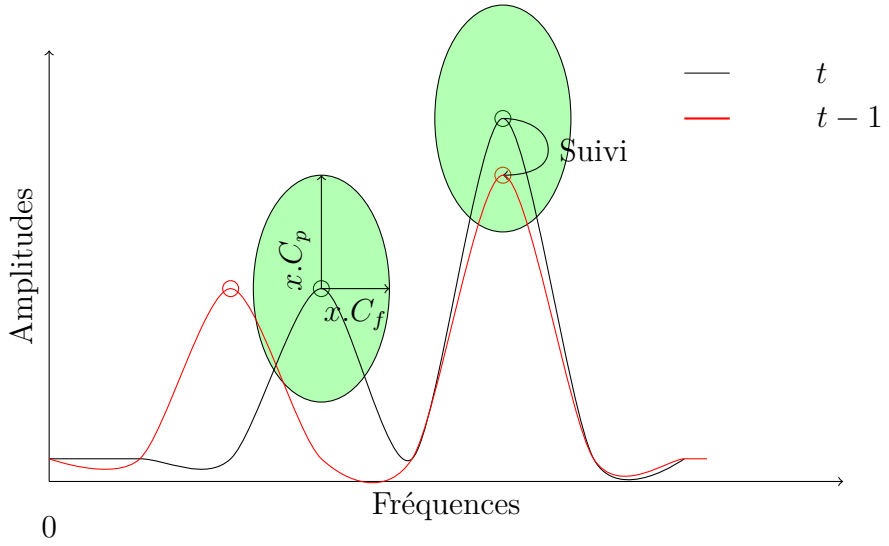


FIGURE 4.14 – Processus de suivi. Les pics des spectres de deux trames consécutives sont comparés grâce à la distance de Tanigushi.

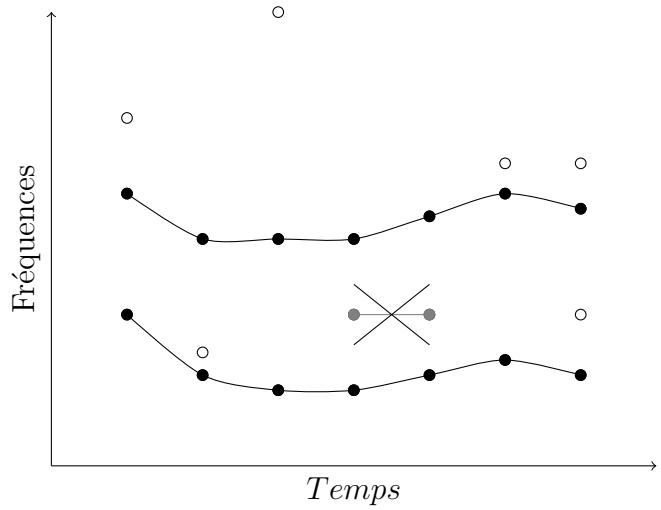


FIGURE 4.15 – Tracés des segments sinusoïdaux au sein du nuage de pics candidats. Les pics reliés (pleins) seront conservés, alors que les pics restés isolés (en creux) ne seront pas conservés pour la suite de l'analyse. Les segments trop courts ($longueur < th_{long}$) sont éliminés (ici en gris).

En revanche, il est fort improbable que ce lien fortuit se poursuive sur un nombre élevé de trames consécutives. Un seuil th_{long} sur la longueur minimale des *segments sinusoïdaux* est utilisé afin de ne garder que les plus significatifs.

Le voisinage est fixé pour être suffisamment restrictif et ne permettre que de faibles variations, rendues possibles par la grande proximité temporelle des trames d'analyse. Néanmoins, en cas de possibilité de lien avec plusieurs pics, seul le plus proche est sélectionné.

Les segments sinusoïdaux ainsi extraits suivent donc l'évolution des principaux pics du spectrogramme et permettent le suivi des zones de fortes amplitudes comme l'illustre la figure 4.16.

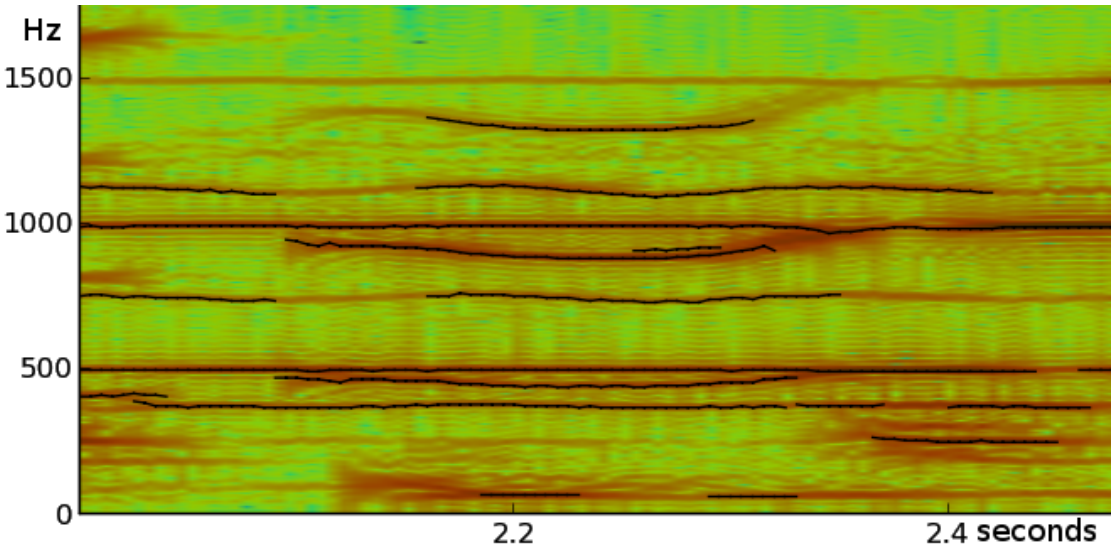


FIGURE 4.16 – Segments sinusoïdaux superposés au spectrogramme. Sur ce spectrogramme, nous observons les harmoniques des différentes sources et les segments sinusoïdaux correspondants (en noir).

4.3.4 Classification

Une fois les *segments sinusoidaux* extraits, nous considérons que chaque pic p_i localisé dans l'espace temps-fréquence par (t_i, f_i) et appartenant à un segment, correspond à l'un des chanteurs. Nous nommons $nbPts_b(t)$ la fonction retournant à chaque instant t et pour la bande d'analyse b, le nombre de pics p_i présents.

La moyenne des valeurs $nbPts_b(t)$ sur toutes les bandes est ensuite calculée, donnant une idée de la présence, à un instant t, de dédoublement se retrouvant dans les différentes bandes. Cette fonction est nommée $nbPts(t)$. Finalement, un taux de dédoublement est calculé sur l'ensemble du segment d'analyse en calculant la moyenne de $nbPts(t)$ sur l'ensemble des instants correspondant au segment d'analyse. Cette moyenne, nommée $m_{overlap}$ permet d'apprécier le taux de dédoublement présent à l'intérieur de ce segment.

Afin de classer le segment temporel dans les deux classes *Solo* et *Chœur*, une approche par seuil à partir de la valeur de $m_{overlap}$ est utilisée. Deux seuils sont proposés de manière empirique sur les premières expériences réalisées : th_{solo} et th_{choeur}.

Le mécanisme de classification est alors le suivant :

- Si $m_{overlap} < th_{solo}$ alors le segment est classé comme contenant du solo.
- Si $m_{overlap} > th_{choeur}$ alors le segment est dit contenant du chœur.
- Si $th_{solo} < m_{overlap} < th_{choeur}$ alors le segment n'est pas classé car il ne possède pas clairement les propriétés de l'une des deux catégories.

4.3.5 Conclusion

La méthode que nous avons proposée pour la détection de chœur à l'unisson est un affinage de la décision prise par un estimateur monophonie/polyphonie. Dans les zones les plus tenues et aux alentours des harmoniques détectées, nous cherchons à mettre en évidence, grâce à un suivi des fréquences, la présence de *segments sinusoïdaux* multiples, correspondant à différentes sources. L'étape de classification consiste ensuite à définir si le nombre de divergences détectées est suffisant pour conclure à la présence de plusieurs chanteurs ou qu'au contraire, leur faible présence valide la décision de la monophonie.

4.4 Détection de sources multiples

L'approche que nous présentons pour la détection de sources est une méthode originale fondée sur la détection et le suivi des fréquences prédominantes à travers le temps. Elle dérive des travaux que nous avons effectués dans le cadre de la détection solo/chœur présenté dans la partie précédente.

Le cœur de la méthode de détection consiste en une succession d'étapes visant à localiser les sources présentes dans l'espace temps-fréquence. Lorsque différentes sources harmoniques sont présentes, elles s'intercalent fréquentiellement de manière complexe ce qui peut rendre très difficile leur détection dans une analyse trame à trame. Afin d'éviter ce problème, nous développons une méthode basée sur le **suivi** de sources candidates.

Plus précisément, nous détectons les sources dans des zones où elles ne sont pas mélangées et, par continuité, nous espérons parvenir à dissocier dans les zones d'interaction la présence de plusieurs sources.

Le phénomène que nous cherchons à identifier est illustré pour un extrait de parole sur la figure 4.17. Dans cet exemple, un interlocuteur prend le relais d'un autre et les harmoniques qu'ils produisent se recouvrent temporellement. Une décision trame à trame sur les instants situés entre 6,7 et 6,8 n'est pas toujours simple

à réaliser et peut être facilement manquée, *a fortiori* par la durée très courte du phénomène. Une analyse du contexte en effectuant un suivi des fréquences clairement prédominante à l'extérieur de cette zone de superposition permet clairement d'identifier la présence de deux sources différentes. C'est cette utilisation du contexte que nous essayons de mettre en œuvre pour notre détection.

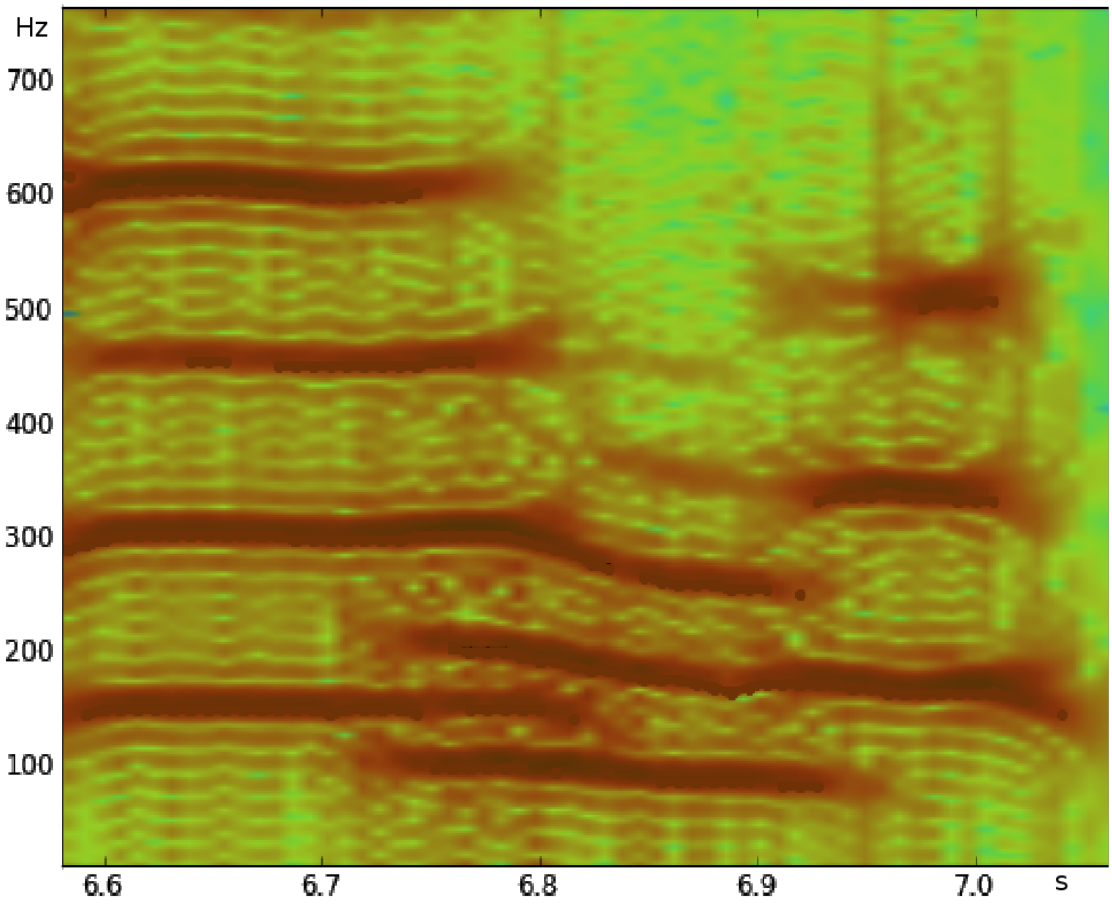


FIGURE 4.17 – Observation d'un phénomène de coexistence de deux sources harmoniques sur un extrait d'une demi seconde de parole. L'analyse en contexte au travers d'un suivi de fréquences permet la distinction entre les deux sources.

L'une des contraintes imposées à notre système est la généricité. En effet, le phénomène de recouvrement de familles harmoniques se retrouve à la fois dans des contenus parlés et musicaux, même s'il s'agit de durées et de fréquences sensiblement différentes. L'information de présence de sources multiples s'avère être structurante dans différents types de contenus, comme ceux rencontrés dans le corpus DIADEMS. Que ce soit en parole (dans les contes ou les rituels par exemple) ou de manière plus évidente en musique, l'arrivée d'une source est souvent décisive

pour déterminer la structure.

Dans ce contexte voulu générique, plusieurs paramètres sont introduits, mais leur ordre de grandeur peut sensiblement varier en fonction du type de contenu analysé. Différentes paramétrisations seront discutées dans la partie dédiée à la validation (section 4.6).

Afin de garantir un suivi le plus robuste possible, différents mécanismes ont été imaginés et mis en place.

Notre système se décompose en trois parties et il est présenté en conséquence :
- la partie **Extraction et suivi des fréquences principales** présente les techniques utilisées pour localiser les phénomènes prédominants du plan temps-fréquence pouvant être produit par l'une des sources harmoniques en présence : les *segments sinusoïdaux*,
- la partie **Détection de familles harmoniques** détaille l'analyse effectuée afin de regrouper les segments sinusoïdaux appartenant à une même source,
- la partie **Localisation de zones multi-sources** décrit comment le regroupement des familles de segments sinusoïdaux est utilisé afin de conclure ou non sur la présence de sources multiples.

4.4.1 Extraction et suivi des fréquences principales

L'extraction des segments sinusoïdaux se déroule en trois étapes :
- l'analyse fréquentielle du signal,
- l'extraction des pics les plus énergétiques de chaque trame,
- le suivi de fréquences afin d'extraire les contours des amplitudes les plus présentes.

Il est à noter que cette méthode s'appuie sur l'hypothèse de la présence de sources harmoniques dans les zones analysées. En conséquence, nous considérons uniquement l'analyse des zones sélectionnées par les méthodes présentées dans le chapitre 2. Ainsi, nous nous focaliserons sur les zones identifiées de manière suffisamment certaine comme de la parole ou de la musique.

Analyse fréquentielle

De façon classique, l'analyse est réalisée sur des trames de longueur $wLen$, et ce avec un décalage de $wStep$ pour permettre un recouvrement et une analyse temporellement plus fine.

Le spectre S_t de chaque trame est calculé sur N points grâce à un algorithme de transformée de Fourier rapide.

Sélection de pics

La première étape visant à la détection des sources consiste à identifier, trame par trame, la position des pics spectraux possédant une énergie significative. De tels pics sont considérés comme ayant été produits par une source en présence et correspond à l'une des harmoniques de la source. Cette sélection est difficile puisqu'aucune indication n'est donnée sur leur localisation fréquentielle *a priori*.

Afin de réduire les erreurs de sélection de pics, nous développons une méthode de sélection plus élaborée que celle réalisée pour la détection de chœur.

Une première sélection simple consiste à repérer les pics dont l'amplitude est strictement supérieure à l'amplitude de ses deux voisins. Or, la majorité des pics extraits de cette façon ne correspondent qu'à des parasites dus à l'enregistrement ou à l'analyse.

Pour extraire les pics les plus vraisemblablement corrélés à l'une des sources en présence, un seuil sur l'amplitude est utilisé. Cette méthode de sélection des pics doit répondre à plusieurs critères :

1. garder un nombre maximum de pics afin de pouvoir assurer une continuité des segments. En effet, le fait de ne pas sélectionner un pic appartenant réellement aux sources conduirait à un suivi interrompu raccourcissant le *segment sinusoïdal* créé.
2. sélectionner un nombre minimal de pics afin de limiter la possibilité de créer des segments liés au bruits et non aux sources,
3. être dynamiquement calculée sur chaque trame d'analyse afin d'assurer la flexibilité sur les différents types et qualités d'enregistrements analysés.

Pour répondre à ces critères, nous avons choisi, pour définir le seuil, d'utiliser une fonction linéaire par morceaux dont les paramètres dépendent du pic maximum du spectre analysé :

$$p_{max} = (f_{p_{max}}, a_{p_{max}}) \tag{4.10}$$

avec $f_{p_{max}}$ sa fréquence et $a_{p_{max}}$ son amplitude.
Cette fonction, notée $th(f)$, est définie comme suit :

$$th(f) = \begin{cases} a_{p_{max}} \left(\frac{(r_{max} - r_{deb})}{f_{p_{max}}} f + r_{deb} \right) & \textbf{pour } f \in [0, f_{p_{max}}] \\ a_{p_{max}} \left(\frac{(r_{fin} - r_{max})}{f_{max} - f_{p_{max}}} f + r_{deb} \right) & \textbf{pour } f \in [f_{p_{max}}, f_{max}] \end{cases} \tag{4.11}$$

f_{max} est la fréquence maximale de l'analyse et les constantes r_{deb}, r_{max} et r_{fin} sont les ratios définissant les points extrêmes de la courbe, pour respectivement

$f = 0$, $f = f_{p_{max}}$ et $f = f_{max}$. Seuls les pics dont l'amplitude est supérieure au seuil sont sélectionnés comme candidats.

La figure 4.18 illustre cette fonction en superposant un spectre à la fonction de seuil dont les points de changements sont explicités.

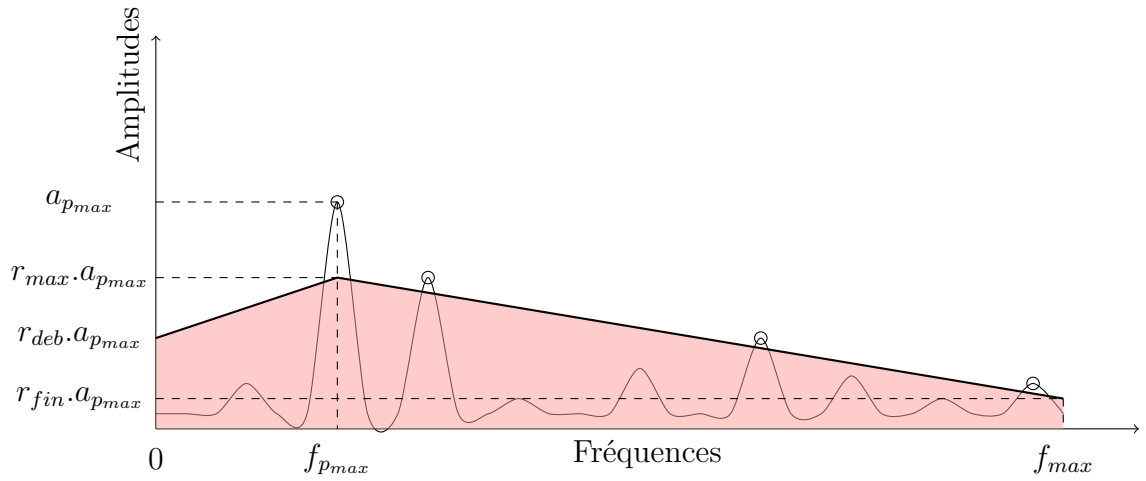


FIGURE 4.18 – Sélection des pics candidats par seuil dynamique. La fonction linéaire par morceaux utilisée comme seuil est définie à partir des coordonnées du pic principal. Seuls les pics d'amplitude supérieure au seuil sont sélectionnés : dans cet exemple, il ne reste que les 4 pics cerclés.

L'avantage principal d'utiliser un seuil adaptatif lié à la fréquence à la place d'un seuil fixe est de prendre en compte la décroissance d'énergie en haute fréquence et ainsi de faciliter la sélection de pics liés aux sources dans les hautes fréquences. Ceci permet donc de maximiser la sélection de pics liés aux sources. De plus, ce seuil étant dépendant des paramètres du pic principal du spectre, il peut s'adapter automatiquement aux conditions de l'enregistrement avec des niveaux pouvant être très différents d'une partie de l'enregistrement à un autre. Même si ce genre de cas se rencontre rarement dans les contenus produits, les enregistrements en milieu ouverts tels ceux présents au sein du projet DIADEMS peut conduire à des fichiers contenant en réalité de multiples sessions d'enregistrement aux conditions variables d'une session à l'autre. Cette technique permet donc de pouvoir s'adapter à ces variations de niveaux sans trop de difficultés.

En revanche, en cas de trop grande différence d'amplitude entre les différentes sources, une adaptation des valeurs de r_{deb}, r_{max} et r_{fin} sera nécessaire ; certaines sources peuvent ne pas avoir d'harmoniques d'amplitude suffisante pour dépasser le seuil. Dans ce cas, la source la plus énergétique masquerait les autres sources.

Cette analyse ne peut cependant être utilisée correctement que dans les zones possédant un rapport signal sur bruit suffisamment élevé, sous peine de voir un grand nombre de pics, liés au bruit, sélectionnés pour les phases suivantes.

C'est pour cette raison que la sélection des zones harmoniques doit être la plus précise possible afin que les pics extraits et les segments qui en découlent soient porteurs de sens.

Suivi

Le suivi de fréquences sert ensuite à relier les pics sélectionnés en segments sinusoïdaux de manière identique à celle décrite dans le paragraphe 4.3.3.

4.4.2 Détection de familles harmoniques

Les segments sinusoïdaux sont analysés afin d'identifier ceux pouvant être regroupés entre eux. Cette analyse se fait en deux temps :

- le calcul d'un critère d'harmonicité afin de lier les segments dont les fréquences sont harmoniques l'une par rapport à l'autre,
- un regroupement des segments afin d'étendre les rapports de liens harmoniques fréquentiellement et temporellement et relier tous les segments appartenant à une source.

4.4.3 Critère d'harmonicité

Les segments sinusoïdaux extraits sont considérés révélateurs de l'ensemble des sources ; il est important de pouvoir estimer quels sont ceux qui ont été produits par la même source. Pour cela nous nous appuyons sur le fait que les sources étant harmoniques, les fréquences observées sont des multiples de fréquences fondamentales. Bien que ces fréquences fondamentales ne soient pas toutes observées, il existe pour certaines fréquences un rapport entier entre elles. Le critère harmonique proposé, d_{harmo}, vise à évaluer si deux segments sont reliés globalement par un même rapport entier.

Le *critère harmonique* est, défini de la façon suivante :

1. Pour chaque trame t commune aux deux segments sinusoïdaux considérés, nous définissons le pic de plus grande fréquence $p_t^{haut} = (f_t^{haut}, a_t^{haut})$ et celui de plus basse fréquence $p_t^{bas} = (f_t^{bas}, a_t^{bas})$. La fonction $ratio(t)$ est ensuite calculée sur l'ensemble des trames communes aux deux segments :

$$ratio(t) = \frac{f_t^{haut}}{f_t^{bas}} \tag{4.12}$$

2. La fonction *ratio* est ensuite lissée en calculant la valeur moyenne sur des fenêtres de longueur $th_{minCommun}$ centrée sur la trame t. La valeur médiane r_{median} sur toutes les fenêtres est ensuite extraite.
3. L'écart de r_{median} à la valeur entière la plus proche est utilisé comme mesure d'harmonicité d_{harmo} entre les deux segments :

$$d_{harmo} = d(\lfloor r_{median} + 0.5 \rfloor, r_{median}) \tag{4.13}$$

La figure 4.19 montre les différentes étapes du calcul aboutissant à la mesure entre deux segments. Ces différentes étapes ont été choisies afin d'assurer un ratio entre les segments qui soit le plus proche d'un entier sur une plus grande partie du recouvrement. En effet, utiliser le ratio moyen global sur tous les instants communs pouvait conduire à la fois à de fausses alarmes comme des non détections, les différentes variations de ratio pouvant se compenser. L'utilisation de fenêtres vise à augmenter la robustesse de la mesure en ne moyennant que sur de petits voisinages avant l'analyse plus globale.

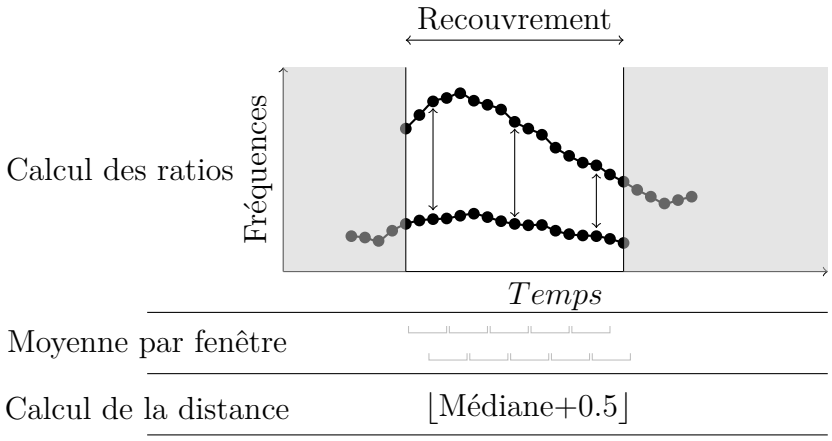


FIGURE 4.19 – Méthode de calcul de la distance entre segments sinusoïdaux pour le regroupement.

Pour ce calcul de cette mesure, nous sélectionnons tous les couples de segments sinusoïdaux ayant un minimum de $th_{minCommun}$ trames en commun.

4.4.4 Regroupement

Une fois toutes les mesures calculées, un graphe non orienté est créé dans lequel chaque nœud représente un segment sinusoïdal. Les mesures inter segments sont ensuite toutes comparées à un seuil th_{Harmo}. Si la mesure est inférieure à ce seuil,

alors un arc est tracé entre les nœuds symbolisant une relation d'harmonicité entre les segments en question.

Après avoir tracé tous les arcs, toutes les composantes connexes en sont extraites. Du fait de la relation d'harmonicité, chaque composante connexe regroupe les segments appartenant à la même source. Au sein d'une même composante, une certaine transitivité est assurée entre les différents segments.

En effet, s'il existe un rapport entier entre les fréquences des harmoniques numéros 1 et numéro 2 et entre les numéros 1 et 3, ce n'est pas le cas entre celles des fréquences 2 et 3. Le fait de créer ce graphe et d'en extraire les composantes connexes permet ainsi de regrouper toutes les harmoniques appartenant à une même famille. Le processus de regroupement par graphe est présenté sur la figure 4.20. Dans cet exemple deux familles apparaissent comme composantes connexes : la famille bleue contenant les segments $S1$, $S2$, $S4$, $S7$ et $S8$ et la famille verte contenant uniquement les segments $S6$ et $S5$. Le segment $S3$ n'étant relié à aucun autre, il ne sera plus utilisé par la suite.

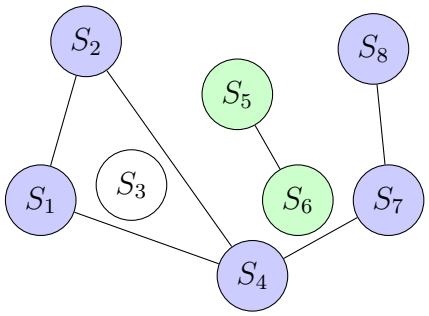


FIGURE 4.20 – Regroupement. À l'aide de tous les couples de mesures entre segments, le graphe est créé et les segments proches reliés.

En outre, la transitivité permet également de propager l'information d'appartenance à une même source à des segments ne se recouvrant pas temporellement. Ainsi même si le suivi d'une harmonique d'amplitude plus faible se trouve haché en plusieurs segments sinusoïdaux, le rapport de chacun de ces segments avec d'autres segments sinusoïdaux mieux définis permet d'assurer le regroupement de tous les suivis en une seule famille.

La figure 4.21 présente un graphe de *segments sinusoïdaux* superposé au spectrogramme correspondant. Sur cet exemple, nous distinguons clairement l'utilité de la transitivité permettant de regrouper des segments autrement séparés en deux sources (ici en gris).

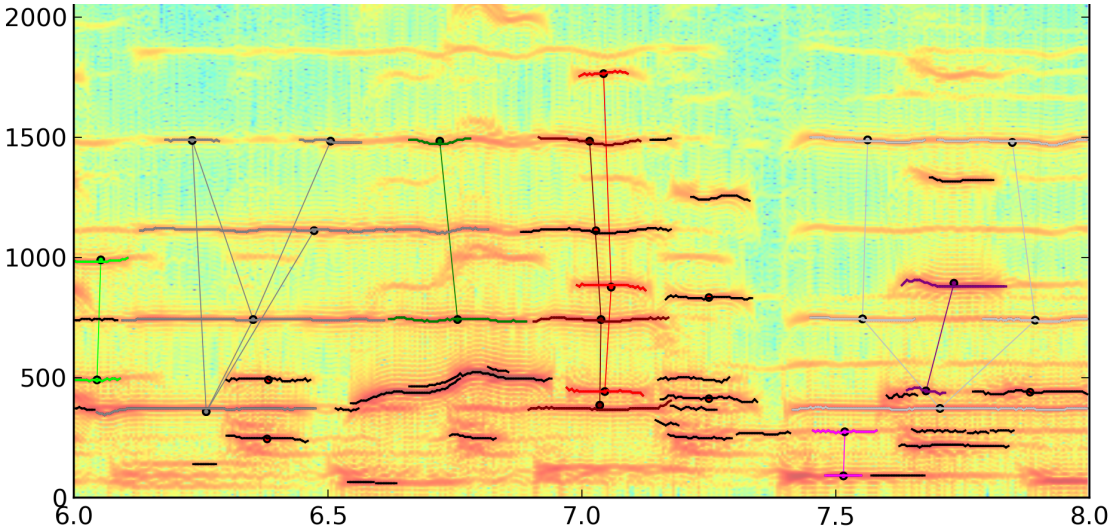


FIGURE 4.21 – Effet du regroupement. Sur cette superposition, le spectrogramme et le graphe des regroupements harmoniques sont représentés. Les segments sinusoïdaux ayant été jugés suffisamment proches sont reliés par un arc et chaque composante connexe est identifiée par sa couleur.

4.4.5 Localisation de zones multi-sources

Après l'extraction des familles harmoniques, nous utilisons les informations de début et fin des segments sinusoïdaux afin de localiser les zones de recouvrement.

Cette étape se déroule en trois parties :
- le décompte du nombre de familles par trame temporelle afin d'estimer le nombre de sources en présence à chaque instant,
- le lissage temporel afin d'éliminer les fausses alarmes et récupérer des zones non détectées en fonction du type de contenu,
- la localisation des zones de recouvrement en utilisant les valeurs lissées du nombre de familles.

4.4.6 Nombre de familles

Une composante connexe du graphe rassemble des *segments sinusoïdaux* correspondant à la réalisation d'une même source ; cette composante est appelée par la suite « famille harmonique ».

Chaque segment étant défini temporellement par un début et une fin, nous pouvons définir les limites temporelles de chacune des familles harmoniques extraites :

– l'instant de départ td_{fh} de la famille fh correspond à l'instant de départ le plus précoce des segments de fh,
– l'instant de fin tf_{fh} de la famille fh correspond à l'instant de fin le plus tardif des segments de f.

En utilisant ces informations, nous pouvons créer la fonction $nb_{clus}(t)$ qui définit pour chaque trame t le nombre de familles harmoniques présentes.

Soit N_{fh} le nombre total de familles harmoniques extraites lors de l'analyse de l'ensemble du morceau étudié :

$$nb_{clus}(t) = \sum_{fh}^{N_{fh}} famille(fh, t) \tag{4.14}$$

Avec $famille(fh, t)$ la fonction indicatrice suivante :

$$famille(f, t) = \begin{cases} 1 \ si \ t \in [td_{fh}, tf_{fh}] \\ 0 \ sinon \end{cases}$$

La figure 4.22 illustre le comportement de la fonction nb_{clus} sur un cas de recouvrement entre deux familles harmoniques. Les deux familles se recouvrent ici dans une forme de *relais* où la première famille présente laisse sa place à la seconde après le recouvrement.

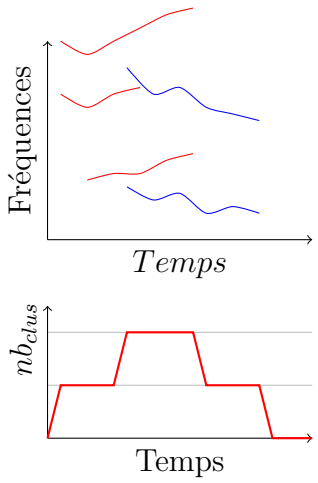


FIGURE 4.22 – Fonction nb_{clus}. Fonction de décompte du nombre de familles harmoniques présentes à un instant donné. Ici deux familles sont présentes, modélisées par leur *segments sinusoïdaux* (en rouge et en bleu).

4.4.7 Lissage

Certains phénomènes de coarticulations peuvent produire des éléments qui donnent une continuation à une famille pour quelques trames alors que celle-ci n'est en réalité plus présente. Afin d'éliminer certaines de ces petites erreurs, un lissage est effectué sur la fonction nb_{clus}.

Ceci n'existant que sur des durées très courtes, un lissage médian sur des fenêtres de largeur l_{liss} permet d'éliminer ces fausses alarmes.

Cette variable doit être réglée en fonction du type de contenu analysé. Pour la parole où les événements de recouvrement sont beaucoup plus courts qu'en musique, il convient de ne pas lisser sur de trop grosses fenêtres sous peine de manquer beaucoup de détections : même dans les contenus de débats, les différents interlocuteurs sont généralement vigilants sur le fait de ne pas trop parler en même temps que les autres. De plus, les animateurs des débats prennent également garde à ce que cela ne dure pas trop longtemps.

En musique en revanche, nous pourrons nous permettre de lisser sur de plus longues fenêtres et ainsi supprimer des fausses alarmes et ajouter des détections manquées.

La figure 4.23 illustre l'effet de la fonction de lissage sur les valeurs de nb_{clus}. Cette étape vise une fois encore à renforcer la décision à prendre en s'appuyant sur le fait que les changements du nombre de présences de sources ne peuvent être trop brefs pour être réellement considérés comme valides.

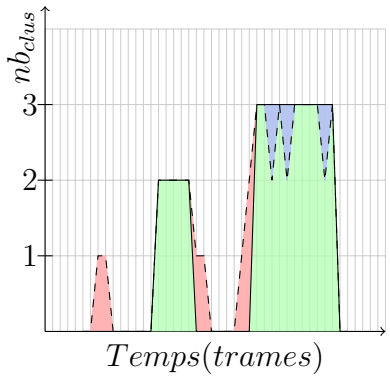


FIGURE 4.23 – Lissage médian. En utilisant un lissage médian sur une fenêtre de largeur l_{liss} sur les valeurs prises par nb_{clus}, nous évitons des fausses alarmes (en rouge) et des oublis (en bleu) sur des durées considérées trop courtes.

4.4.8 Décision

La décision de classification entre mono-source et multi-sources s'effectue en utilisant la fonction lissée nb_{clus}. Pour la détection de la parole, concentrée sur de plus petits événements, tous les instants t tels que $nb_{clus}(t) > 1$ sont détectés comme étant de la parole superposée. En ce qui concerne la musique en revanche, la longueur du phénomène nous permet d'utiliser une décision sur une fenêtre plus large, par exemple sur chaque seconde.

Cette méthode de décision permet une plus grande robustesse de la décision. Cette décision se fait selon les conditions suivantes : soit plus de deux sources sont détectés à certains instants, soit deux sources sont détectées pour une durée suffisamment longue.

Nous formalisons ces conditions avec F la fenêtre d'analyse :

- **si** $\exists t \in F, nb_{clus}(t) > 2$, **alors** F est Multi-sources,
- **si** $card(t > 1, \forall t \in F) > N_{multi}$, **alors** F est Multi-sources,
- **sinon**, F est Mono-Source.

avec N_{multi} le nombre minimal de trames comportant deux sources pour considérer F comme polyphonique.

Une nouvelle fois, les différents paramètres se fixent en fonction du contexte d'analyse, parole ou musique, dans le but d'identifier des segments de recouvrement plus ou moins courts. Pour la recherche de zones multi-sources en musique, le phénomène étant beaucoup plus long (rarement moins d'une seconde), une décision à la majorité dans un voisinage d'une seconde permet d'ajouter un nouveau lissage à la décision et ainsi éviter beaucoup d'erreurs.

4.4.9 Conclusion

La méthode que nous proposons vise à être une approche générique de détection de superpositions de sources harmoniques. L'approche est basée sur le simple suivi des fréquences de plus grandes amplitudes, et la mise en relation entre elles sur le plan harmonique. Cette approche vise à être robuste à des conditions d'enregistrement et des contenus très variés et de ce fait, cette méthode doit s'appliquer à la base de données du projet *DIADEMS*. En effet, cette base étant composée d'archives ethnomusicologiques de tout style et de toute qualité, la robustesse de l'approche est primordiale.

Nous concluons ce chapitre sur la validation des approches proposées au travers de leur comportement sur quelques enregistrements.

4.5 Validation : le chœur à l'unisson

4.5.1 Paramètres

Les paramètres utilisés pour cette expérience sont :
– une largeur de bande bw de 0.16.
les paramètres C_f, C_p et th_{tani} sont fixés aux valeurs proposées par tanigushi, à savoir respectivement : 100, 3 et 1.

4.5.2 Corpus d'étude

Pour la validation de notre approche, nous avons extrait un enregistrement du corpus DIADEMS très caractéristique du phénomène étudié ; il contient une alternance entre des zones de chant solo et des reprises en chœur à l'unisson. Il s'agit d'un enregistrement de 143 secondes réalisé sur le terrain, titré « Griots du Lamido ». Soumis à l'analyse par l'approche Monophonie/Polyphonie de l'équipe SAMOVA [48], le signal est en grande majorité estimé monophonique alors qu'il comporte à part égale des zones de solo et de chœur à l'unisson. Comme dit précédemment, l'application de notre méthode sur cet extrait est une remise en cause de cette décision *a priori* monophonique. Nous observerons le comportement du suivi sur des zones de chœur à l'unisson et de solo afin de démontrer un changement de comportement qui pourrait être discriminant.

4.5.3 Résultats et discussions

Une analyse qualitative des résultats sur cet enregistrement justifie les hypothèses que nous avons faites pour la distinction entre les zones de chœur et les zones de solo. Des exemples caractéristiques de ces deux cas sont présentés sur les figures 4.24 et 4.25. De manière générale, des embranchements apparaissent clairement dans les bandes de plus haute fréquence lorsque le chœur existe. En revanche, même lorsque nous observons des harmoniques de rang élevé, les bandes des zones de solo ne comportent bien qu'un seul suivi ne se séparant pas.

La figure 4.24 présente une longue note tenue en solo. Les segments verticaux noirs sont ceux de la segmentation forward-backward. Les segments labellisés **1** ont été rejetés de l'analyse par leur taille, ceux labellisés **2** par leur inharmonicité. Le long suivi reste bien unique même dans les bandes d'analyse de plus haute fréquence. Le gain de précision fréquentielle dans ces bandes ne révèle pas de source proche ayant pu être masquée. Quelques dédoublements surviennent entre les lobes secondaires de la 4ème harmonique, mais cela reste épisodique.

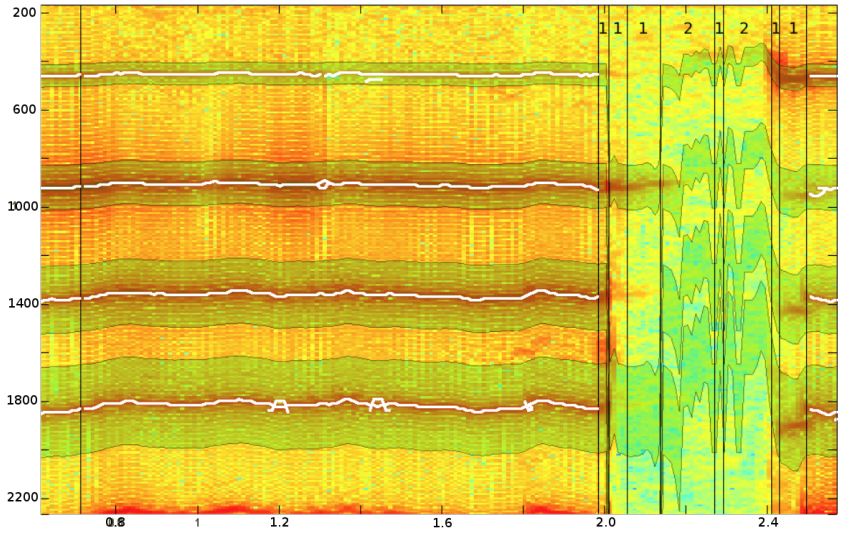


FIGURE 4.24 – Résultat du suivi de fréquences sur un extrait de solo de l'enregistrement « Griots du Lamido ». Le sens de graduation des fréquences est inversé. Sur chaque zone d'analyse nous distinguons clairement un seul suivi (en blanc).

La figure 4.25 présente quant à elle un exemple clair de chœur. Si le suivi reste unique dans les premières bandes d'analyse, les bandes de fréquences supérieures révèlent le masquage. Dans les bandes des harmoniques 2 et 4, le suivi distingue parfaitement des harmoniques différentes et les relie entre elles. Nous pouvons observer la présence « d'œillets » lors de grandes divergences (autour de 37,2 s) mais également un va-et-vient entre les différentes harmoniques formant un tissage, preuve de la proximité des différentes sources, au sens de Tanigushi.

La distinction solo/cœur est donc faisable en considérant surtout l'analyse des harmoniques de haut rang. Les erreurs d'estimation de fréquence fondamentale peuvent poser un problème dans le cas où celle-ci est sous-estimée. Dans cette situation, une bande de fréquence sur deux ne contient pas de phénomène harmonique récupérable et le suivi se fait sur une bande de bruit. En revanche, dans le cas d'une sur-estimation de $f0$, ceci ne produit pas d'erreur puisque les phénomènes restent présents dans les bandes d'analyse. De plus le fait d'utiliser un lissage des valeurs de $f0$ sur un segment homogène tend à réduire fortement ces erreurs d'estimation de fréquence fondamentale.

Comme nous l'avons dit, la méthode que nous avons développée pour la détection de superposition de sources harmoniques est générique (cf. section 4.4), mais sa mise en œuvre avec le réglage de certaines variables est différente selon le contexte de parole ou de musique. De ce fait nous étudions ci-après deux mises en œuvre différentes selon ce contexte.

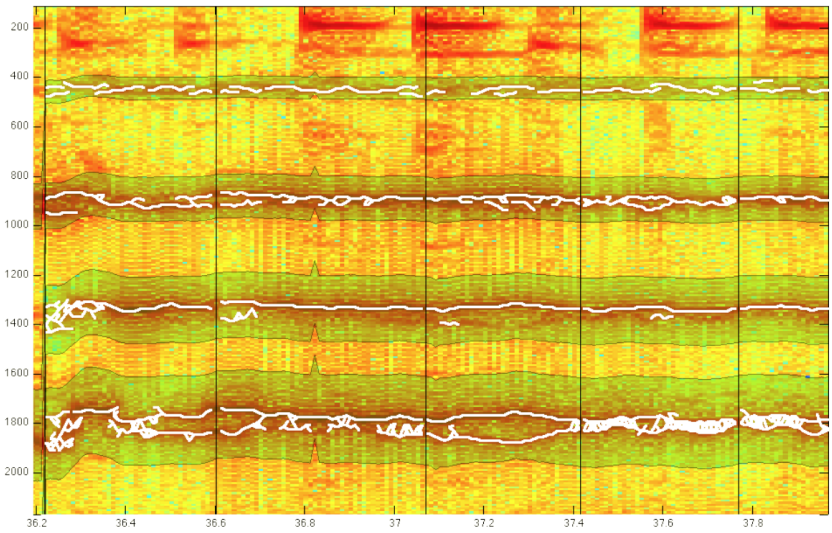


FIGURE 4.25 – Résultat du suivi de fréquences sur une partie de chœur de l'enregistrement « Griots du Lamido ». Sur les bandes d'analyse les plus élevées (en bas), nous distinguons clairement des divergences entre les harmoniques des différents chanteurs. Ces motifs sont caractéristiques de ce chant en chœur à l'unisson.

4.6 Validation : méthode de recherche de superpositions en contexte musical

Afin d'évaluer la validité de notre approche, nous allons comparer les résultats des trois systèmes que nous avons implémentés.

De plus afin de tester le comportement des résultats entre l'approche de détection de chœurs à l'unisson et l'approche détection de superpositions, nous avons utilisé le même enregistrement « Griots du Lamido » que pour la détection de chœur, à savoir 143 secondes de chant (voir section 4.5.2).

Ce fichier a été annoté manuellement entre les classes « Solo » et « Multi ». Ces classes sont représentées de manière équilibrée (52% de zones annotées « Solo » et 48% annotées « Multi ») et sont présentes en alternance. Les segments annotés « Multi » contiennent les zones de chœur à l'unisson.

4.6.1 Paramètres

Les paramètres utilisés sont :

– Pour le calcul des segments sinusoïdaux, les paramètres C_f, C_p et th_{tani} sont fixés aux valeurs proposées par tanigushi : respectivement 100, 3 et 1.

– Les paramètres de sélection des pics r_{max}, r_{deb} et r_{fin} sont fixés à 0.18, 0.20 et 0.09.
– Le recouvrement minima pour deux segments $th_{minCommun}$ est fixé à 8. Le seuil de regroupement th_{Harmo} est fixé à 1%.
– Le nombre de trame minimal polyphonique pour la décision N_{multi} est fixé à 30%.

4.6.2 Résultats de notre système

Le contexte musical nous permet de fixer la longueur de la fenêtre de décision à 2 secondes (durée de F) et prendre cette décision indépendamment d'une fenêtre à l'autre. Chaque fenêtre est identifiée comme Multi-sources si au moins la moitié des trames de cette fenêtre sont classées Multi-sources (N_{multi}). Dans le cas contraire, elle est considérée comme Solo. Les résultats de l'évaluation montrent un score de bonne classification de **69 %**.

La répartition des erreurs par rapport à la vérité terrain est décrite par la figure 4.26.

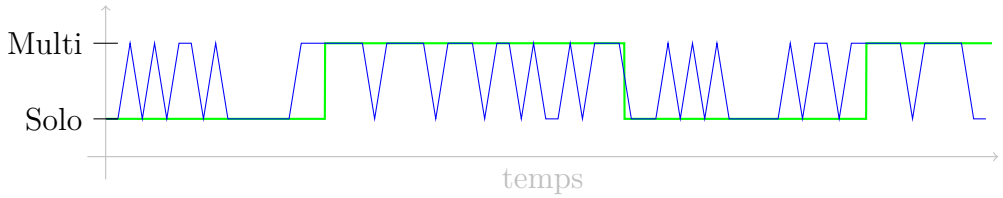


FIGURE 4.26 – Résultats de la classification de notre méthode (en bleu) par rapport à l'annotation manuelle (en vert).

Pour cette méthode, les résultats sont satisfaisants avec une décision globalement assez stable même si certaines zones sont instables avec une alternance entre les deux classes. Plusieurs zones Solo considérés comme Multi le sont à cause de la présence forte de variation rapide de la fréquence fondamentale ou à cause de la trop forte prédominance de tambour, dont certaines fréquences sont suivies et alors également identifiées comme une source. Nous pouvons estimer que ces erreurs sont explicables, mais elles posent le problème de la distinction entre le fond et les phénomènes d'intérêt, lorsque les deux se trouvent à des intensités proches.

4.6.3 Résultats du système Klapuri

Nous avons implémenté l'algorithme de Klapuri tel que décrit dans le premier paragraphe de ce chapitre, l'objectif étant la transcription du morceau analysé en un format MIDI. Pour atteindre ce but, plusieurs post-traitements avaient été ajoutés (fusion de notes identiques sur plusieurs trames, élimination de doublons, etc.). Nous avons éliminé cette partie afin de ne garder que le cœur de la méthode à laquelle nous ajoutons une contrainte de travail : deux fréquences détectées avec un rapport harmonique entre elles ne comptent que comme une seule source. Ces fréquences peuvent donc être identiques ou liées par un facteur entier, elles ne seront comptabilisées que comme une famille harmonique.

Cette contrainte supplémentaire est nécessaire pour éviter une sur-estimation du nombre de sources dans la mesure où l'algorithme retourne un nombre toujours élevé de fréquences présentes. Ces estimations étant très souvent liées, nous pouvons ne pas altérer le système principal et obtenir le résultat voulu par ce post-traitement.

Pour l'algorithme de Klapuri, nous classons en Multi-sources dès que le nombre de sources estimées est supérieur strictement à 1. Le score de bonne classification sur l'enregistrement « Griots du Lamido » est de **78%**.

La répartition des erreurs par rapport à la vérité terrain est décrite par la figure 4.27.

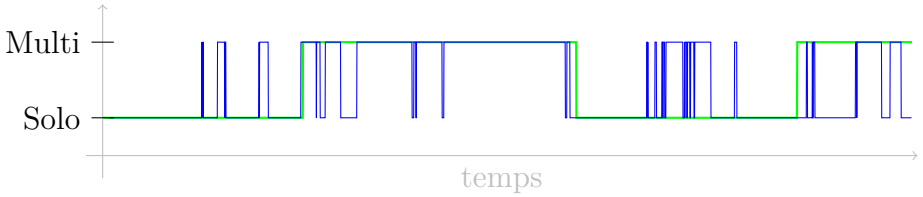


FIGURE 4.27 – Résultats de la classification de la méthode de Klapuri (en bleu) par rapport à l'annotation manuelle (en vert).

4.6.4 Résultats du système Liénard

Nous avons réalisé une version adaptée de la méthode proposée par Liénard afin de créer un détecteur Solo/Multi-sources en contexte musical. L'enchaînement des différentes étapes est présenté sur la figure 4.28.

L'enchainement des peignes est appliqué sur le spectre de chaque trame, seules les fréquences correspondant à des fréquences fondamentales présentes, se retrouvent avec une amplitude conséquente.

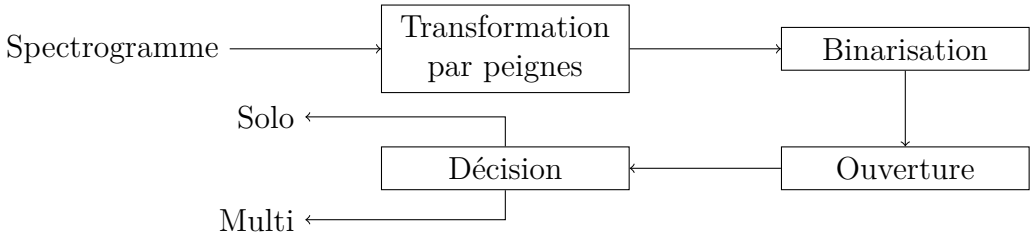


FIGURE 4.28 – Diagramme de flux de l'adaptation de la méthode d'analyse par peignes. Ce système permet, en partant du système d'extraction de fréquences fondamentales, de créer une méthode de décision Solo/Multi-sources.

Une binarisation du résultat du filtrage de chaque trame est effectuée afin de calculer le nombre de fréquences fondamentales estimées. Chaque spectre est normalisé afin que la valeur du pic maximum soit invariablement fixée à 1. Pour chaque spectre, seules les fréquences ayant une amplitude supérieure à un seuil Amp_{min} voient leur amplitude fixées à 1. Toutes les autres amplitudes sont fixées à 0. Le spectrogramme obtenu est binaire et représente l'activation des fréquences fondamentales avec le temps.

Une ouverture est effectuée sur le spectrogramme binarisé, afin d'effectuer un lissage et limiter les effets de seuil. Elle est effectuée par un élément structurant carré plein paramétré par la longueur N_{struct} de son côté. Elle permet la suppression de petits éléments isolés de taille inférieure à N_{struct} et donc d'obtenir un spectre binarisé beaucoup plus propre et débarrassé du bruit impulsionnel.

La prise de décision nécessite un ultime traitement. Même nettoyé par les deux méthodes précédentes, le spectre binarisé continue à activer des fréquences harmoniques entre elles, même en contexte de solo. Afin de prendre en compte ce phénomène, un score s_{multi} est calculé pour refléter le caractère harmonique ou non des fréquences activées par trame. Ce score est calculé comme la somme cumulée des écarts entre fréquences activées et leur arrondi à l'entier le plus proche :

$$s_{multi} = \sum_{f1}^{A} \sum_{f2}^{A} |round(|f2 - f1|) - |f2 - f1|| \tag{4.15}$$

avec A l'ensemble des fréquences restant activé sur la zone étudiée. Si les fréquences contenues dans A sont des harmoniques d'une même fréquence, cette somme tend vers 0 (théoriquement elle est égale à 0) alors qu'en présence de plusieurs sources elle est largement supérieure.

Le système de Liénard modifié, appliqué sur l'enregistrement « Griots du Lamido » donne de mauvais résultats : **50,15%** de taux de bonne classification ; il

n'y a pas de différence significative avec une décision prise au hasard. La figure 4.29 illustre la classification de cette méthode vis-à-vis de l'annotation manuelle.

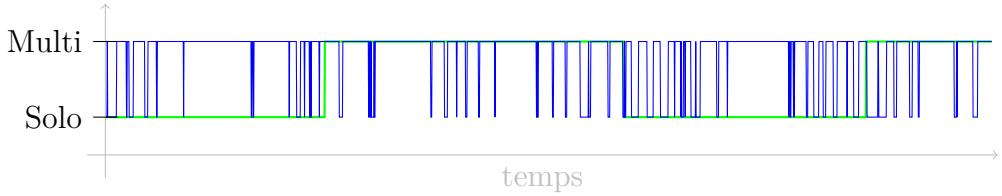


FIGURE 4.29 – Résultats de la classification de la méthode de Liénard modifiée (en bleu) par rapport à l'annotation manuelle (en vert).

Nous pouvons supposer que les résultats obtenus sur cet exemple sont mauvais en raison des conditions d'enregistrement de l'extrait. En effet, cette méthode, très fine, semble sensible aux bruits ambiants et autres perturbations dues à l'enregistrement « sur le terrain ».

4.6.5 Discussions et fusion

Puisque notre approche et celle de Klapuri fournissent des performances similaires, il est intéressant de comparer leurs erreurs. Comme illustrées sur les figures 4.27 et 4.26, les erreurs de ces deux systèmes n'ont pas lieu aux mêmes moments. Une combinaison des deux approches peut donc se révéler intéressante pour une amélioration des performances. *A contrario*, l'approche des peignes semble être trop sensible aux conditions des enregistrements et offre de mauvais résultats sur cet exemple. Nous choisissons donc de ne pas l'utiliser dans le cadre d'une fusion avec notre méthode.

Afin de comparer les deux approches, nous réalisons un histogramme. Celui-ci représente, pour chaque seconde, le nombre de trames pour chacune des valeurs possibles de sources estimées, à savoir entre 1 et 5. Ces histogrammes sont présentés sur la figure 4.30.

Le nombre de sources estimées au sein d'une même seconde peut être très variable, surtout pour l'algorithme de Klapuri. Ceci est principalement dû aux zones de silence, aux zones non harmoniques ou de transition présentes dans le signal. En revanche, notre système tend à donner un nombre de sources élevé seulement pour les zones polyphoniques et donne plus rarement plus de 2 sources simultanément.

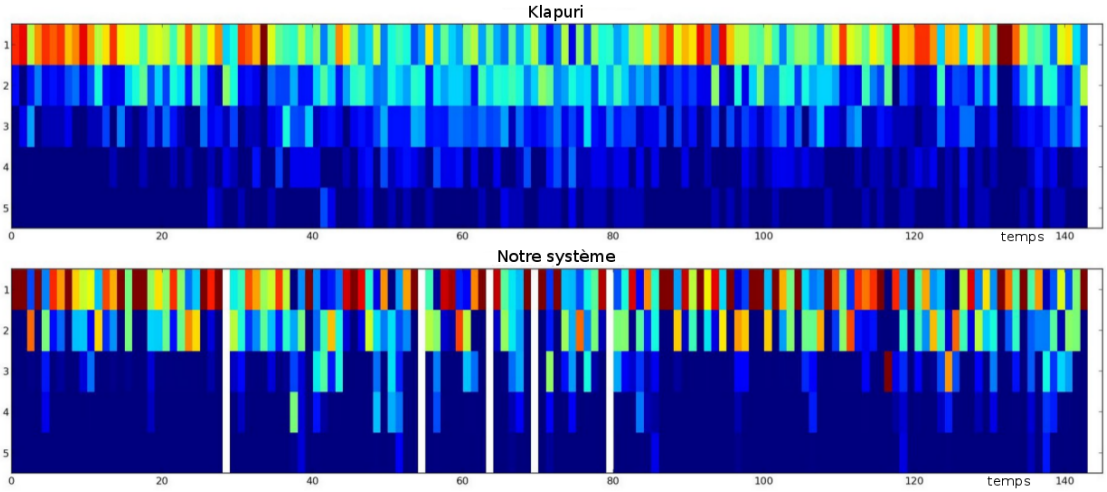


FIGURE 4.30 – Histogrames du nombre de trames par seconde associées à un nombre de sources. En haut par la méthode de Klapuri, en bas, par notre approche.

Pour obtenir une décision plus performante, nous créons une tolérance sur l'identification des zones multi-sources. Sur chaque seconde analysée, si 75% du temps une seule source est détectée, alors cette seconde est considéré comme monophonique. Sinon, elle est considérée comme contenant plusieurs sources. Cette décision est appliquée pour les deux algorithmes. En cas de désaccord, la seconde est considérée comme polyphonique. Un lissage médian sur trois secondes est ensuite effectué pour obtenir la décision finale comme illustré sur la figure 4.31.

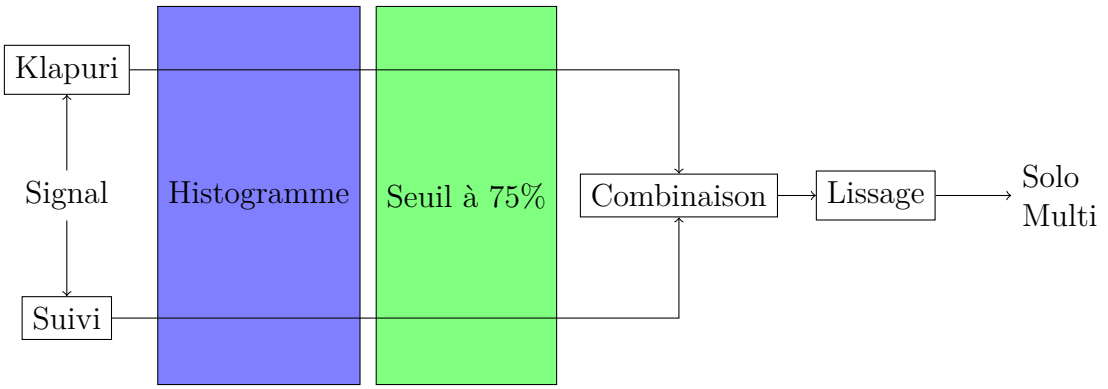


FIGURE 4.31 – Diagramme de flux de la stratégie de combinaison de notre approche avec celle de Klapuri.

Cette prise de décision conduit à une amélioration des performances. Les performances du système fusionné passent alors à 82% de bonne classification, ce qui montre bien l'apport de chacune des méthodes. La figure 4.32 illustre la répartition de l'estimation par rapport à l'annotation.

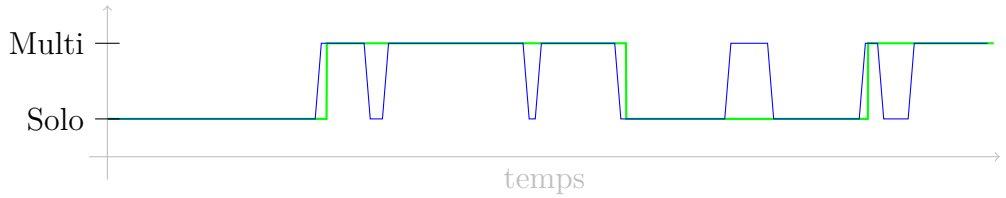


FIGURE 4.32 – Résultats de la classification de la fusion de notre méthode avec celle de Klapuri (en bleu) par rapport à l'annotation manuelle (en vert).

Cette augmentation des performances par une étude du comportement sur une seconde puis un lissage médian sur 3 secondes consécutives montre l'intérêt d'analyser le nombre de sources sur un contexte suffisamment grand en musique. En effet, une analyse trop fine peut conduire à des erreurs dans les zones de transitions qui créent des données aberrantes et isolées. L'utilisation d'un contexte de quelques secondes est de plus compréhensible en musique puisqu'il est peu probable que de réels phénomènes Solo/Multi-sources existent à l'échelle d'une seconde.

4.7 Validation de la méthode de recherche de superpositions en contexte parole

4.7.1 Paramètres

Les paramètres utilisés sont les mêmes que pour le chapitre précédent à l'exception des suivants :

- r_{max} et r_{deb} sont fixés respectivement à 0.15 et 0.09 afin de sélectionner plus de pics.
- la taille minimale de recouvrement $th_{minCommun}$ est descendue à 4 pour autoriser des segments plus courts.

4.7.2 Corpora

Pour cette expérience, nous avons utilisé les données du projet ANR ETAPE. L'une des tâches de ce projet consistant à évaluer la détection de parole superposée, une partie du corpus a été annotée en ce sens.

Le corpus initial, que nous appelons *Corpus 1* est composé de 22 heures 44 minutes de flux audio divisé en 35 fichiers. Ces fichiers durent de 16 minutes à 1 heure 6 minutes. Les types de contenu incluent des jeux, des débats et des émissions d'information.

Afin de valider plus finement les principes de notre méthode, nous avons restreint ce corpus à un ensemble de passages constitués uniquement de parole. Nous avons extraits 60 passages de *Corpus 1*. Les zones extraites durent entre 4 et 1.8 secondes et contiennent uniquement de la parole ou de la parole superposée. Les deux classes sont équitablement réparties dans ce corpus appelé par la suite *Corpus 2*.

4.7.3 Évaluation sur *Corpus 1*

Le score que nous choisissons d'utiliser est un taux de bonne détection *bd*. Il s'exprime par la formule suivante :

$$bd = \frac{\sum evenements_{ok}}{\sum evenements} \tag{4.16}$$

où $evenements_{ok}$ sont les événements de l'annotation possédant une intersection avec les événements estimés.

Ce calcul des scores est réalisé par un script fourni par l'un des partenaires du projet, le *LIMSI* et il est utilisé classiquement pour ce genre d'évaluations. Malgré notre participation à la campagne d'évaluation, les résultats que nous présentons ont été calculés en interne puisque ceux de notre soumission à la campagne n'ont jamais été dévoilés.

Résultats qualitatifs et discussions

L'analyse des résultats de notre système donne un taux de bonne détection des événements $evenements_{ok}$ de **34%**. Une partie des erreurs (fausses détections) est due à la difficulté qu'éprouve notre système de détection de la parole à estimer correctement les zones de superpositions comme de la parole. En effet, l'un des paramètres utilisé pour la détection de zones de parole est lié à la détection d'un débit syllabique de 4 Hertz. La superposition des deux locuteurs entraîne une diminution de l'énergie autour de cette fréquence, conduisant la zone à être classée comme non parole. Le taux de bonne classification parmi les événements ayant été détectés comme étant de la parole monte à **42%**.

Parallèlement, nous observons de nombreuses zones de fausses alarmes, détectées sur des zones bruitées ou contenant des zones de brouhaha en fond sonore. Nous pouvons noter que les zones de brouhaha peuvent entrer dans la définition des zones de paroles superposées, mais leur utilisation présentant peu d'intérêt,

elles n'ont pas été annotées comme telles, posant une nouvelle fois la question de l'analyse du fond sonore.

Le fort taux de fausses alarmes sur l'évaluation du corpus ÉTAPE remet en question la sensibilité au bruit de notre système. Il semble que la paramétrisation choisie, si elle permet la détection de courts événements de parole superposée implique une trop grande sensibilité au bruit. Contrairement à l'analyse de la musique, nous ne pouvons pas effectuer un lissage sur plusieurs secondes lors de la décision, puisque les événements que nous cherchons à localiser sont très courts. De manière générale, dans la majorité des zones où les locuteurs ne se détachent pas clairement du fond sonore, trop de pics sont sélectionnés et le suivi de fréquences perd tout son sens, chaque pic trouvant à se lier avec l'un de ceux de la trame voisine.

4.7.4 Évaluation sur *Corpus 2*

Protocole

La décision de classification se fait sur la présence ou non de zones détectées comme parole superposée lors de l'analyse de l'extrait. La présence de zones de superposition conduit à la classification globale de l'extrait comme *Superposé*, son absence comme *Solo*

L'évaluation sur *Corpus 2* se fait en terme de taux de bonne classification *bc* d'extraits.

$$bc = \frac{\sum extraits_{ok}}{\sum extraits} \tag{4.17}$$

où $extraits_{ok}$ désigne un extrait classé dans sa classe d'annotation.

Résultats et discussions

Sur cette deuxième évaluation de notre système, les performances de notre système atteignent 78 % de bonne classification.

Une analyse plus fine des résultats montre qu'un seul fichier de la classe *Solo* a été classé en tant que *Superposé*. Ce résultat va dans le sens de la première expérience : le système semble avoir tendance à générer des fausses alarmes. Les zones de transitions ou de bruit génèrent toujours des fausses détections.

Cette deuxième expérience semble valider la faisabilité de notre approche de détection de parole superposée avec un score acceptable de bonne classification. Elle souligne également le problème de la recherche de sources harmoniques dans les régions très courtes.

4.8 Conclusion

Dans ce chapitre nous présentons une approche de détection de sources harmoniques multiples, fondée sur un suivi des fréquences de plus fortes amplitudes dans le spectre. L'originalité de notre approche repose sur l'utilisation des rapports entre fréquences et la définition de groupes à partir d'un graphe relationnel ; chaque groupe généré représente l'empreinte d'une source harmonique. La présence de plusieurs empreintes à un instant donné est utilisée comme détecteur de sources superposées.

La mise en œuvre de cette stratégie permet, en musique, la localisation de zones multi-sources, en opposition aux solos et une variante est utilisée pour la détection de zones de chœur à l'unisson. En parole, notre approche est utilisée, avec un seuillage de détection beaucoup plus fin, pour la localisation de zones de paroles superposées.

Les expériences de validation sur des corpus réduits montrent la faisabilité de nos trois approches. Néanmoins, elles identifient également leur faiblesse : cette approche ne peut, en l'état n'être utilisée que dans les zones ou des sources harmoniques sont présentes et suffisamment prédominantes par rapport au bruit de fond. Notre système de sélection des pics fait en effet explicitement l'hypothèse que le pic de plus forte intensité appartient à une source et donc qu'une source est bien présente. Le suivi de fréquences dans des zones ne contenant pas de sources crée des suivis chaotiques ou de nombreux segments sinusoïdaux sont créés et il en découle une zone faussement détectée comme superposée.

Ce problème, relativement restreint en analyse de la musique devient critique pour l'analyse de la parole. En effet, les zones de recouvrement de la parole sont généralement très courtes : elles correspondent aux uniques instants où les sons voisés des deux locuteurs se recouvrent. Les autres correspondent au recouvrement de deux sons non voisés ou d'un son voisé et d'un son non voisé. L'ajout d'un système précis de qualification de l'harmonicité de la trame analysée semble donc primordial.

Néanmoins notre approche fournit de bons résultats sur l'exemple de validation issus du corpus DIADEMS. Une expérience plus approfondie sur les données de ce corpus est présentée dans la dernière partie de ce manuscrit.

Chapitre 5

Applications directes

5.1 Introduction

Les travaux de cette thèse, ont débutés lors de la conception du projet DIADEMS et de ce fait ils y sont extrêmement liés.

Le corpus du projet présente une très vaste hétérogénéité. Il consiste, en effet, en un très grand nombre d'enregistrements des années 1900 à nos jours, comportant des enregistrement musicaux comme des interviews et des contes, en studio comme sur le terrain, dans de nombreuses ethnies à travers le monde, comme par exemple, au Gabon, au Yémen, en Algérie, en Bolivie, au Brésil, au Burkina-Faso, en Indonésie... Une telle diversité implique également une grande disparité de qualité, liée au matériel utilisé, au passage du temps comme à celui de l'analogique au numérique. Toute cette diversité rend finalement ce corpus aussi difficile que passionnant à étudier.

Sur le plan scientifique, les préoccupations de recherche des collègues ethnomusicologues et ethnolinguistes ont été sources d'inspiration et sources de nouvelles problématiques ; nombre des contributions développées dans ce manuscrit en sont une réponse.

Néanmoins, au-delà de l'aspect scientifique, sur le plan opérationnel, il s'agit de mettre à disposition de ces mêmes collègues ces nouveaux algorithmes. Dans ce chapitre, nous présentons les différentes étapes scientifiques et techniques qui ont conduit à l'intégration de nos techniques théoriques au sein de la librairie de calcul du projet. Nous exposons deux cas d'utilisation avec un profil de type ethnomusicologue. Au moment où ces lignes sont écrites, le projet est en cours de développement et il est certain que des ajustements seront effectués d'ici sa fin en fonction des retours qu'auront effectués les ethnomusicologues.

Ce chapitre s'articule en deux parties :

- une partie décrivant la plateforme **Telemeta**, développée pour la navigation et le partage des données ethnomusicologiques ainsi que la librairie *TimeSide*, conçue pour le calcul en ligne d'outils d'indexation et de recherche. Nous y décrivons également la philosophie de développement liée à l'implémentation de nos algorithmes au sein de cette librairie,
- une seconde partie listant les cas d'utilisation du point de vue ethnomusicologue afin de montrer d'un point de vue applicatif comment nous répondons, à l'aide de nos techniques, au besoin que les chercheurs ethnomusicologues expriment.

5.2 Contexte technologique

5.2.1 *Telemeta*

L'objectif principal du projet est de fournir des outils d'aide à la recherche ethnomusicologique pour exploiter pleinement l'immense fond d'archives sonores du CNRS. Ces archives sont mises à disposition au travers de la plateforme *Telemeta*, développée par la société PARISSON. Sa construction est antérieure à celle du projet DIADEMS ; c'est un outil fédérateur de la communauté ethnomusicologue et ethnolinguiste. Sa philosophie première est de proposer aux utilisateurs un affichage clair et indexé du fond de données du projet en utilisant les informations disponibles pour chacun de ces enregistrements. Ces informations sont principalement constituées des annotations prises par les ethnomusicologues qui les ont produits lors d'enquêtes.

Le but du projet DIADEMS est d'augmenter la visibilité et l'utilisation de ce fond exceptionnel par la communauté ethnomusicologique internationale. Il a donc inclus naturellement cette plateforme comme cadre scientifique et technique de développement. Mais au-delà de la simple réalisation d'un service de diffusion, le projet DIADEMS vise à proposer des outils automatiques afin d'enrichir les informations annotées et notamment en termes de segmentation et de caractérisation du contenu.

La plateforme *Telemeta* a donc été enrichie d'outils d'indexation automatique ou semi-automatique des contenus ethnomusicologiques sonores. Il s'agit non seulement d'intégrer les outils dans le cadre informatique de la plateforme (briques logicielles compatibles...), mais de les adapter au contexte d'utilisation (paramétrisation, utilisation en mode automatique ou semi-automatique, profil de l'utilisateur...).

Telemeta est une plateforme web développée en langage *Python*, en utilisant le framework web *Django* [1]. Ce framework *Python* permet une mise en œuvre rapide de site web en utilisant un haut niveau d'abstraction et des jeux de templates de présentation. L'écriture en *Python* de la plateforme *Telemeta* permet d'intégrer directement des modules de calcul et de profiter de la puissance et du grand nombre de bibliothèques standards disponibles pour ce langage.

5.2.2 *TimeSide*

Afin de réaliser les calculs des différentes méthodes proposées, la société PARISSON a développé une bibliothèque *Python* de calcul en flux de données audio : *TimeSide* [2]. Cette bibliothèque est conçue pour fournir un cadre de travail permettant l'analyse de grandes quantités de données audio, le transcodage, la visualisation et l'interaction avec des données.

Cette bibliothèque est couplée avec le programme d'analyse audio en flux GStreamer [3] ; cette dépendance permet de réaliser des décodages de n'importe quel type de format audio standard. Elle assure également l'analyse asynchrone du flux en traitant les trames en parallèle du décodage. Elle permet enfin le ré-encodage des données vers un autre type, bien que cette fonctionnalité soit peu utile dans le cadre de notre projet.

La bibliothèque fournit des interfaces de programmation pour deux types de traitement :

- les *Graphers* permettent d'implémenter des représentations graphiques du signal pour l'affichage. Des représentations traditionnelles sont déjà implémentées comme la forme d'onde ou le spectrogramme, mais il est vraisemblable que des représentations plus spécifiques à l'analyse ethnomusicologique seront développées.
- les *Analyzers* opèrent des calculs aboutissant à la création de méta données (segmentations, courbes ou instants particuliers). Ces éléments constituent le cœur de notre travail au sein du projet.

La figure 5.1 présente les principaux objets de traitement utilisés au sein du projet. Il s'agit de la chaîne de traitement typique qui est utilisée pour chaque analyse.

1. `http://www.django-fr.org`
2. `https://github.com/yomguy/TimeSide`
3. `http://gstreamer.freedesktop.org/`

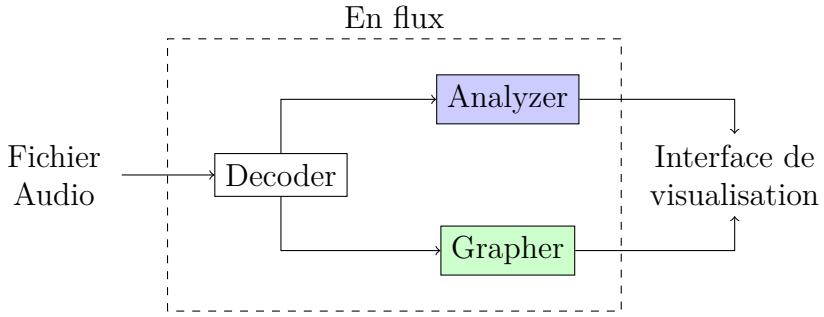


FIGURE 5.1 – Diagramme de flux classique d'une analyse par *Telemeta* et relation des différentes entités (Analyzer, Decoder and Grapher).

5.2.3 Problématiques d'intégration

Dans le contexte de calcul en ligne et en flux, de nouvelles contraintes sont posées. Afin de s'y adapter, le processus de calcul doit pouvoir se réaliser trame à trame, avec un éventuel post-traitement une fois toutes les trames analysées. Si certains des traitements que nous proposons correspondent bien à ce type de schéma, ce n'est pas toujours le cas. L'exemple le plus évident est celui de la segmentation par divergence Forward-Backward qui implique une série d'aller-retour au sein du signal.

De plus, même si les méthodes que nous proposons ne sont pas concernées, il semble évident que certaines autres méthodes ont besoin d'un accès non séquentiel à l'intégralité du signal. À cette fin, la bibliothèque permet également d'accéder à l'intégralité du signal. Néanmoins, la librairie n'étant pas optimisée pour ce type d'accès, les performances en terme de temps de calcul s'en trouvent dégradées. Or, dans un contexte de calcul en ligne il est primordial de conserver un bon niveau de performance afin de conserver une utilisation fluide de la plateforme. C'est pourquoi nous nous sommes efforcés de suivre au maximum les spécifications d'analyse en flux de la classe *Analyzer*.

5.2.4 Patron de conception

Au vue des contraintes liées à l'intégration de nos algorithmes dans le projet, j'ai proposé d'appliquer, pour le développement, le patron de conception **Modèle-Vue-Contrôleur** (cf. figure 5.2). La philosophie principale de ce patron de conception est de favoriser la modularité en séparant le calcul (le **Modèle**) de la représentation des données (la **Vue**). Tous les échanges entre ces deux entités passent par des méthodes standards définies par le **Contrôleur** (changements de paramètres, demandes de calcul ou envois de résultats).

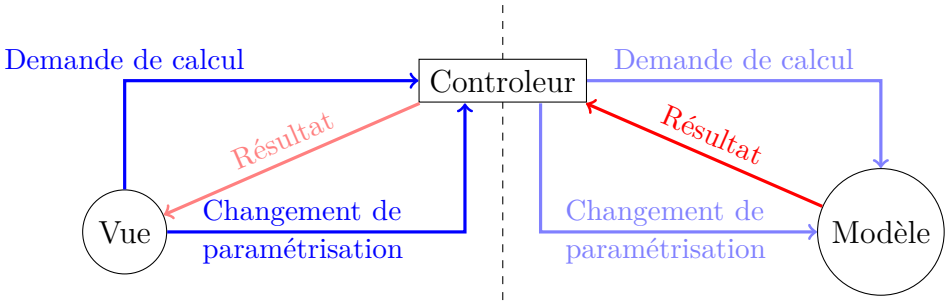


FIGURE 5.2 – Principe du patron de conception Modèle-Vue-Contrôleur. Toutes les interactions entre la Vue (affichage des données) et le Modèle (calculs) passent par le Contrôleur. Cette modularité permet de changer facilement la Vue ou le Modèle en fonction du besoin applicatif.

Ce modèle nous semble le plus adapté à l'ajout dans *TimeSide* puisqu'il décorrèle les calculs que nous implémentons, de la représentation des données qui sera, elle, gérée directement par *Telemeta*.

Afin de pouvoir les intégrer au sein de *Telemeta*, tous les algorithmes présentés dans le chapitre 2 ont été réécrits en *Python* afin de correspondre à ce patron. D'autres algorithmes de l'équipe ont également été recodés de la même façon, ce qui a conduit à une petite standardisation des outils de traitement sonore de l'équipe.

SamoPlay : un produit dérivé de *Telemeta*. L'un des principaux avantages de ce patron de conception est d'offrir une totale modularité et de pouvoir substituer un Modèle ou une Vue par une autre. Ainsi, grâce à la normalisation des algorithmes de l'équipe, et en parallèle de mes travaux de thèse, j'ai conçu un service web pour l'équipe. Celui-ci permet à tout membre identifié, en local ou à distance, d'analyser des fichiers sonores ainsi que de visualiser les résultats, en codant sa propre Vue grâce à un serveur *Python* et *Javascript*.

Ce service web, baptisé **SamoPlay**, sera prochainement accessible à l'adresse suivante : `http://samoplay.irit.fr`.

Ce service offre à l'équipe une vitrine pour ses algorithmes de traitement audio ainsi qu'une possibilité, lors de réunions à l'extérieur, de tester directement les performances de ses algorithmes sur quelques exemples ou d'en faire la démonstration (voir exemple de la figure 5.3). Sur cet exemple, le signal, son spectrogramme associé et une annotation parole/non-parole sont affichés sur la gauche. Sur la droite, se trouve une courbe de modulation de l'énergie à 4 Hertz, calculée et tracée sur l'intégralité du signal (12 secondes).

Enfin, les fonctionnalités d'ajouts de nouveaux algorithmes ont été simplifiées et documentées afin de permettre aux futurs étudiants d'intégrer leurs travaux et rendre le service d'autant plus utile et évolutif.

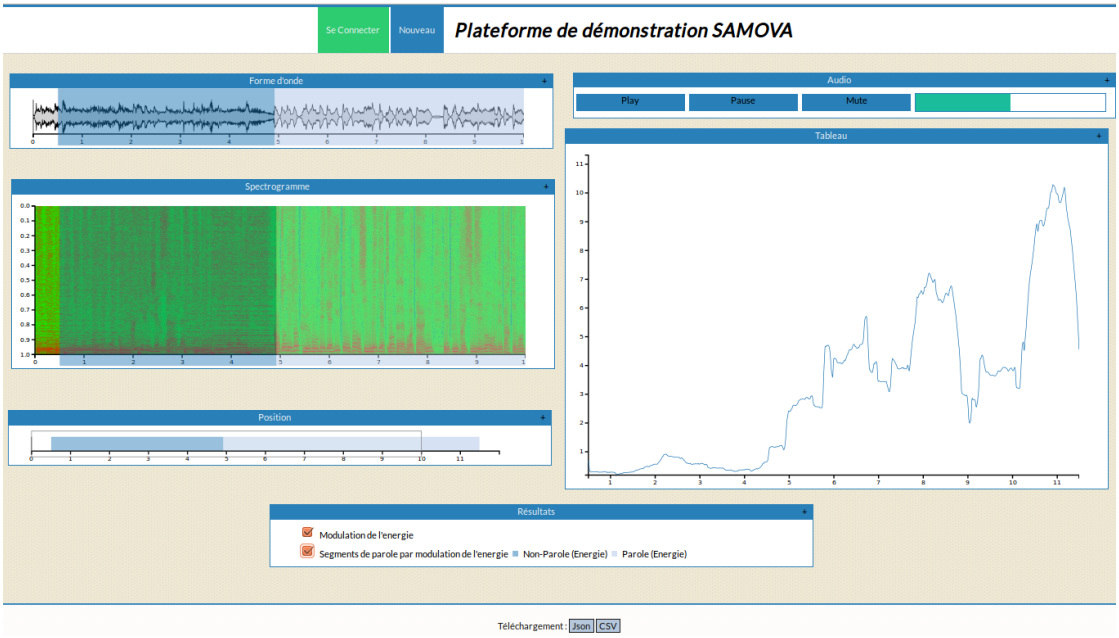


FIGURE 5.3 – Interface de visualisation développée pour le service web *SamoPlay*. Elle permet d'écouter un signal sonore, de le traiter et d'effectuer l'affichage des différents résultats de calcul.

5.3 Comportements sur les données du projet

Notre philosophie a été de proposer des approches les plus génériques possibles pour s'affranchir de la nature de l'audio étudié (musique, parole). Nous avons cependant vu que, dans certains cas, des variables d'ajustement sont nécessaires selon que nous traitons des morceaux de parole ou de musique ; nous avons alors défini selon les contextes des variables par défaut à partir de corpus restreints d'apprentissage. Néanmoins, les spécificités et l'hétérogénéité du corpus sonore du projet DIADEMS, que ce soit en termes de contenu ou de formes, laissent à penser que les approches génériques ne pourront pas éviter tous les écueils. Nous espérons que leur comportement soit suffisamment cohérent pour permettre une extraction d'informations, qui, si elle n'est pas obtenue avec la précision d'une annotation manuelle, soit suffisamment cohérente pour donner à l'utilisateur l'accès à une information pertinente.

C'est pourquoi, afin d'analyser l'adéquation entre les besoins des futurs utilisateurs du projet et nos différentes technologies de traitement sonore, nous proposons deux cas d'usage du point de vue ethnomusicologique :

- la recherche de superpositions de chanteurs,
- la caractérisation des voix intermédiaires.

Les cas d'usage ont été précisés par nos collègues ethnomusicologues ; pour chacun d'entre eux, nous précisons la chaîne de traitement et nous l'appliquons à des données sélectionnées et étiquetées, contenant les phénomènes d'intérêt repérés par les ethnomusicologues du projet.

5.3.1 Recherche de superpositions de chanteurs

Dans les archives sonores, sont enregistrées de nombreuses cérémonies au cours desquelles nous pouvons observer une alternance entre un soliste guidant un chant et des groupes reprenant ou complétant le chant. L'analyse d'une telle manifestation passe par sa structuration Solo/Chœur et il est intéressant d'identifier chacune de ses deux phases afin de caractériser leur relation pour ensuite les interpréter et comprendre le rituel associé.

Les différentes méthodes mises en œuvre sont présentées dans la figure 5.4.

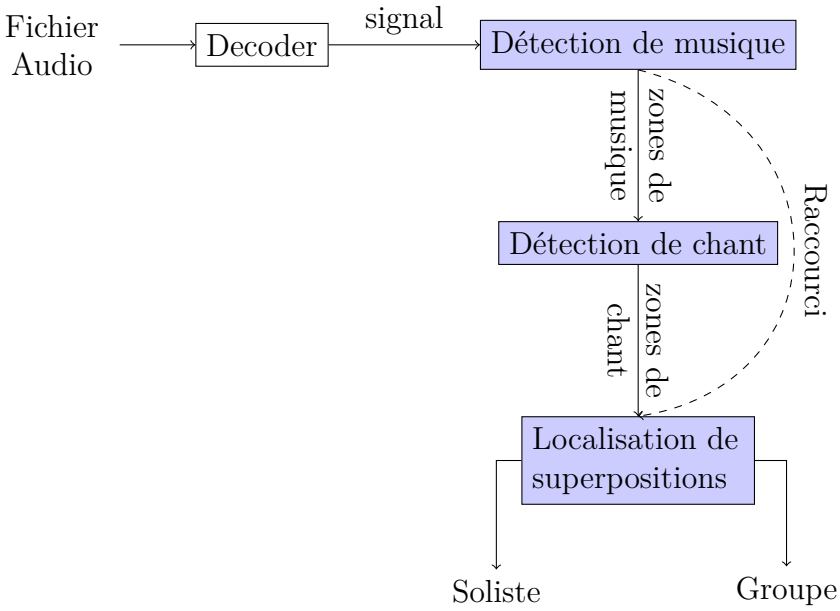


FIGURE 5.4 – Schéma de traitement en flux de la bibliothèque *TimeSide* dans son utilisation pour atteindre les passages de « Chant Solo » et « Chant chœur ».

Le repérage dans le flux audio des productions « Chant Solo » et « Chant chœur » se traduit dans un premier temps comme la recherche de superpositions ou non des sources harmoniques.

Pour ce faire, nous avons enchaîné les deux traitements suivants :
- la détection des segments de type musique,
- la localisation de zones monophoniques/polyphoniques par la méthode de localisation de superpositions en contexte musical (cf. chapitre 4.4).

À noter que ce second traitement est réalisé sur les segments de musique qu'il s'agisse de zones détectées comme du chant ou non : la détection du chant pourra se faire en parallèle ou ultérieurement.

De plus, puisque les différents enregistrements pointés par les ethnomusicologues comme intéressants dans la recherche de superpositions ne comportent que de la musique chantée, nous n'avons pas effectué, pour nos cas d'utilisation, de détection de chant : ceci explique la flèche « Raccourci » sur la figure 5.4.

Nous nous sommes focalisés sur le résultat des suivis de fréquences afin d'en appréhender les points forts et les points faibles. Les différents paramètres de notre méthode appliquée à la musique, présentée dans le chapitre précédent, ont été conservés à l'identique.

Points forts de la chaîne de traitement

La phase de détection de musique fonctionne correctement avec une bonne localisation des zones de musique harmonique. Malgré certaines erreurs sur les zones d'annonces (parlées), la segmentation exclut les zones ne comportant que des percussions et englobe la totalité des zones chantées.

Cette segmentation est intéressante dans la mesure où elle permet clairement d'analyser par suivi des zones comportant majoritairement une source harmonique, hypothèse fondamentale pour la suite des traitements.

La détection de zones monophoniques par la méthode recherche de superposition se passe généralement bien avec un bon suivi des zones prédominantes en contexte non bruité (voir exemple sur la figure 5.5). La méthode regroupe généralement plusieurs harmoniques ce qui renforce la fiabilité de ce choix.

Le suivi des différentes sources en zones polyphoniques fonctionne correctement si les différents chanteurs chantent à des niveaux semblables et qu'il n'y a pas trop de bruit de fond parasite. La figure 5.6 illustre un passage où les deux chanteurs sont particulièrement bien séparés : nous distinguons clairement les différentes voix regroupées.

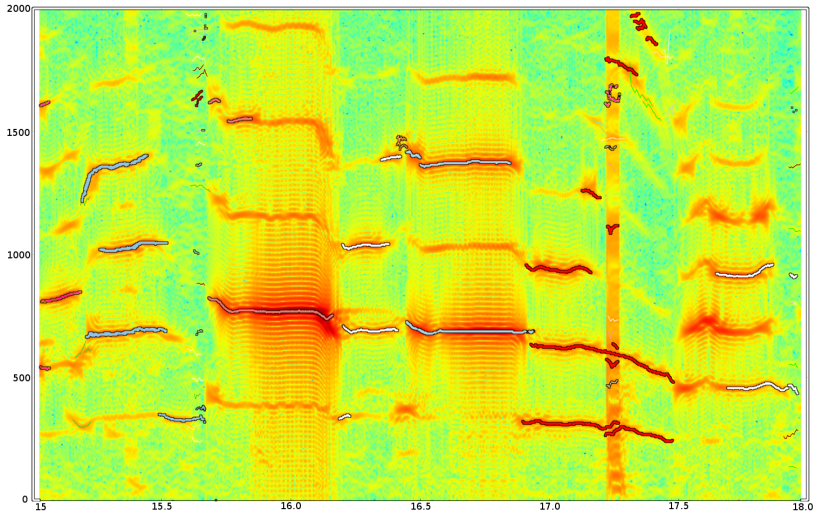


FIGURE 5.5 – Suivi de fréquences sur zone monophonique de 3 secondes. Plusieurs harmoniques du chanteur ont correctement été suivies et le regroupement s'effectue en une seule famille.

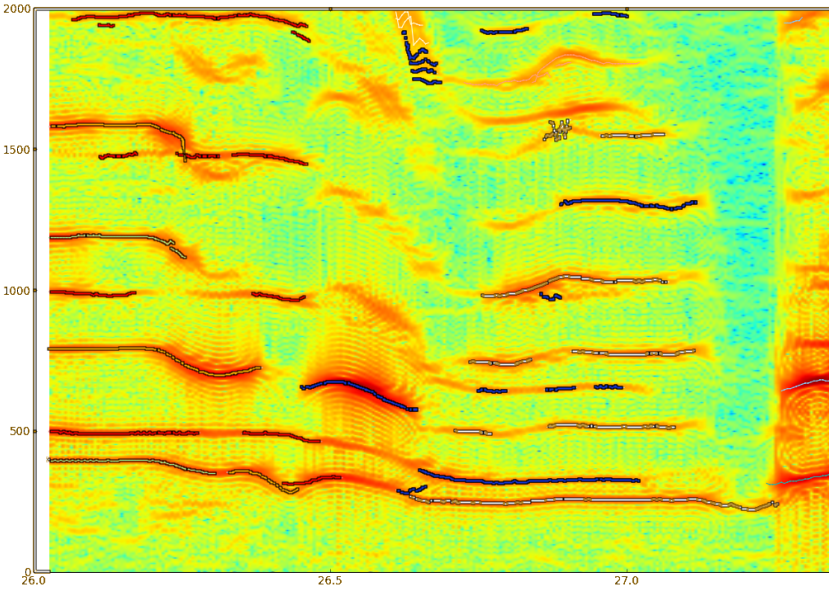


FIGURE 5.6 – Suivi de fréquences sur zone polyphonique de 1,5 secondes. Les différentes sources sont bien suivies et regroupées. Les différentes couleurs illustrent les différentes familles harmoniques.

Points faibles de la chaîne de traitement

L'un des points faibles de notre système de détection de superpositions, mis en valeur sur ces extraits, est la difficulté d'obtenir les segments incluant des sources secondaires, lorsque celles-ci sont trop différentes en terme d'intensité ce qui implique des harmoniques trop faibles.

Le seuillage de sélection des pics, conçu pour ne pas accumuler trop de pics liés au bruit de fond, dessert la méthode en ne sélectionnant plus assez de pics et en masquant certaines harmoniques lors du regroupement. Si cet effet peut conduire à « oublier » des sources, il peut également en résulter une sur- estimation artificielle du nombre de sources.

En effet, si le suivi correspondant au f_0 n'est pas correctement effectué, alors le lien entre les harmoniques paires et impaires peut ne plus se faire. Il en résulte une détection de deux groupes distincts qui, pourtant, correspondent à la même source.

Ce phénomène est particulièrement visible dans la figure 5.7 où une même source est séparée en trois groupes, principalement entre les instants 15,5 et 16 !

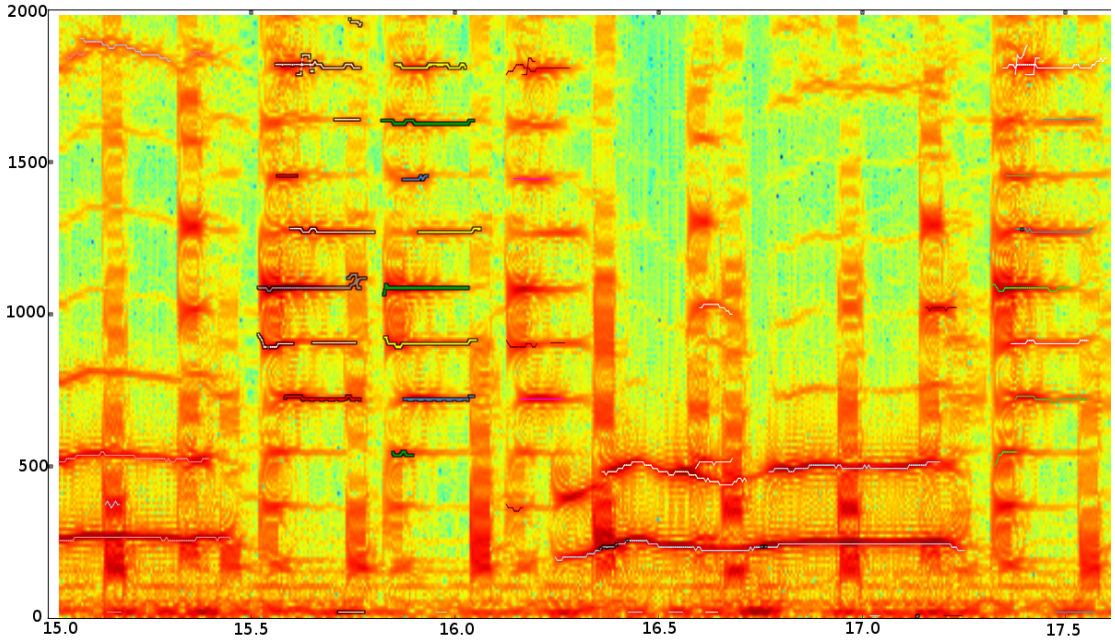


FIGURE 5.7 – Erreurs de regroupement par manque de suivi sur les premières harmoniques. La seule source présente est alors séparée de manière erronée en trois groupes entre les instants 15,5 et 16. Les différentes couleurs illustrent les différentes groupes harmoniques. Chaque groupe constitue une partie des harmoniques de la même source.

La technique de regroupement peut également montrer ses limites dans les zones où sont présents des sons harmoniques, de basse fréquence ; leur suivi fait qu'ils peuvent servir de point de jonction entre deux familles qui ne seraient pas liées par ailleurs. Plus le son est de basse fréquence et plus ce risque est grand.

L'exemple présenté par la figure 5.8 illustre bien ce problème : de nombreux segments générés par le tambour permettent de faire le lien entre les harmoniques de deux familles qui ne devraient pas être liées.

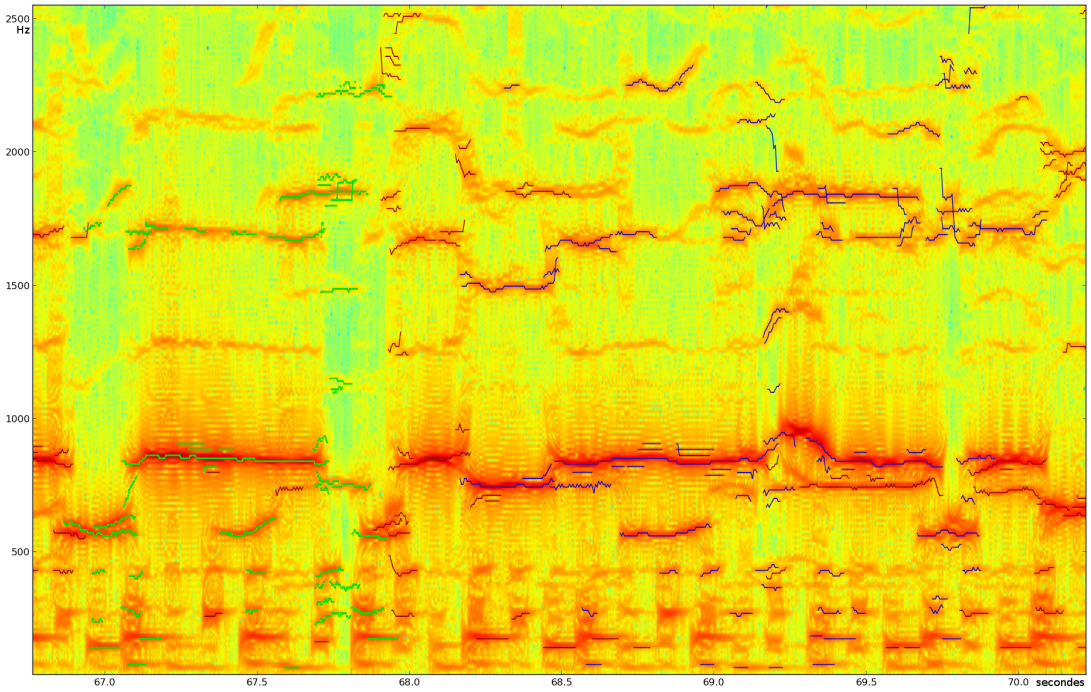


FIGURE 5.8 – Erreurs de regroupement en une seule famille de deux sources. Les suivis de très basses fréquences générées par le tambour font un pont entre les deux familles d'harmoniques.

De manière générale, la paramétrisation de la sélection de pics semble être le point le plus sensible de notre méthode. En fonction du contexte sonore, de la différence entre l'énergie des différentes sources, voir entre les harmoniques d'une même source, la sélection peut engendrer des erreurs qui ne peuvent être rattrapées par les étapes suivantes.

5.3.2 Voix intermédiaires

Ce deuxième usage a émergé au cours du projet et témoigne de l'intérêt des outils de traitement développés à des fins de recherche en ethnomusicologie et ethnolinguistique.

La définition des différents types de voix intermédiaires est un problème actuellement en débat au sein des ethnomusicologues et ethnolinguistes : il est question de voix parlée, récitée, racontée, psalmodiée, chantée... Mais les caractéristiques de ces différents types de voix ne sont pas clairement établies. Les descriptions faites par les chercheurs ont conduit à penser que le rythme pouvait être un indice pertinent dans cette caractérisation. Afin de les aider dans cette exploration, nous avons proposé une chaîne de traitement en flux qui permet de visualiser le « Tempogramme » de la parole. Comme cela a été décrit au chapitre 3, la suite des traitements inclut la segmentation de divergence Forward-Backward et le calcul du tempogramme lui-même (voir figure 5.9). Cette chaîne de traitement a été appliquée à plusieurs types de voix, caractérisées *a priori* par les chercheurs.

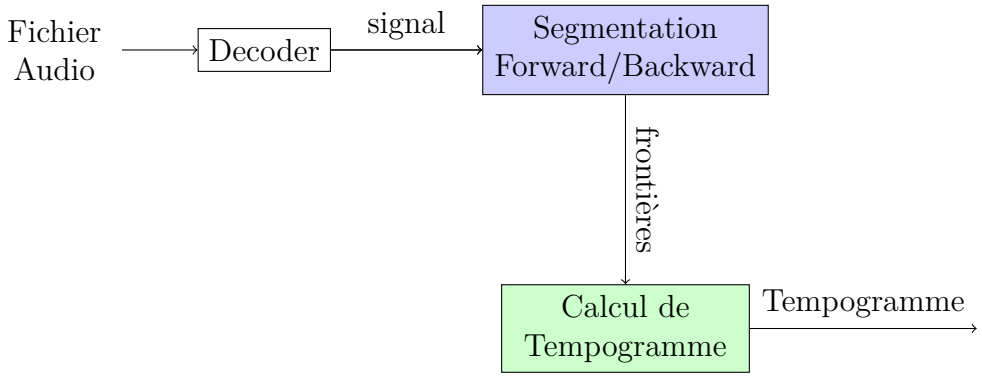


FIGURE 5.9 – Schéma de traitement en flux de la bibliothèque *TimeSide* dans son utilisation pour la caractérisation des voix intermédiaires.

Quelques résultats intéressants et prometteurs

Le résultat le plus intéressant de cette étude qualitative est la distinction de deux sous-groupes au sein des enregistrements étiquetés par les ethnomusicologues comme contenant de la voix psalmodiée. Au sein de cette catégorie, qui regroupe notamment des enregistrements de prières religieuses, nous distinguons deux catégories :

1. celle où la rythmique est importante ; la parole est scandée à rythme fixe,
2. celle où les variations de hauteurs de fréquence fondamentale sont accentuées et sans régularité rythmique particulière.

La figure 5.10 montre le tempogramme d'un enregistrement de psalmodie avec rythme fixe. Ici nous voyons nettement cette régularité avec une forte énergie de manière constante autour de 1 Hertz. Nous distinguons également clairement que l'enregistrement comporte deux parties de rythme différent, la jonction s'effectuant autour de la 45ième seconde.

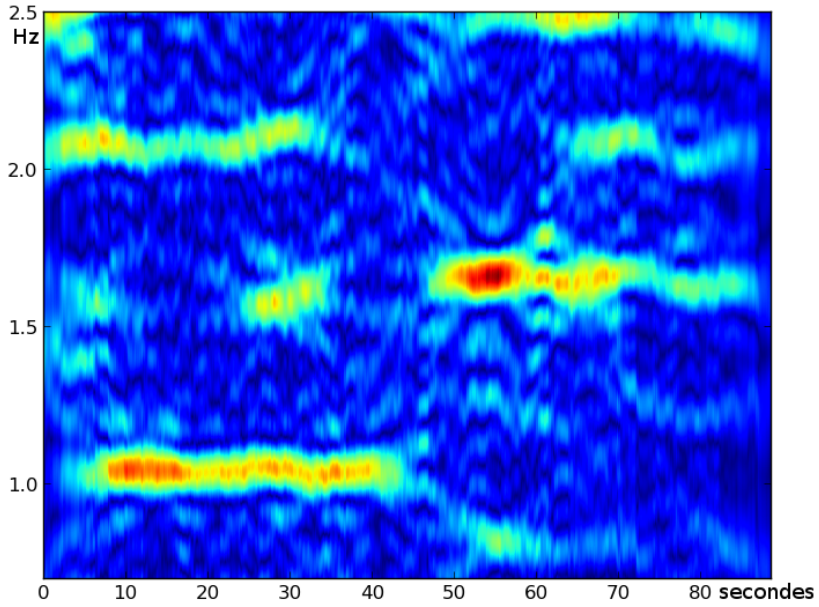


FIGURE 5.10 – Tempogramme d'un enregistrement de psalmodie enchaînant deux rythmes différents. Le passage d'un rythme à un autre, à partir de la 45ième seconde, est clairement visible.

Il est important de noter que cet enregistrement, soumis aux détecteurs de parole et de musique (cf. section 2.2) est sur sa grande majorité classé comme « Parole » (76% du temps) et « Non musique » (86% du temps). Cette configuration laisse donc à penser que la classe « Psalmodie rythmique » pourrait être identifiée en recherchant des zones d'enregistrement à tempo stable tout en comportant de la parole et non de la musique.

La sous-classe de psalmodie non rythmée semble clairement se rapprocher du chant avec un allongement très marqué des voyelles et la présence de vibrato. Cette propriété est confirmée par le détecteur de chant : 26% du temps des fichiers monophoniques. Ces mêmes fichiers sont également considérés comme contenant de la musique sur 80% du temps. Cette sous-catégorie semble donc difficilement séparable de la voix chantée en l'état de nos méthodes car elle en partage la plupart des caractéristiques.

En revanche, le tempogramme obtenu lors de l'analyse ne révèle pas de réelle structure rythmique, et son aspect est proche de celui obtenu sur les contenus récités/racontés.

Les autres catégories de voix intermédiaires offrent toutes des tempogrammes d'un aspect proche dans lequel ne se détache pas de régularité rythmique. Les figures 5.11 et 5.12 montrent les tempogrammes issus de l'analyse respectivement d'un enregistrement de type « raconté » et de type « récité ».

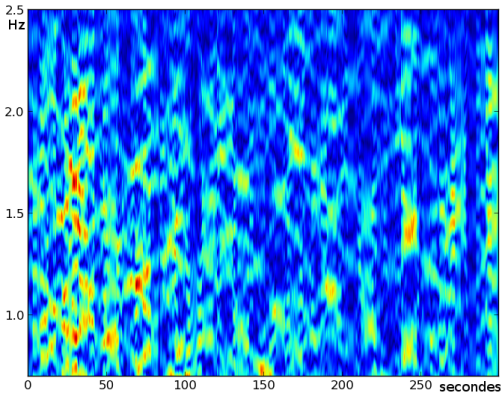


FIGURE 5.11 – Tempogramme d'un enregistrement de type « Raconté ».

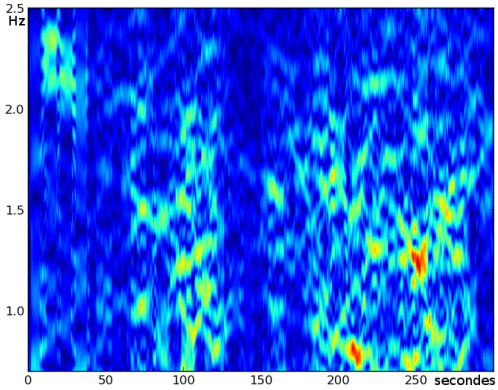


FIGURE 5.12 – Tempogramme d'un enregistrement de type « Récité ».

Cette analyse a permis de valider une hypothèse sur l'importance du rythme pour une sous-catégorie de voix psalmodiée mais également d'affirmer que la seule information rythmique ne suffit pas à clairement délimiter ces trois classes. Dans le même esprit d'exploration/validation d'hypothèses, une analyse d'autres paramètres prosodiques, comme le débit syllabique, la fréquence fondamentale, la durée des silences... pourrait valider d'autres caractéristiques pertinentes.

5.3.3 Conclusion et améliorations possibles

Dans ce chapitre, nous avons appliqué les méthodes développées lors de cette thèse dans le cadre de cas d'utilisation du projet DIADEMS. Nous avons cherché à savoir si nos algorithmes, appliqués aux données du projet pouvaient être une aide et apportaient des informations dans des cas concrets d'analyse par les ethnomusicologues et ethnolinguistes.

Si le suivi de fréquences montre ses limites en cas de milieu trop bruité, il reste intéressant si le niveau de bruit est bas et les différentes sources d'intensité proches. L'analyse par Tempogramme parvient également à montrer son apport dans le cadre de l'analyse de voix psalmodiée : la régularité rythmique apparaît clairement et nous détectons même la présence de plusieurs phases correspondant

à des changements de rythme du locuteur. En revanche, il est mis en évidence que le rythme n'est pas suffisant pour la distinction de toutes les classes.

Il nous semble donc que nos approches, même sans être parfaitement robustes, peuvent contribuer à l'extraction d'informations d'intérêt et ces quelques cas d'usage ouvrent de nouvelles perspectives.

Pour une meilleure analyse par suivi de fréquences, un pré-traitement pourrait être envisagé afin d'aiguiller automatiquement la paramétrisation vers certaines valeurs en fonction de paramètres de plus bas niveau (niveau de bruit, harmonicité...) et ainsi optimiser les résultats.

Il semble intéressant à l'avenir de se pencher sur des techniques de débruitage du signal afin de limiter l'impact du fond sonore qui peut lui aussi comporter des sources harmoniques mais dont l'analyse n'est pas d'intérêt dans notre contexte. Cette problématique ouvre la voie à la problématique de la distinction entre la partie « intéressante » du signal et celle relevant de l'« ambiance ». Que ce soit d'un point de vue du traitement de signal comme de l'analyse ethnologique, la distinction semble dépendre des objectifs et est clairement un problème délicat à aborder.

Chapitre 6

Conclusion et Perspectives

6.1 Conclusion

Dans ce travail de thèse, nous avons exploré diverses méthodes d'indexation des contenus sonores musicaux comme de parole. Nous avons concentré notre recherche sur l'estimation de rythmes et la détection de zones de présences conjointes de sources harmoniques. En privilégiant un principe de généricité, nous nous sommes focalisés sur des analyses extrayant des informations communes et pertinentes à la fois sur les contenus de parole comme sur les contenus de musique. Seules les variables de mise en œuvre des algorithmes ont été adaptées au contexte sonore. L'environnement de développement, réunissant différents algorithmes d'extraction audio de l'équipe SAMOVA, a permis de finaliser un système d'indexation automatique de référence, puis de l'enrichir par l'ensemble des contributions de cette thèse. Le contexte applicatif du projet DIADEMS d'indexation de données ethnomusicologiques a non seulement permis de valider nos méthodes sur des extraits du corpus du projet afin d'en définir les points forts et les sources d'amélioration ; mais il a également démontré une fois encore l'intérêt d'un véritable travail pluridisciplinaire.

6.1.1 Vers un système d'indexation complet

Un des objectifs étant la réalisation d'un système d'indexation à plusieurs niveaux, une grande partie de nos travaux de thèse a été consacrée à la standardisation de plusieurs méthodes d'indexation développées antérieurement par l'équipe SAMoVA. Ce même formalisme a été appliqué aux méthodes que nous avons proposées pour enrichir ce système.

Le système d'indexation ainsi défini est unifié, utilisable facilement, tout en ayant une interaction simplifiée : il permet de créer une chaîne de traitement constituée de la séquence de n'importe lesquels de nos algorithmes. Un exemple au

travers de la chaîne de traitement pour la détection de superpositions harmoniques a été détaillée dans le chapitre 2. Elle est constituée de l'enchaînement d'un détecteur Parole / Musique suivi d'une classification Monophonie / Polyphonie, pour enfin terminer par la recherche de superpositions. Une autre chaîne semblable peut facilement être construite en mettant en œuvre la détection de rythme.

6.1.2 Rythme

La méthode que nous proposons pour l'estimation du rythme repose sur une segmentation en zones homogènes du signal. Cette méthode, initialement conçue pour la segmentation de la parole en zones sub-phonétiques permet un découpage de la musique en segments proches des phases de la note. Dans ces deux cas de parole et de musique, les unités atteintes par ce découpage sont les unités de base de la structure. Notre contribution porte sur l'analyse fréquentielle de cette segmentation. Nous proposons une transformation de Fourier des instants de rupture ainsi détectés, pondérés par la variation locale de l'énergie. Nous appelons ce résultat de l'analyse de Fourier : *spectre de rythme*.

La première application directe de ce processus est l'estimation de la valeur du rythme. En effet, calculées sur l'intégralité d'un morceau, les fréquences les plus énergétiques du spectre de rythme sont toutes fortement liées au tempo. Néanmoins, la composante fréquentielle la plus énergétique ne correspondant pas toujours à la valeur correcte du tempo, nous avons mis en place différentes méthodes de décision afin d'éviter d'atteindre le multiple ou sous-multiple de la réalité rythmique. Cette approche a montré de très bonnes performances sur des contenus de musique au rythme très marqué.

Nous avons prolongé l'utilisation possible du spectre de rythme en effectuant l'analyse sur des fenêtres glissantes (à la manière d'un spectrogramme), nous avons obtenu des « *tempogrammes* » avec lesquels nous pouvons effectuer le suivi du tempo pour des morceaux au tempo changeant.

Ces deux analyses servent à estimer la valeur du rythme dès lors qu'un rythme existe. Mais nous pouvons également utiliser cette analyse afin de déceler des zones comportant un rythme. Nous avons ainsi montré des différences dans la structure rythmique entre différents types de voix : parlée, lue et scandée.

Une validation en estimation de tempo en musique et en recherche de zones à rythme stable en parole a été présentée.

6.1.3 Superpositions harmoniques

Les méthodes de détection de zones de superpositions de sources harmoniques que nous avons proposées sont basées sur un suivi des fréquences prédominantes

trame à trame. Le contexte d'application oblige à une exploitation différenciée de ce suivi :

- dans les zones identifiées comme musique monophonique, une vérification par suivi de fréquences permet, notamment sur les harmoniques de haut rang, de mettre en évidence la présence ou non de chœurs. En effet, les petites divergences entre les chanteurs ressortent, malgré leur volonté de chanter à la même fréquence. La présence de divergences dans le suivi ainsi que la présence de suivis multiples autour d'une même harmonique se révèlent caractéristiques de ce type de chant.
- Dans les zones de musique comme de parole, le suivi des fréquences est couplé à un regroupement, sur des critères harmoniques. Cette mise en relation entre les suivis conduit à la définition d'un graphe dont les composantes permettent l'identification de la trace fréquentio-temporelle d'une source harmonique. En fonction du contexte, elle caractérise donc un locuteur ou un instrument. La présence simultanée de plusieurs sources est ensuite utilisée comme détecteur de zones de superpositions.

Nous avons montré que le suivi des fréquences prédominantes et leur groupement par des critères harmoniques est une bonne empreinte d'une source harmonique. Si cette approche montre ses limites en contexte trop bruité, (nous n'avons utilisé que des contraintes simples pour la sélection des pics pouvant être reliés), il est certain que des traitements plus élaborés peuvent grandement augmenter la fiabilité du suivi et sa robustesse au bruit. Cette caractérisation des sources propose l'avantage de pouvoir être utilisée pour identifier des sources musicales comme des sources de parole sans apprentissage spécifique de leurs caractéristiques.

Une validation de notre approche sur des contenus de parole et de musique est présentée afin d'appuyer nos théories.

6.1.4 Applications et mise en œuvre au sein de DIADEMS

Le contexte applicatif propre au projet DIADEMS ainsi que les contraintes techniques qu'il impose a été une grande source d'approfondissement sur le plan scientifique, informatique et en termes d'usage.

Les besoins des ethnomusicologues en termes d'outils ont permis d'avancer dans la réalisation du système d'indexation, de privilégier le développement d'outils et d'en clarifier les vrais usages. Les résultats de nos algorithmes sur les enregistrements d'intérêt pointés par les ethnomusicologues du projet montrent des résultats prometteurs et marquent également les configurations qui peuvent poser problèmes. Le niveau de bruit de fond, et particulièrement s'il contient des sources harmoniques (brouhaha, instruments d'accompagnement...) est très problématique pour notre analyse de superpositions. La détection du rythme fonctionne bien

même si elle ne peut suffire à elle seule à identifier un type précis de parole.

Pour que nos outils soient utilisables au sein du projet, nous avons dû les intégrer à la librairie de calcul *TimeSide*. Cette librairie de calcul en flux a imposé un décloisonnement entre les différents modules de calcul et nous a fourni l'opportunité de standardiser différents algorithmes d'analyse audio utilisés au sein de l'équipe. Cette standardisation s'est effectuée autant sur le plan du langage (puisque tous les algorithmes ont été recodés en *Python*) que sur celui de l'ingénierie (en utilisant le patron de conception Modèle - Vue - Contrôleur).

Ce travail nous a offert l'occasion de regrouper une grande partie des algorithmes de traitement audio de l'équipe pour les intégrer au sein d'un service de calcul en ligne propre à SAMOVA.

6.2 Perspectives

Les travaux effectués lors de ce doctorat ont permis de répondre à quelques questions. Mais ils ont également déclenchés de nombreuses autres questions auxquelles nous n'avons pas eu le temps de répondre. Ces questions sont abordées ici sans leur apporter de réponse mais avec l'espoir qu'elles feront l'objet de futurs travaux.

6.2.1 Étude du rythme

Les améliorations les plus immédiates pouvant être faites sur notre approche de rythme concernent certainement la méthode de décision de tempo à partir du spectre de rythme. En effet, les expériences de validation que nous avons effectuées montrent que le tempo annoté est lié aux pics les plus énergétiques dans 97% des cas. Or, notre technique de décision, ne permet encore que 78% de tempi strictement correctement estimés. Si ce résultat est bon, une bonne marge de progression est possible et une décision plus élaborée pourrait sans doute augmenter ce score. Le système de pondération des frontières de la segmentation Forward-Backward peut également être amélioré. En l'état actuel, nous partons de l'hypothèse que les frontières les plus importantes comportent une forte augmentation locale de l'énergie. Dans un contexte musical, cette hypothèse peut être remise en cause et être substituée par d'autres heuristiques basées sur des connaissances musicologiques ou couplées avec d'autres types d'approches d'estimation de rythme. Cette amélioration pourrait peut-être, dans l'analyse de musiques au tempo moins marqué, offrir de meilleures performances. Bien qu'annoncé, l'analyse par tempogramme n'a pas été appliquée pour l'analyse de musiques au tempo variable, il reste à explorer cette analyse afin d'estimer les changements de rythme, les différentes phases musicales et ainsi contribuer à la découverte de la structure du morceau.

À un autre niveau rythmique, les ethnomusicologues du projet sont fortement intéressés par la détection d'*ostinato* ; des formules harmoniques ou rythmiques répétitives. En utilisant d'autres types de frontières (par exemple en repérant tous les débuts d'une même note), nous pourrions mettre en évidence des régularités révélatrices de la présence d'ostinato.

Enfin, l'extraction du tempo, couplée aux positions des frontières de plus forte intensité pourrait assez directement mener à un extracteur de pulsations au sein du morceau. L'idée est que, pour obtenir les pulsations, il faut extraire la série de frontières dont les distances entre voisins sont liées au tempo tout en ayant des poids forts qui indiquent qu'ils correspondent vraisemblablement aux *onsets*. La figure 6.1 illustre une sélection des frontières en suivant ce principe.

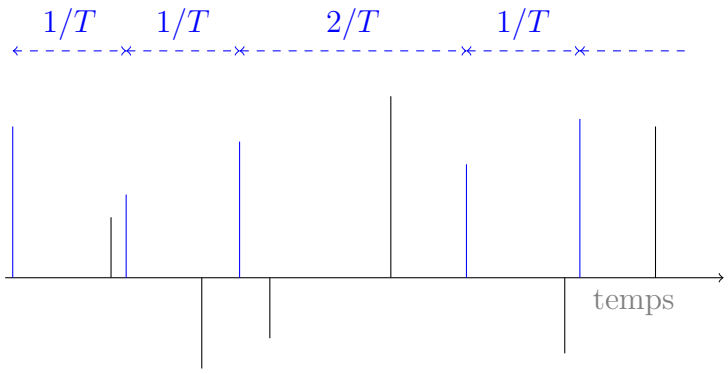


FIGURE 6.1 – Détection de la pulsation à partir du tempo T et de la position des frontières.

Les séries ainsi extraites sur l'ensemble du morceau doivent correspondre à différentes pulsations (temps/contre temps...). Une première étude a déjà été effectuée sur le sujet et semble perceptivement cohérente. Une analyse plus approfondie des résultats reste à effectuer.

Dans le contexte de l'analyse de la parole, les perspectives directes vont vers la conception d'un système de classification des types de parole sur le continuum entre voix parlée et voix chantée. Nous avons montré que la composante rythmique est importante dans cette discrimination, mais d'autres paramètres devront certainement entrer en jeux pour une topologie fine de ces types de voix. Un travail conjoint au sein du projet DIADEMS est en cours avec les ethnomusicologues et ethnolinguistes afin de définir clairement ces catégories. Je veux croire que la conception d'outils automatiques comme les nôtres serve aussi à la conception de définitions basées sur des critères objectifs.

Enfin, si notre méthode a fait ses preuves sur l'analyse de la musique et de la parole, il est intéressant de la confronter également à d'autres types de sons

réguliers. Nous avons ainsi effectué une première analyse pour la recherche de bruits de pas dans des enregistrements urbains. En utilisant un détecteur d'impact au lieu de la segmentation Forward-Backward et en réalisant un *tempogramme* de leurs positions, nous obtenons des zones rythmiquement stables. La figure 6.2 illustre le comportement de cette approche sur un enregistrement contenant des pas. Les zones de fortes intensités du *tempogramme* correspondent aux zones annotées manuellement comme contenant des pas (bandes grises). De plus, les motifs harmoniques des pas réguliers ressortent clairement.

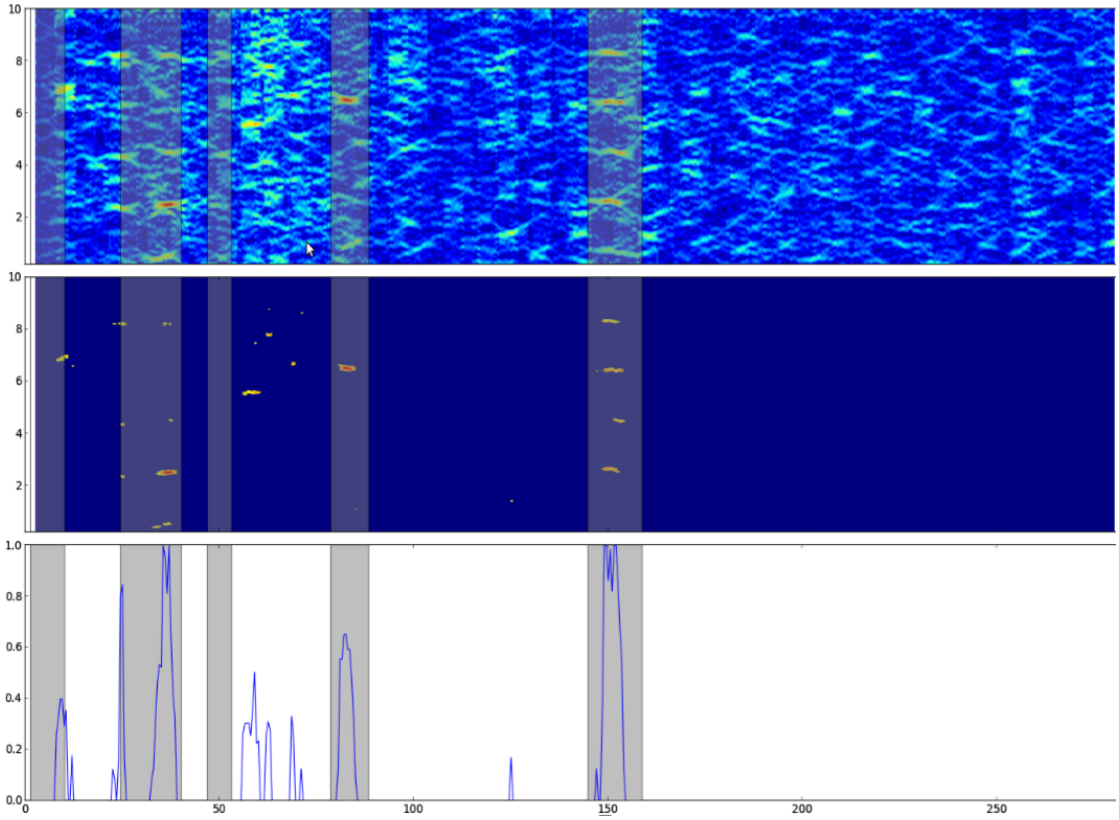


FIGURE 6.2 – Analyse par tempogramme d'un enregistrement de 4 min 30 s contenant des pas. En haut le tempogramme, au centre le tempogramme « seuillé » pour ne faire ressortir que les zones de fortes intensités. En bas, une proposition de courbe de détection basée sur l'énergie cumulée par colonne sur le tempogramme « seuillé ».

Une zone ne comportant pas de « pas » s'active également. Elle contient en réalité des chocs de vaisselles repérés par le détecteur d'impacts. En revanche, ne possédant pas de régularité rythmique, son *tempogramme* est nettement plus bruité. Nous pourrions utiliser cette caractéristique afin de l'éliminer.

6.2.2 Superpositions harmoniques

La méthode que nous proposons pour l'analyse des zones de superpositions harmoniques étant relativement nouvelle, il reste beaucoup à faire pour augmenter son niveau de robustesse et de performance. L'une des premières améliorations est sans aucun doute de revoir la stratégie de choix des pics. Ce dernier se révèle en effet critique puisque l'oubli de sélection d'un pic lié à une source conduit à interrompre le suivi, ce qui peut avoir comme conséquence la non prise en compte de plusieurs portions de segments sinusoïdaux ; ils sont considérés comme trop courts alors que si ils étaient tous reliés, ils seraient parfaitement cohérents et informatifs. Afin de résoudre ce type de problématique, nous pouvons imaginer la mise en place d'un lissage amélioré du résultat des suivis afin d'être capable, à partir de leur dynamique de combler les trous d'une ou deux trames entre les suivis.

L'autre avantage résultant d'une amélioration de la saisie des pics, serait la diminution (voire l'élimination dans le meilleur des cas) des pics liés au bruit. Nous avons vu à quel point des pics liés au bruit peuvent se retrouver liés entre eux et ainsi créer des segments sinusoïdaux sans lien avec les sources ; ils viennent pour autant interférer grandement, en faisant par exemple le lien entre deux familles différentes.

Le système de création des segments sinusoïdaux peut lui aussi être amélioré en vue de notre objectif. Si seules deux trames contiguës sont prises en compte pour relier les différents pics sélectionnés, il pourrait sembler intéressant d'étendre l'analyse pour considérer les positions des pics dans un plus grand voisinage. Ceci renforcerait le suivi en prenant en compte une plus longue dynamique qui serait plus difficilement mise en échec par les pics liés au bruits.

Enfin, un pré-traitement consistant à éliminer le bruit de fond pour ne conserver que le « bruit d'intérêt » doit être envisagé comme une possibilité d'améliorer sérieusement nos performances. Ce problème est néanmoins loin d'être simple et prend probablement trop en compte le niveau sémantique pour être entièrement automatisé.

D'un point de vue applicatif, nous pouvons penser à différentes perspectives.

Alors que nous nous sommes simplement concentrés sur la détection de zones de superposition, nous pourrions caractériser plus précisément la relation entre sources et ainsi identifier par exemple des zones de chant en quinte, tierce ou autre, voire détecter d'autres constantes dans la relation entre les sources dans des contenus ethnomusicologiques.

Les segments sinusoïdaux estimés correctement pourrait être utilisés comme signature d'une source afin de la reconnaître tout au long du morceau.

Différentes informations pourraient être extraites de ces segments pour carac-

tériser les sources. Nous pourrions par exemple nous inspirer des travaux de Salamon [49] afin d'extraire la mélodie de l'accompagnement et ainsi faire un premier pas dans l'extraction de la partie « d'intérêt » du signal.

Nous pensons notamment à la relation entre les amplitudes des différentes harmoniques qui est liée au timbre. Cette information pourrait également, être utilisée pour caractériser la source en fonction de catégories d'instruments.

En parole, les ethnolinguistes ont pointé une analyse automatique qui leur serait précieuse et qui peut découler de notre méthode de recherche de superpositions, la localisation de *tuilages*. Le phénomène de *tuilage* intervient lorsque un locuteur A n'a pas terminé son intervention, qu'un locuteur B intervient et continue ensuite seul (voir figure 6.3).

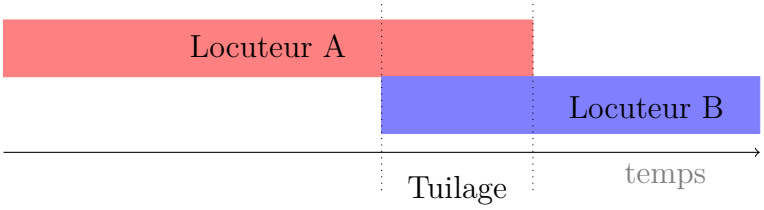


FIGURE 6.3 – Illustration du phénomène de *tuilage* entre le « Locuteur A » et le « Locuteur B ».

Ce phénomène intervient fréquemment dans certains rituels maya et semble structurant, d'après les linguistes alors que leurs apparitions semblent aléatoires pour un néophyte. Notre suivi de fréquence, couplé à une reconnaissance du locuteur dans les zones « Solo », pourrait permettre de localiser de telles configurations afin d'en analyser leur répartition et de mieux comprendre ces rituels.

6.2.3 Conception

Les travaux que nous présentons ont été réalisés dans le cadre du projet DIADEMS. Ce projet continue après cette thèse et a pour objectif la conception d'un site d'accès et d'analyse de données ethnomusicologiques. À cet effet, les différentes méthodes d'analyse automatiques proposées par les partenaires doivent être intégrées à la librairie *TimeSide*. Il me semble important de profiter de cette occasion d'utiliser toutes ces méthodes différentes au sein de cette librairie pour augmenter les interactions entre les différentes méthodes afin de toutes les améliorer. D'un point de vue scientifique comme technique, pouvoir s'appuyer sur des décisions fiables, même de « bas niveau » d'autres méthodes, peut permettre de relâcher des contraintes qui complexifient parfois énormément les méthodes. Il s'agit dès lors de faire coopérer et partager les résultats d'algorithmes issus de différents laboratoires...

Bibliographie

[1] Julien Pinquier, Jean-Luc Rouas et Régine André-Obrecht : Robust speech/music classification in audio documents. *In Seventh International Conference on Spoken Language Processing*, 2002.

[2] Régine André-Obrecht : A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 36(1):29–40, 1988.

[3] Hélène Lachambre, Régine André-Obrecht et Julien Pinquier : Caractérisation de la voix chantée en contexte monophonique et polyphonique. *In XXIIe colloque GRETSI (traitement du signal et des images), Dijon (FRA), 8-11 septembre 2009*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 2009.

[4] Alain De Cheveigné et Hideki Kawahara : Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002.

[5] Barry L. Vercoe : *Timing is of the essence : Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm*. Thèse de doctorat, Massachusetts Institute of Technology, 1993.

[6] Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle et Pedro Cano : An experimental comparison of audio tempo induction algorithms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1832–1844, 2006.

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

[8] Miguel A. Alonso, Gaël Richard et Bertrand David : Tempo and beat estimation of musical signals. *In ISMIR*, 2004.

[9] Anssi Klapuri : Multiple fundamental frequency estimation by summing harmonic amplitudes. *In ISMIR*, pages 216–221, 2006.

[10] Christian Uhle, Jan Rohden, Markus Cremer et Juergen Herre : Low complexity musical meter estimation from polyphonic music. *In Audio Engineering Society Conference : 25th International Conference : Metadata for Audio*. Audio Engineering Society, 2004.

[11] Simon Dixon, Elias Pampalk et Gerhard Widmer : Classification of dance music by periodicity patterns. *In ISMIR*, 2003.

[12] George Tzanetakis et Perry Cook : Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[13] Alexander Grossmann et Jean Morlet : Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis*, 15(4):723–736, 1984.

[14] Geoffroy Peeters : Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, 2007, 2006.

[15] Patrick Flandrin : *Time-frequency/time-scale analysis*, volume 10. Academic Press, 1998.

[16] Andrew J. Viterbi : Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.

[17] Albert Di Cristo : *La prosodie de la parole*. Voix, parole, langage (Solal). De Boeck Supérieur, 2013.

[18] Katarina Bartkova, Denis Jouvet *et al.* : Automatic detection of the prosodic structures of speech utterances. *In SPECOM 2013-International Conference on Speech and Computer*, 2013.

[19] Anna Hjalmarsson Gabriel Skantze, Catharine Oertel : User feedback in human-robot interaction :prosody, gaze and timing. *In INTERSPEECH 2013*, pages 1901–1905, 2013.

[20] Dimitrios Ververidis et Constantine Kotropoulos : Emotional speech recognition : Resources, features, and methods. *Speech communication*, 48(9): 1162–1181, 2006.

[21] Björn Schuller, Stefan Steidl et Anton Batliner : The interspeech 2009 emotion challenge. *In INTERSPEECH*, pages 312–315, 2009.

[22] Franck Ramus et Jacques Mehler : Language identification with suprasegmental cues : A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, 105:512, 1999.

[23] Masayuki Suzuki Ibuki Nakamura, Nobuaki Minematsu *et al.* : Development of a web framework for teaching and learning japanese prosody : Ojad (online japanese accent dictionary). *In INTERSPEECH 2013*, 2013.

[24] Maxime Le Coz, Hélène Lachambre, Lionel Koenig et Régine André-Obrecht : A segmentation-based tempo induction method (regular paper). *In International Society for Music Information Retrieval Conference*, pages 27–31, http ://drops.dagstuhl.de/, 2010.

[25] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura et Ryuichi Oka : Rwc music database : Popular, classical and jazz music databases. *In ISMIR*, volume 2, pages 287–288, 2002.

[26] Masataka Goto : Development of the rwc music database. *In Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, volume 1, pages 553–556, 2004.

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

[28] Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel et Olivier Galibert : The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. *In International Conference on Language Resources, Evaluation and Corpora*, 2012.

[29] Lori F. Lamel, Jean-Luc Gauvain, Maxine Eskenazi *et al.* : Bref, a large vocabulary spoken corpus for french. *training*, 1991.

[30] Jean-Sylvain Lienard, Claude Barras et François Signol : Using sets of combs to control pitch estimation errors. *In Proceedings of Meetings on Acoustics*, volume 4, page 060003, 2008.

[31] François Signol : *Estimation de fréquences fondamentales multiples en vue de la séparation de signaux de parole mélangés dans un même canal.* These, Université Paris Sud - Paris XI, décembre 2009.

[32] Oshry Ben-Harush, Hugo Guterman et Itshak Lapidot : Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization. *In Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, pages 1–6. IEEE, 2009.

[33] Claude Elwood Shannon : A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1947.

[34] Delphine Charlet, Claude Barras et Jean-Sylvain Lienard : Impact of overlapping speech detection on speaker diarization for broadcast news and debates. *In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7707–7711. IEEE, 2013.

[35] Hynek Hermansky : Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738, 1990.

[36] Steven Davis et Paul Mermelstein : Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[37] Ravichander Vipperla, Jurgen Geiger, Simon Bozonnet, Dong Wang, Nicholas Evans, Bjorn Schuller et Gerhard Rigoll : Speech overlap detection and attribution using convolutive non-negative sparse coding. *In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4181–4184. IEEE, 2012.

[38] Daniel D. Lee et Hyunjune Sebastian Seung : Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[39] Sree Harsha Yella et Fabio Valente : Speaker diarization of overlapping speech based on silence distribution in meeting recordings. *In INTERSPEECH*, 2012.

[40] Anssi Klapuri et Manuel Davy : *Signal processing methods for music transcription*. Springer, 2006.

[41] Chunghsin Yeh, Axel Roebel et Xavier Rodet : Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1116–1126, 2010.

[42] Stanislaw Andrzej Raczynski, Nobutaka Ono et Shigeki Sagayama : Multiple frequency estimation for piano recordings with stitched regularized harmonic nmf. *In Proceedings of MIRex 2009*, 2009.

[43] Daichi Sakaue, Takuma Otsuka, Katsutoshi Itoyama et Hiroshi Okuno : Bayesian nonnegative harmonic-temporal factorization and its application to multipitch analysis. *In ISMIR Proceedings*, pages 91–96, 2012.

[44] Kazuyoshi Yoshii et Masataka Goto : Infinite latent harmonic allocation : A nonparametric bayesian approach to multipitch analysis. *In ISMIR*, pages 309–314, 2010.

[45] Judith C. Brown : Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89:425, 1991.

[46] Karin Dressler : Sinusoidal extraction using an efficient implementation of a multi-resolution fft. *In Proc. of 9th Int. Conf. on Digital Audio Effects (DAFx-06)*, pages 247–252, 2006.

[47] Toru Taniguchi, Akishige Adachi, Shigeki Okawa, Masaaki Honda et Katsuhiko Shirai : Discrimination of Speech, Musical Istruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals. *In Interspeech - European Conference on Speech Communication and Technology*. ISCA, septembre 2005.

[48] Hélène Lachambre, Julien Pinquier et Régine André-Obrecht : Distinguishing Monophonies from Polyphonies using Weibull Bivariate Distributions. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6):1837–1842, août 2011.

[49] Justin Salamon, Geoffroy Peeters et Axel Röbel : Statistical characterisation of melodic pitch contours and its application for melody extraction. *In ISMIR*, pages 187–192, 2012.

Résumé

Les travaux de cette thèse portent sur des méthodes permettant de retrouver automatiquement des informations dans des enregistrements sonores. Les données que nous analysons sont fournies par les archives du Musée de l'Homme de Paris : il s'agit de milliers d'heures d'enregistrements musicaux et d'interviews de 1900 à nos jours. Nous proposons deux types d'analyse conçues pour fonctionner aussi bien sur de la musique que sur de la parole. Le premier permet d'extraire le rythme de l'enregistrement à partir de la répartition des zones stables du signal à l'aide d'un « spectre de rythme ». Le second effectue un suivi sur les fréquences les plus présentes et cherche à les regrouper par source pour détecter si plusieurs personnes ou instruments sont présents. Ces analyses peuvent permettre, entre autres, de retrouver la structure d'un chant en fonction du nombre de sources ou savoir si une personne parle, raconte, récite en encore scande en utilisant le rythme présent dans la parole.

Mots-Clef

Musique, Parole, Rythme, Suivis de fréquence, Superpositions harmoniques

Abstract

This thesis aims at designing methods to automatically extract information on sound signals. The sound archives we analyse are provided by the Musée de l'Homme of Paris : they are compounded of thousands of hours of musical recording and interviews from year 1900 to nowadays. We propose two different types of analysis designed to work on music as well as speech. The first system aims at extracting rhythm according to the repartition of stable areas of the signal using a "rhythm spectrum". The second uses a frequency tracking of the most predominant frequencies to group them into source-related clusters to detect if different people or instruments are present. Those techniques may extract different kind of information such as structuring a song using the number of singers or automatically knowing if a record contains someone speaking, reciting or even chanting.

Key Words

Musique, Speech, Rhythm, Frequency tracking, Harmonic overlapping