Rapport de Stage de Master 2 ATIAM

Séparation de la partie percussive d'un morceau de musique

Stage effectué à l'IRCAM
(Institut de Recherche et Coordination Acoustique/Musique)
dans l'équipe Analyse/Synthèse

du 01 mars au 30 juin 2010

Encadré par
Axel Röbel, Geoffroy Peeters et Mathieu Lagrange

François Rigaud

Contexte du stage : l'équipe Analyse/Synthèse de l'IRCAM

Avec la profusion des données numériques disponibles depuis les années 90 une communauté de scientifique s'est attachée à mettre en place des moyens de caractérisation et de classification automatiques. Pour le cas de flux audionumériques, le MIR (Music Information Retrieval) est né avec une première conférence en 2000 : ISMIR (Internationnal Symposium on Music Information Retrieval) et très rapidement une communauté s'est imposée dans ce domaine de recherche autour de la campagne d'évaluation MIREX (Music Information Retrieval Evaluation eXchange). La liste des tâches qui font l'objet de "compétitions" entre chercheurs sur des bases de données cachées fait apparaître des aspects de détection, d'estimation et d'extraction (attaques, accords, battue, hauteurs fondamentales), de classification automatique (identification du genre et de "l'humeur", de l'auteur, étiquetage automatique), des aspects concernant la similarité (requête par chantonnement, identification de variantes, alignement sur la partition, similatité entre séquences symboliques) ou encore d'extraction automatique de mélodie principale, qui s'apparente à de la séparation de sources.

A l'IRCAM, une partie de l'équipe Analyse/Synthèse travaille dans ce domaine et le sujet de stage proposé "Séparation de partie percussive d'un morceau de musique" s'insère entre plusieurs projets :

- Une application intéressante pour l'équipe est d'effectuer un pré-traitement pour un algorithme d'estimation multipitch [1] développé par Chunghsin Yeh [1]. Si l'on considère que les percussions ne participent pas au contenu harmonique d'un spectre alors on effectue en quelque sorte un débruitage en les retirant du signal avant une analyse de hauteurs.

- Pour effectuer cette extraction/soustraction de partie percussive, un algorithme de détection d'onsets [2] développé par Axel Röbel [2] [3] est disponible. Ce traitement peut servir d'étape préalable à l'extraction afin de limiter la recherche des sons de percussions à ces instants précis dans le signal.

En dehors d'un contexte de pré-traitement, l'extraction de piste percussive permettrait un grand nombre d'applications directes comme le suivi de tempo (ou beat tracking), la transcription automatique [3] de partie rythmique, l'identification de genre, le remixage, ...

Pour cette étude nous considèrerons une seule classe de percussions : la batterie. Cet instrument est présent aujourd'hui dans de nombreux genres musicaux populaires et possède un nombre limité d'éléments que nous détaillerons dans la présentation du sujet.

1. Estimation des partiels/composantes fréquentielles d'un signal audio.
2. Instants marquant le début d'une note ou d'un son.
3. Génération d'une partition à partir d'un flux audio

Table des matières

I Présentation du sujet **4**

1 Problématique de séparation-transcription **4**

2 Présentation de la batterie **5**
2.1 Les éléments 6
2.2 Caractéristiques spectro-temporelles 6

3 Bases de données **8**

II Etat de l'art **10**

1 Separation Aveugle de Sources **10**
1.1 Présentation du problème 10
1.2 Analyse en Sous-espaces Indépendants 11
1.3 Factorisation en Matrices Non-Négatives 12
1.3.1 Application à la transcription de batterie (illustration NMF) . 13
1.3.2 Application à l'extraction de piste de batterie 16

2 Approche du type débruitage **17**
2.1 Estimation de l'espace bruit par la méthode Haute Résolution 17
2.2 Diffusion complémentaire de spectrogramme (CDS) 18

3 Méthodes du type *Match and adapt* **21**

III Algorithmes développés **22**

1 Modèle d'enveloppe et de trajectoire percussive **22**
1.1 Des fonctions d'activations temporelles obtenues par NMF vers un modèle de trajectoire moyenne percussive 22
1.2 Modèle paramétrique de l'enveloppe temporelle du son produit par la percussion d'une peau 25
1.3 Modèle paramétrique de la trajectoire temporelle moyenne d'un bin percussif 26

2 Comparaison de trajectoires **29**
2.1 Utilisation de la détection d'onset pour la définition des zones de comparaisons 29
2.2 Mesure de distance entre deux trajectoires 29
2.3 Précision du calage temporel avant la mesure de distance 31

3 Algorithme 1 : Extraction rapide mais limité **31**
3.1 Limitation aux éléments grosse-caisse, caisse-claire, charleston fermé . 31
3.2 Prise en compte des effets dégradant la partie harmonique 31

4 Algorithme 2 : Prise en compte de la mesure du niveau de bruit pour l'extraction des toms et cymbales (développement en cours) 33
4.1 Recherche du alpha optimal pour chaque bin 33
4.2 Utilisation du spectrogramme de niveau de bruit 34
4.3 Deux approches en cours de développement 34

IV Mesures de performances et comparaison avec certains algorithmes de l'état de l'art 37

1 Présentation du corpus 37

2 Mesures de performances 37
2.1 Définition du Rapport Signal sur Résiduel 37
2.2 Evaluation des performances de l'algorithme 1 38
2.3 Comparaison avec d'autres algorithmes de séparation de partie percussive 38

3 Evaluations plus pertinentes de séparation de sources (futur travail) 41

V Conclusion et perspectives 42

A Schéma de principe de l'algorithme 1 43

B Schéma de principe de l'algorithme 2 44

Première partie
Présentation du sujet

1 Problématique de séparation-transcription

Les problèmes de séparation et de transcription de la piste de batterie d'un morceau de musique sont liés et peuvent êtres traités conjointement. Selon les traitements on distingue plusieurs types d'approches :

• Dans un contexte de **séparation** (cas du stage), une approche "naturelle" consiste à extraire directement la batterie du morceau en utilisant des algorithmes de séparation de sources avec ou sans *a priori* sur les sources à estimer.

Pour cela, les méthodes de Séparation Aveugles de Sources comme ISA ou NMF présentées partie II.1. permettent de décomposer le signal en sources indépendantes et il est possible de retrouver la batterie dans une ou plusieurs de ces sources. Ce type de méthode possède en général le défaut de nécessiter la connaissance du nombre de pistes à extraire.

Une autre idée est de décomposer le signal en parties harmonique + percussion. On suppose via cette décomposition que les deux parties ne se recouvrent pas spectralement (ce n'est pas le cas en pratique, mais on peut faire l'hypothèse pour un instant et une bande de fréquence donnés dans le spectrogramme, qu'une des deux parties est prépondérante). Cette approche de type débruitage est détaillée partie II.2..

• Dans le cadre d'une **transcription** deux classes de méthodes sont recensées par Gillet [4] :

Le *segment and classify*, consiste à segmenter le morceau en détectant les onsets, en extraire des informations (features) et classer les données à l'aide des techniques d'apprentissage automatique. Ce type de méthode est reconnu pour donner de bons résultats dans le cas ou le signal comporte uniquement de la batterie mais semble poser plus de problème dans le cas de signaux polyphoniques.

Le *separate and detect*, nécessite une première étape de séparation de partie percussive (cf. premier point), suivit d'une étape de détection des évènements de chaque élément de la batterie. Une version alternative de la NMF, la Décomposition en Matrice Non-négatives peut être utilisée pour le second traitement (présenté dans l'illustration de la NMF partie II.1.3.1.).

• Enfin, pour traiter conjointement les problèmes de **séparation** et de **transcription**, les méthodes du type *match and adapt* sont proposées. Celles-ci consistent à partir de templates de référence (modèles temporels ou spectraux) pour chaque élément de la batterie à rechercher dans le signal audio des zones de forte ressemblance et itérativement affiner le template jusqu'à reconstruire le timbre correspondant. Ces méthodes sont présentées partie II.3. .

Dans ce document, nous présentons une étude bibliographique détaillée de certains algorithmes de l'état de l'art (NMF+SVM partie II.1 et Diffusion Complémentaire de Spectrogramme partie II.2.2). Ces méthodes implémentées sous MATLAB^{TM} serviront de références pour la comparaison des performances du nouvel algorithme mis en oeuvre lors du stage. Nous présentons plus succinctement (algorithmes plus complexes à implémenter) la séparation de partie percussive par la méthode Haute Résolution partie II.2.1 et les méthodes de type *match and adapt* partie II.3.

2 Présentation de la batterie

La batterie est un instrument apparu au début du XX^{eme} siècle avec la naissance du Jazz. Ses éléments constitutifs ont été rassemblés à partir de diverses percussions déjà existantes de part le monde. La caisse claire et la grosse caisse étaient utilisées par les fanfares militaires dès le XVII^{eme} en Europe. Les toms sont inspirés des percussions africaines et amérindiennes. Les cymbales quand à elles proviennent d'orient et sont probablement parmis les instruments les plus anciens du monde.

La batterie s'est rapidement imposée comme instrument de référence pour marquer la rythmique d'un morceau et est aujourd'hui présente dans toutes les musiques populaires (sous sa forme acoustique ou synthétisée). Dans le milieu de la musique savante, Edgard Varèse a largement participé à son introduction avec une des premières pièces composée uniquement pour percussions en occident, *Ionisation* (1929-1931) jouée par 13 percussionnistes.

FIGURE 1 – Set de batterie.

2.1 Les éléments

Sur la *figure 1* nous présentons un "set" classique de batterie.

La grosse caisse, les toms et la caisse claire sont formés de fûts recouverts de peaux sur les faces avant et arrière. La grosse caisse, de diamètre plus élevé, est l'élément produisant le son le plus grave, elle se joue au pied avec une pédale et marque souvent les temps forts. La caisse claire comporte sur la face inférieure un timbre (composé de filaments de fer tendus) qui donne la résonance à l'instrument. Les toms ont un diamètre inférieur à celui de la grosse caisse et sont frappés avec les baguettes. Ils sont généralement accordés en réglant la tension des peaux (problème avec l'hypothèse non harmonique des percussions).

Les cymbales sont des disques ou des cloches de métal. Les plus courantes sont la *crash*, utilisée pour marquer les temps forts ou les variations d'un morceau, la *ride* produisant trois sons distincts selon la zone de frappe (le corps pour un son léger et clair, la cloche pour un son claquant et précis, et le bord qui possède un son plus gras et lourd) et le *charleston* qui sert généralement à marquer le tempo, composé de deux cymbales dont l'espacement est réglé par l'appuie sur une pédale (en charleston fermé les résonances sont très brèves).

D'une manière générale, les fûts produisent des sons impulsifs rapidement amortis, alors que les cymbales frappées entrainent de longues résonances chaotiques [5] (modélisations physiques non-linéaires). Nous pouvons déjà prévoir qu' une hypothèse du type "impulsion brève d'énergie" sera limitée au cas des fûts.

2.2 Caractéristiques spectro-temporelles

Nous visualisons *figure 3* le spectrogramme des sons produits par différents éléments d'une batterie. De gauche à droite, coups de : grosse caisse, tom basse, caisse claire, charleston fermé et cymbale crash. *Figure 2* on détaille les spectres d'amplitude calculés sur la durée de chaque son.

• Hypothèse de non-harmonicité des percussions : validité ?

On vérifie que les sons de grosse caisse et de tom sont spectralement très localisés en basse fréquence. Le tom possède une résonance plus importante que la grosse caisse du fait de son spectre très étroit, ceci nous donne la sensation de hauteur. La caisse claire présente une série de pic harmoniques correspondant à la résonance du timbre (multiples de la 100ène de Hertz). Enfin, les sons de charleston et de cymbale crash sont très étalés spectralement. On visualise de nombreux partiels sur le spectrogramme de la cymbale [4].

Ces deux figures nous montrent que l'hypothèse de non-harmonicité du spectre des sons de percussions n'est pas correcte pour tous les éléments de la batterie. En effet, les membranes et les cymbales possèdent des modes de résonance qui se retrouvent dans les sons. Il est possible qu'une telle hypothèse entraine une limitation des performances pour l'extraction de la caisse claire, des toms ou des cymbales.

4. D'après la théorie, comme les cloches, les cymbales produisent un spectre inharmonique. Ici ceci est difficile à vérifier à l'oeil nu.

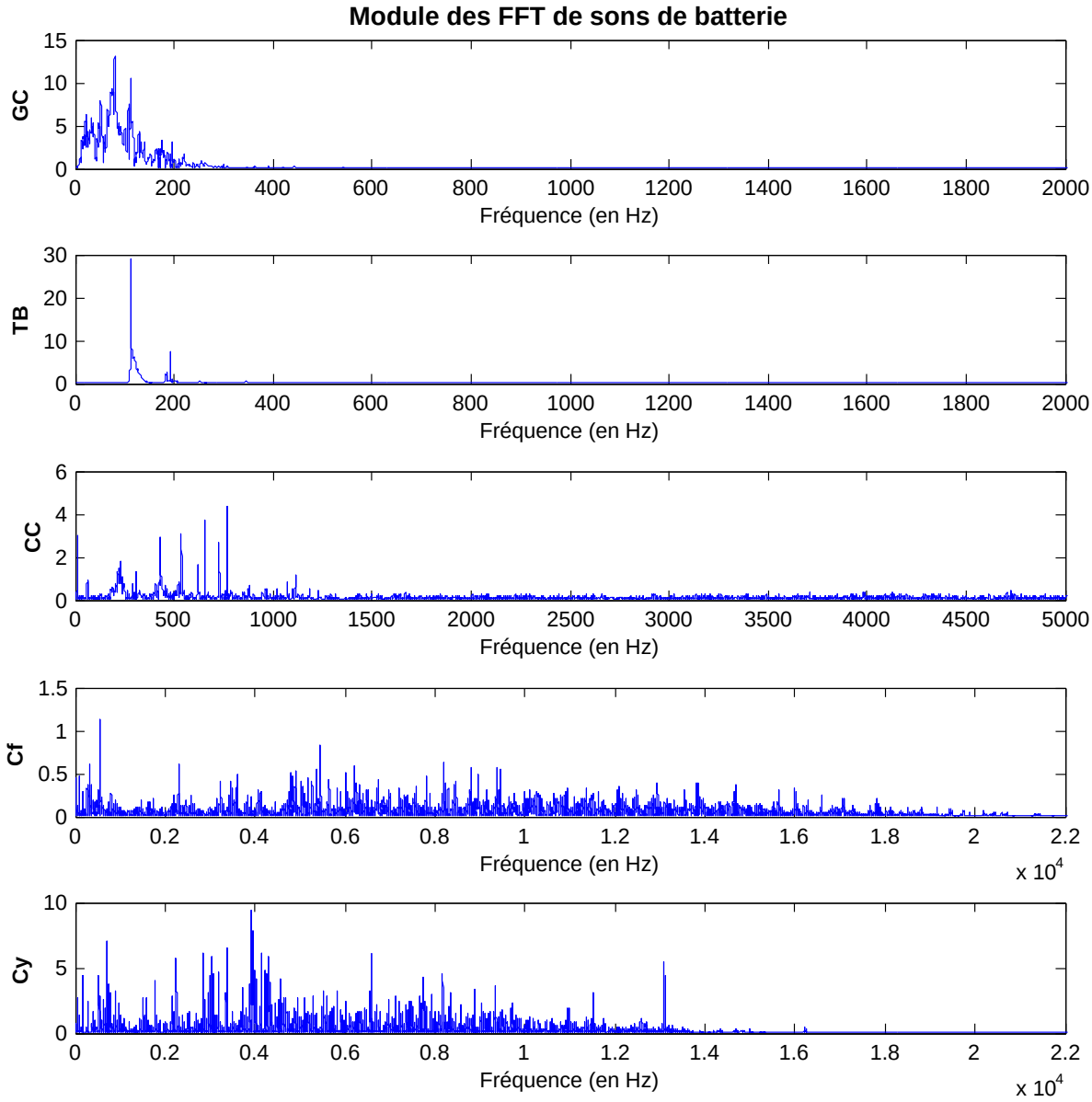


FIGURE 2 – Spectres d'amplitude (calculés avec fenêtrage de hanning sur la durée de chaque son) d'un coup de grosse caisse, tom basse, caisse claire, charleston fermé et cymbale crash.

• Hypothèse d'impulsion énergétique : validité ?

Certains algorithmes d'extraction de batterie [6] [7] [8] font l'hypothèse que les sons percussifs possèdent une attaque très brève, quasi instantanée et une décroissance très rapide (inférieure à $200ms$). On mesure sur la *figure 3* les temps de décroissance des enveloppes à 95% de la valeur maximale : environ $100ms$ pour la grosse caisse, $410ms$ pour le tom, $175ms$ pour la caisse claire, $260ms$ pour le charleston ouvert et $1,8s$ pour la crash. L'ordre de grandeur reste correct pour tous les éléments sauf la cymbale. Une hypothèse de ce genre risque de poser problème pour son extraction.

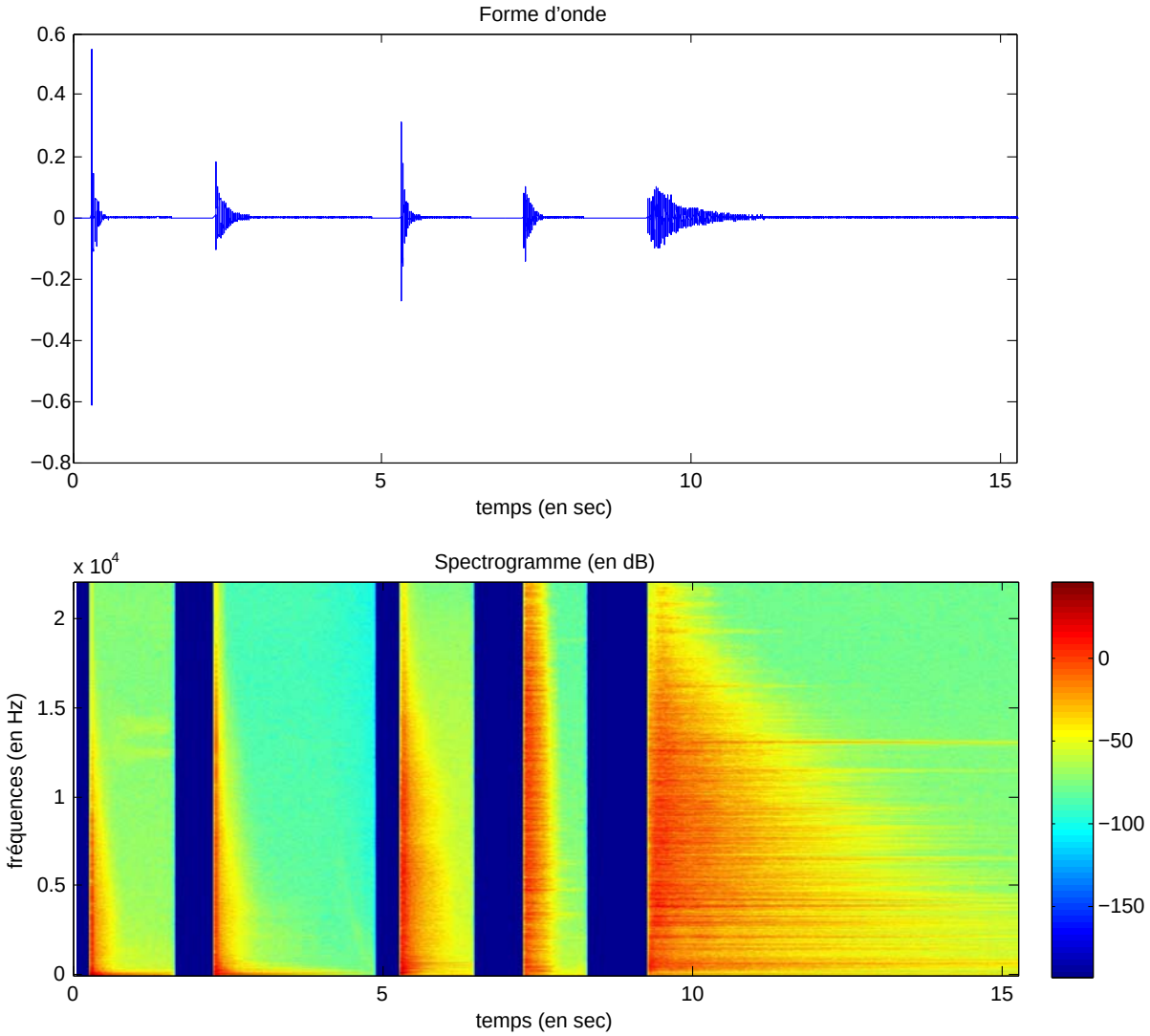


FIGURE 3 – Formes d'onde et spectrogramme d'un coup de grosse caisse, tom basse, caisse claire, charleston fermé et cymbale crash. Spectrogramme calculé avec fenêtre de hanning de $50ms$ et recouvrement de 75%

3 Bases de données

Pour comparer les résultats d'extraction-transcription de batterie deux bases de données sont classiquement utilisées : *ENST Drums* [9] et *RWC Music Database* [10].

Sur la *figure 4*, Paulus [11] présente le pourcentage d'apparition de chaque élément de la batterie dans les deux bases de données. On vérifie que les trois éléments les plus présents en moyenne sont, le charleston, qui marque généralement chaque temps, et la grosse caisse/caisse claire qui marquent souvent un temps sur deux. Il est à prévoir selon les hypothèses effectuées sur les modèles de percussions que tous les éléments de la batterie ne pourront pas être correctement extraits. Nous privilégierons dans un premier temps pour ce travail l'extraction des éléments à courtes résonances, soit la caisse claire, la grosse caisse et le charleston.

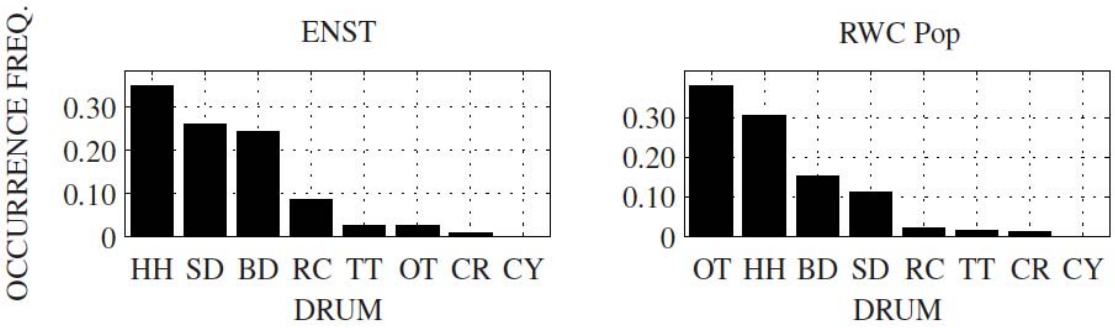


FIGURE 4 – Pourcentage d'apparition des différents éléments de batterie dans les bases de données ENST Drums et RWC Music Database : BD (Bass Drum), CR (all Crash cymbals), CY (other Cymbals), HH (open and closed Hi-Hat), RC (all Ride cymbals), SD (Snare Drum), TT (all Tom-Tom) et OT (Other instruments : cow bell, triangle, tambourine). Extrait de [11]

Pour l'élaboration de notre algorithme nous utiliserons des enregistrements multi-pistes, c'est à dire des enregistrements dont on possède les pistes séparées de chaque instrument. On pourra ainsi regarder l'effet de l'extraction sur chaque piste (on extraira inévitablement une partie harmonique avec les percussions et inversement on oubliera des sons percussifs dans la partie harmonique) et calculer les résiduels harmoniques et percussifs.

Deuxième partie

Etat de l'art

1 Separation Aveugle de Sources

1.1 Présentation du problème

La séparation aveugle de sources est une méthode de traitement du signal qui consiste à extraire les sources composant un signal avec peu ou pas d'informations *a priori*. Pour ce faire, on dispose en pratique de P enregistrements de N sources mélangées.

Une méthode permettant la résolution de ce problème est l' Analyse en Composantes Indépendantes (ACI ou ICA en anglais). L'hypothèse sous-jacente est que les sources sont des variables aléatoires mutuellement indépendantes statistiquement. [5]

Les mélanges de sources peuvent être linéaires ou non-linéaires. Parmi les mélanges linéaire, on trouve les mélanges linéaires instantanés, les mélanges atténués et décalés dans le temps et les mélanges convolutifs. Nous nous intéresserons ici uniquement au premier cas, c'est à dire aux signaux dont les observations s'écrivent comme des combinaisons linéaires des sources. Le problème peut alors se traduire sous la forme matricielle suivante,

$$x(t) = A \cdot s(t)$$

avec $x(t)$ le vecteur des observations (au temps t) de taille $P \times 1$, $s(t)$ le vecteur des sources de taille $N \times 1$ et A la matrice de mélange de dimension $N \times P$ possédant des coefficients constants.

Trois cas sont possibles selon le nombre d'observations P et de sources N :

- $P = N$: On parle alors de mélange déterminé. Pour résoudre le problème il suffit (plusieurs algorithmes existent) d'estimer la matrice de mélange A carrée, et de l'inverser.

- $N > P$: Le mélange est sous-déterminé. On peut choisir d'extraire uniquement P sources qui seront un (autre) mélange des sources initiales ou alors poser des hypothèses (de parcimonie par exemple) afin de tenter d'extraire toutes les sources.

- $P > N$: Le mélange est dit sur-déterminé, on peut se ramener à un mélange déterminé en utilisant uniquement N observations ou effectuer une Analyse en Composantes Principales (ACP).

5. En plus de l'hypothèse de mutuelle indépendances des sources, l'ACI stipule qu'au plus une des sources peut suivre une distribution gaussienne. En effet, pour des variables aléatoires gaussiennes l'indépendance se réduit à la décorrélation, soit à l'annulation de la corrélation. Cependant la corrélation étant symétrique, $Corr(X_i, X_j) = Corr(X_j, X_i)$, le nombre de contraintes est inférieur au nombre de degrés de liberté du système si l'on considère au moins deux variables aléatoires gaussiennes. Il possède alors une infinité de solutions.

1.2 Analyse en Sous-espaces Indépendants

Dans le cas de la séparation de sources à partir d'un signal monophonique (un seul flux) on ne dispose que d'une seule observation, soit $P = 1$. Il n'est pas possible de résoudre le problème par la ACI. Une méthode alternative est l' Analyse en Sous-espaces Indépendants (ASI ou ISA en anglais) appliquée au spectrogramme (amplitude de la Transformé de Fourier à Court Terme $X[r\ k]$, avec r les indices de trames et k les bins fréquentiels). L'idée est de séparer le spectrogramme en plusieurs spectrogrammes indépendants correspondants à chaque source. La décomposition s'exprime sous la forme suivante,

$$X = \sum_{i=1}^{N} X_i$$

Pour effectuer cette décomposition on suppose que le spectrogramme de chaque source se factorise par le produit d'un vecteur colonne f contenant un spectre stationnaire et d'un vecteur temporel ligne t^T correspondant à la modulation en amplitude du spectre dans le temps. Soit :

$$X_i = f_i \quad t_i^T$$

Cette méthode est appliquée à la séparation de partie percussive dans les articles [12] [6] [13].

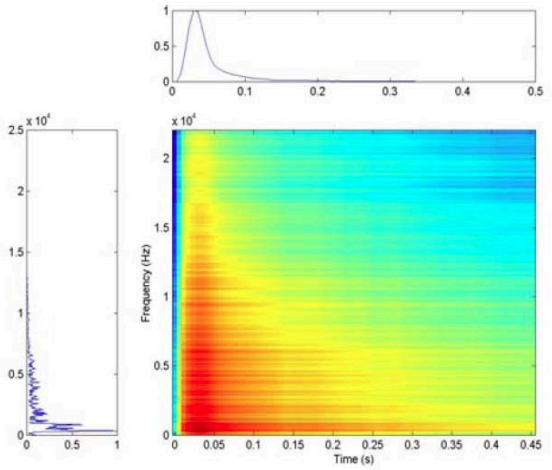


FIGURE 5 – Décomposition d'un spectrogramme de son de caisse claire, extrait de [12]

Remarque 1 : Il est à noter que cette méthode est limitée à des sources stationnaires, ce qui n'est jamais exactement le cas en principe.

Remarque 2 : L'indépendance des variables aléatoires de ACI porte ici sur les vecteurs f_i pour la décomposition.

Remarque 3 : Ce type de décomposition n'exclut pas le fait d'obtenir des spectres d'amplitude f ou des fonctions temporelles de modulation t négatifs, ce qui pose problème pour la resynthèse (et qui n'a pas de sens physique).

Remarque 4 : En théorie, les "spectrogrammes complexes" des différentes sources sont additifs mais pas les "spectrogrammes d'amplitude" (linéarité de la Transformée de Fourier, mais pas de l'opérateur module). Pour la resynthèse, une méthode simpliste mais rapide consiste à garder la matrice des phases lors de l'étape de la Transformation de Fourrier à Court Terme et à l'appliquer aux spectrogrammes des signaux à reconstruire.

1.3 Factorisation en Matrices Non-Négatives

Une manière de prendre en compte la non-négativité des fonctions de base f_i et t_i est de factoriser le spectrogramme initial en ajoutant une contrainte. Renommons (simple convention liée à la NMF) V le spectrogramme, W le "dictionnaire" formé de la concaténation des vecteurs f_i et H la matrice d'activation formé de la concaténation des vecteurs t_i^T. La décomposition proposée par ISA peut alors se mettre sous la forme :

$$V = W \cdot H$$

avec V de taille $n \times m$ (n bins fréquentiels et m trames temporelles), W de taille $n \times r$ (r le nombre de sources à extraire) et H de taille $r \times m$.

La Factorisation en Matrices Non-négatives (NMF en anglais) consiste à estimer les matrices W et H composées uniquement de coefficients positifs approximant au mieux V. Soit,

$$\widehat{V} \approx W \cdot H$$

en minimisant la distance $D(V|W \cdot H)$.
Plusieurs distances sont proposées, la plus classique étant le carré de la distance Euclidienne :

$$\|V - W \cdot H\|^2 = \sum_{ij} (V_{ij} - [W \cdot H]_{ij})^2$$

Dans le cas de l'extraction/transcription de percussion, l'introduction de la divergence [6] de Kullback-Leibler a montré de bons résultats [14]. Celle-ci se définie de la manière suivante :

$$D(V|W \cdot H) = \sum_{ij} V_{ij} \cdot \log\left(\frac{V_{ij}}{[W \cdot H]_{ij}}\right) - V_{ij} + [W \cdot H]_{ij}$$

L'algorithme :

L'algorithme NMF propose d'initialiser aléatoirement les matrices W et H avec des coefficients positifs et de converger vers un minimum en affinant tour à tour W et H de manière itérative (en gardant la contrainte de positivité).
Le problème de minimisation peut s'exprimer sous la forme suivante :

$$W \leftarrow \underset{W \in \mathbb{R}_+^{n \times r}}{\operatorname{argmin}} D(V|W \cdot H) \qquad H \leftarrow \underset{H \in \mathbb{R}_+^{r \times m}}{\operatorname{argmin}} D(V|W \cdot H)$$

L'arrêt de l'algorithme s'effectue en fixant un critère sur l'évolution de la fonction coût (variation relative inférieure au 1/1000 par exemple pour notre programme).

6. le terme de distance n'est plus applicable car cette fonction n'est pas symétrique en V et $\widehat{V} \approx W \cdot H$

Pour la minimisation, plusieurs techniques d'optimisation sont disponibles. Une alternative à la "descente de gradient" est proposé par Lee et Seung [15] : les "règles de mises à jour multiplicative" (multiplicative update rules). Cette méthode contient implicitement le fait que les matrices doivent rester positives (pas d'additions/soustractions). De plus elle présente un bon compromis entre temps de calcul et complexité de programmation. Pour chaque itération on effectue les calculs suivants (cas de la divergence KL) :

$$W \leftarrow W. \times \frac{(V./WH)H^T}{1_{size(V)}S^T} \qquad H \leftarrow H. \times \frac{W^T(V./WH)}{W^T 1_{size(V)}}$$

$.\times$ et $./$ représentent la multiplication et la division terme à terme de deux matrices de tailles identiques.

1.3.1 Application à la transcription de batterie (illustration NMF)

Le but de la transcription est de déterminer les instants d'occurence de chaque élément de la batterie. Pour cela nous devons dans un premier temps apprendre le dictionnaire W (i.e. les spectres stationnaires de chaque élément) sur un morceau de référence ou seule la batterie est présente et où les éléments sont joués séparément. La méthode est détaillée dans [16]. La *figure 6* illustre le résultat obtenu pour un signal d'apprentissage contenant un coup de grosse caisse, de caisse claire et de charleston fermé.

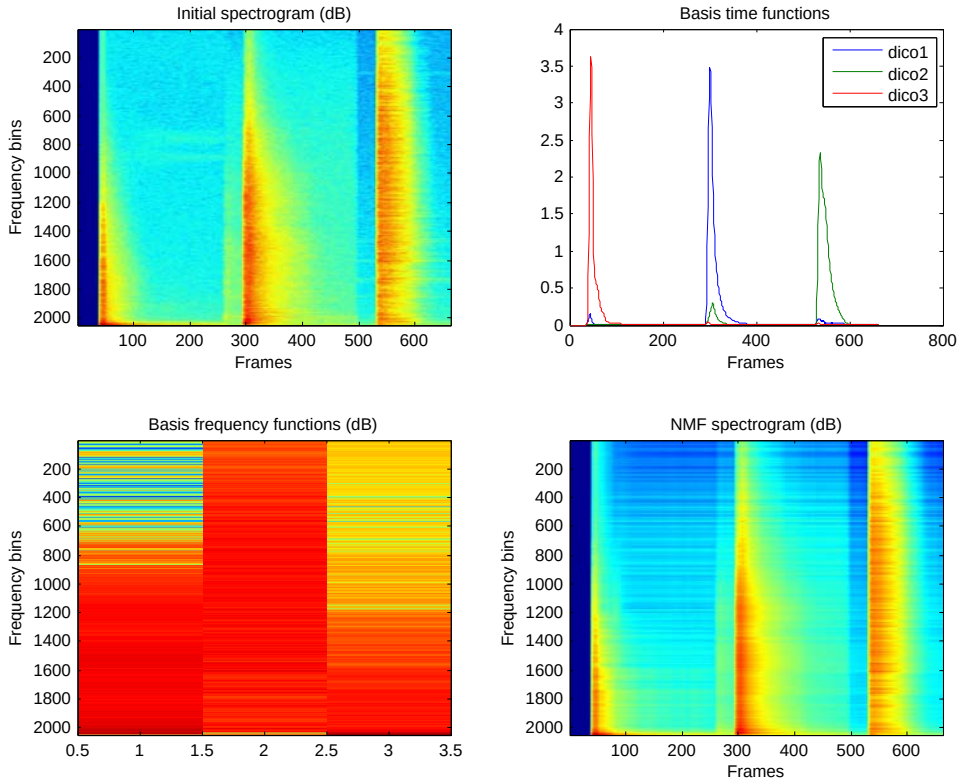


FIGURE 6 – En haut à gauche le spectrogramme initial Vcalculé par fenêtrage de hanning de $50ms$ et pas d'avance de $1/16$ (valeurs choisies pour l'illustration). En bas à droite le sprectrogramme NMF $\widehat{V}$. En haut à droite les fonctions temporelles d'activation. En bas a gauche les spectres stationnaires appris.

On remarque que les différents éléments ne sont pas appris dans l'ordre d'apparition sur le signal audio de départ (la grosse caisse correspond au $3^{\text{ème}}$ spectre). Dans un but de transcription une étape de classification automatique sera ensuite appliquée. Autre remarque, la NMF n'est pas unique, les fonctions de bases sont définies à une constante multiplicative près.

Une fois le dictionnaire appris, on relance l'algorithme sur un signal musical en fixant W et donc en mettant à jour uniquement H. Cette opération porte le nom de Décomposition Non-Négative. Pour illustrer l'algorithme nous "programmons" en MIDI une boucle de batterie où les 3 éléments appris apparaissent (même kit de batterie, ce qui n'est pas le cas en pratique) et nous superposons un piano jouant sur les mêmes temps que la batterie (cf. *figure 7*). On visualise le résultat de la NMF sur les *figures 8* et *9* .

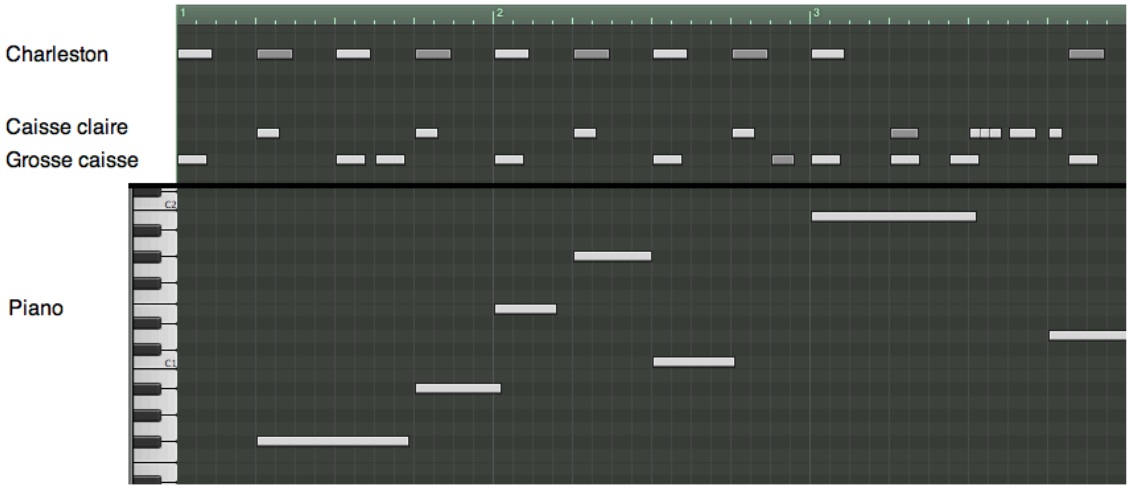


FIGURE 7 – Partition MIDI de la boucle à transcrire.

On remarque que les attaques de piano apparaissent dans les pistes de grosse caisse et de caisse claire. Il serait possible d'effectuer un post-traitement portant sur le taux de décroissance des "releases" pour les supprimer. Plus problématique, la caisse claire et le charleston ont tendance à se recouvrir. Ceci poserait surement problème pour la transcription de différentes cymbales. Des hypothèses de parcimonie peuvent être employées pour améliorer la séparation entre pistes , celles-ci sont ajoutées via l'introduction de contraintes dans la fonction coût, voir [17].

Une autre amélioration possible réside dans la prise en compte de la non stationnarité des signaux réels. Au lieu de travailler avec des spectres de base on peut généraliser la NMF à des spectrogrammes de base. Ceci est proposé sous le terme de "Non-negative Matrix Factor 2-D Deconvolution" dans [18].

La resynthèse est possible mais le son correspondra à la batterie qui a servie à l'apprentissage et non pas à la batterie du morceau original. Ceci ne présente pas d'intérêt pour l'application prévue dans le stage mais illustre bien l'algorithme de NMF.

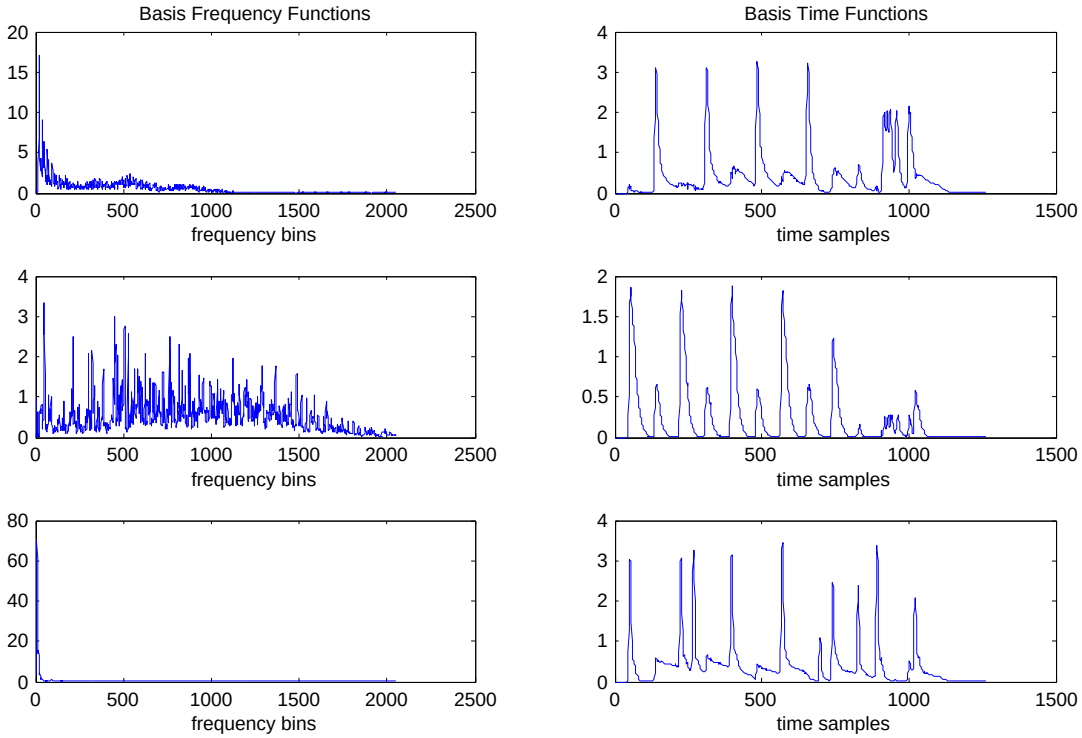


FIGURE 8 – Fonctions de bases extraites des matrices W et H. De haut en bas : caisse claire, charleston fermé et grosse caisse

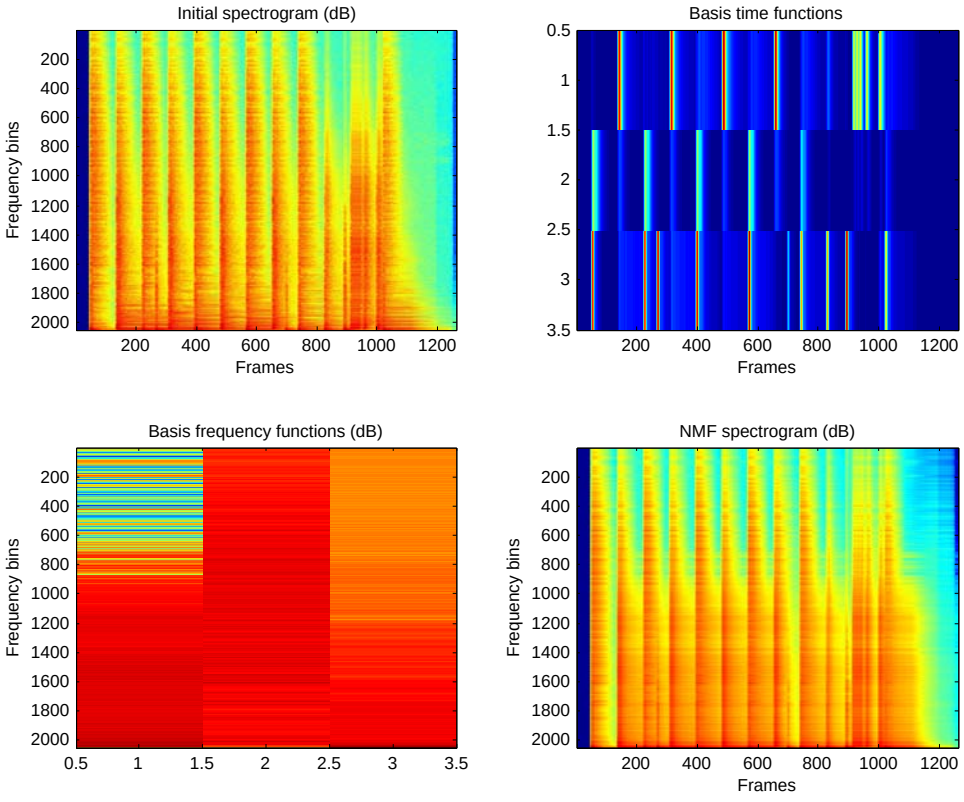


FIGURE 9 – NMF de la boucle musicale (batterie + piano) proposée.

1.3.2 Application à l'extraction de piste de batterie

L'utilisation première de la NMF correspond à une Séparation Aveugle de Sources, il est possible de l'appliquer à l'extraction de piste de batterie. Dans l'article [14] Hélen et Virtanen proposent de décomposer le signal audio en 20 composantes (choix arbitraire). Certaines contiennent des fragments de sons d'éléments de batterie et d'autres des éléments de la partie harmonique. Pour récupérer les 2 parties distinctes une étape de classification automatique avec apprentissage est indispensable. La méthode généralement proposée pour ce type d'application est la SVM (Machines à Support de Vecteur ou Support Vector Machine en anglais). Le classement des données s'effectue sur plusieurs *features* comme les coefficient cepstraux, le centroïde spectral, ...

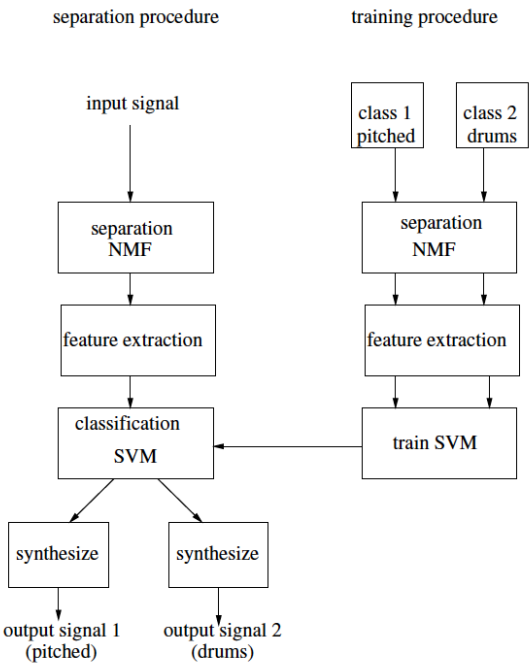


FIGURE 10 – Algorithme de séparation de batterie par NMF + SVM, extrait de [14]

Pour la comparaison finale des différents algorithmes implémentés lors du stage, nous effectuerons la classification des sources "à l'oreille" (étape d'apprentissage complexe à mettre en oeuvre dans la durée du stage).

2 Approche du type débruitage

2.1 Estimation de l'espace bruit par la méthode Haute Résolution

A l'origine, les méthodes Haute Résolution ont été développées pour faire face aux limitations de l'analyse de Fourier lors de l'estimation de sinusoïdes.

Considérons un modèle de signal constitué de sinusoïdes modulées exponentiellement (EDS pour Exponentially Damped Sinusoids) et d'un bruit additif. Pour chaque trame d'analyse le signal réel $s[n]$ peut être explicité sous la forme suivante :

$$s[n] = \sum_{k=0}^{K-1} a_k \cdot e^{\delta_k n} \cdot cos(2\pi f_k n + \phi_k) + b[n]$$

avec pour chaque partiel k, $a_k > 0$ l'amplitude , $\delta_k \in \mathbb{R}$ le coefficient de modulation exponentielle, $f_k \in]-\frac{1}{2}, \frac{1}{2}]$ la fréquence instantanée et $\phi_k \in]-\pi, \pi]$ la phase à l'origine.

Lors d'une analyse par FFT, les pics correspondant à des sinusoïdes pures dans le spectre ont une résolution spectrale limitée par le choix du type de la fenêtre d'analyse et de sa longueur. De plus, la précision spectrale dépend du nombre de points de calcul de la FFT. A ces effets s'ajoute l'élargissement du pic lors de la modulation temporelle de la sinusoïde.

L'idée des méthodes Hautes Résolution est de trouver deux espaces disjoints, l'un contenant la partie harmonique et l'autre la partie bruit, afin de permettre l'extraction des paramètres simplement. On parle alors d'espace signal et d'espace bruit. Plusieurs applications sont possibles [19] [20] : l'analyse haute résolution de signaux musicaux (tracé du HR-ogram [7]) et leur modification/resynthèse.

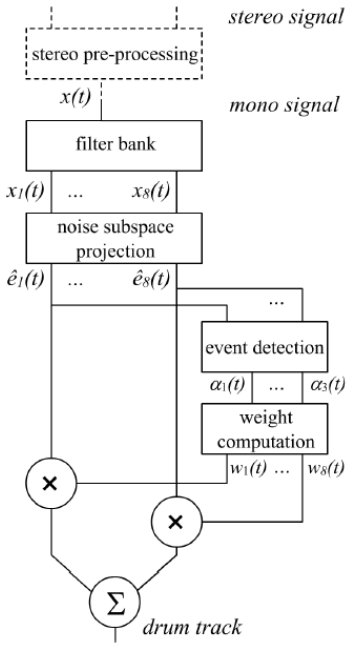
Cette méthode peut être appliquée à la séparation de piste percussive en effectuant l'hypothèse que celle-ci est composée uniquement de bruit. L'estimation des paramètres des sinusoïdes est alors laissée de côté pour ne s'occuper que de la reconstruction de $b[n]$ à partir de l'espace bruit.
Il est a noter cependant que l'algorithme nécessite l'estimation du nombre de sinusoïdes pour définir la taille de l'espace signal. Une méthode est *a priori* disponible pour effectuer automatiquement cette estimation mais l'algorithme actuel pose arbitrairement le nombre de sinusoïdes à estimer par sous bande.

Nous résumons les grandes lignes de l'algorithme proposé par Gillet [21] :

- Un pré-traitement est proposé pour les signaux stéréo afin de pondérer le mixage des deux voies de manière à faire ressortir les éléments impulsifs du morceau de musique.

- Une décomposition en 8 sous bandes est effectuée à l'aide d'un banc de filtre dyadique (ou "d'octaves") qui découpe l'axe fréquentiel en octaves en partant de la bande $[\frac{Fs}{4}, \frac{Fs}{2}]$.

7. Spectrogramme haute résolution de la partie harmonique tracé à partir des sinusoïdes estimées



– La projection du signal audio sur l'espace bruit est ensuite réalisée. Pour le détail du calcul théorique et algorithmique des espaces signal et bruit le lecteur pourra se reporter à la théorie des méthodes Hautes Résolutions (entre autres [19] et [20]). Le signal résultant contient la batterie mais aussi toutes les attaques des notes des autres instruments. Un post-traitement est alors mis en oeuvre.

– Le post-traitement consiste à effectuer un masquage temporel sur les signaux "bruits" reconstruits dans chaque sous bandes aux instants où la batterie n'est pas présente. Pour cela, un simple seuil est appliqué sur l'amplitude de l' enveloppe temporelle des signaux reconstruits. 3 seuils sont fixés et associés à trois classes d'éléments de la batterie : un pour la grosse caisse dans les 2 bandes basses fréquences, un pour la caisse claire dans les 2 bandes moyennes fréquences et un pour les cymbales dans les 4 bandes hautes fréquences.

– Finalement le signal de batterie est reconstruit en additionnant les contributions dans chaque bande.

Le seuil sur l'amplitude n'étant pas très robuste, un post-traitement amélioré est proposé dans [4] . Celui-ci propose de récupérer l'enveloppe des signaux percussifs temporels dans chaque sous bandes et dans un premier temps faire correspondre aux décroissances des courbes exponentielles décroissantes, permettant ainsi de définir des enveloppes moyennes de références pour chacun des différents éléments de batterie (grosse caisse, caisse claire et charleston dans l'article). Dans un second temps, ces enveloppes moyennes sont comparées aux signaux temporels et un seuil est appliqué sur la fonction de comparaison. Une fois l'étape de comparaison terminée, l'enveloppe de chaque sous bande est reconstruite à partir des enveloppes de références (on réalise ainsi une transcription) et est appliquée aux signaux correspondants. Le masque temporel de post-traitement réalisé n'est alors plus binaire.

2.2 Diffusion complémentaire de spectrogramme (CDS)

L'article [8] propose une méthode rapide et simple à programmer pour séparer la partie percussive d'un morceau de musique. Celle-ci se base sur l'hypothèse que la partie percussive et la partie harmonique occupent des domaines complémentaires sur le spectrogramme. La partie harmonique est généralement caractérisée par des raies horizontales dont l'amplitude varie lentement par rapport aux sons de percussions. Ces dernières, au contraire, sont très localisées en temps, elles concentrent l'énergie dans des fines bandes verticales.

L'idée de l'algorithme est de favoriser l'apparition d'une direction sur le spectrogramme, selon si l'on veut récupérer les sons percussifs où harmoniques. Pour cela, on modifie le spectrogramme de manière à minimiser la norme du gradient, celui-ci caractérisant les variations d'amplitude selon les axes temporels et fréquentiels. En ajoutant un poids sur une des composantes du gradient on a la possibilité de favoriser une direction.

Ono & al. proposent d'effectuer le traitement sur le spectrogramme de puissance "compressé" :

$$W[r,k] = |X[r,k]|^{2\gamma} \quad \text{avec} \quad (0 < \gamma \leq 1)$$

Celui-ci se décompose en une partie harmonique et une partie percussive réparties de façon complémentaire dans le spectrogramme :

$$W[r,k] = H[r,k] + P[r,k]$$

avec comme contraintes $H[r,k] \geq 0$ et $P[r,k] \geq 0$.
En s'intéressant simplement aux variations de chaque partie selon son axe privilégié (r correspond à l'indice de trame et k aux bins fréquentiels), le module du gradient discrétisé peut s'exprimer par la somme suivante :

$$\sum_{r,k} \big(H[r-1,k] - H[r,k])^2 + (P[r,k-1] - P[r,k]\big)^2 .$$

Au final la fonction coût à minimiser est quadratique pour toutes les variables, elle possède un unique minimum.

$$J(H,P) = \frac{1}{2\sigma_H^2} \cdot \sum_{r,k} (H[r-1,k] - H[r,k])^2 + \frac{1}{2\sigma_P^2} \cdot \sum_{r,k} (P[r,k-1] - P[r,k])^2$$

Les coefficients σ_H et σ_P permettent de favoriser la diffusion du spectrogramme selon une direction. En pratique, résoudre les équations $\frac{dJ}{dH[r,k]} = 0$, $\frac{dJ}{dP[r,k]} = 0$ conduit à autant d'équations que d'échantillons du spectrogramme. Pour contourner ce problème, une méthode itérative est proposée dans l'article, nous ne la détaillons pas ici.

Au final l'algorithme présente 3 paramètres d'entrée : le nombre d'itérations, γ le paramètre de compression du spectrogramme de puissance et, $\alpha = \frac{\sigma_P^2}{\sigma_H^2 + \sigma_P^2}$. Pour $0 < \alpha \leq 0.5$ on favorise la séparation de la partie percussive, pour $0.5 \leq \alpha < 1$ on favorise la séparation de la partie harmonique.

Ono & al., remarquent que la voix reste très présente dans la partie percussive séparée. Ceci est dû au fait que celle-ci ne peut pas être considérée comme stationnaire sur les trames d'analyse. De même, on remarque que tous les instruments effectuant des glissendi et des vibratos (évolution de partiels non horizontale) se retrouvent dans la partie percussive.

Nous présentons sur la *figure 11* les spectrogrammes obtenus pour, dans le premier cas, favoriser l'extraction de la partie percussive et, dans le second cas, la partie harmonique d'un extrait de musique. Le spectrogramme de puissance comprimé ($\gamma = 0.3$) est présenté en échelle linéaire pour une meilleure illustration de l'algorithme. On remarque que le bruit "hors des partiels" est contenu dans la partie percussive. Celui-ci peut contenir une partie des sons de cymbales.

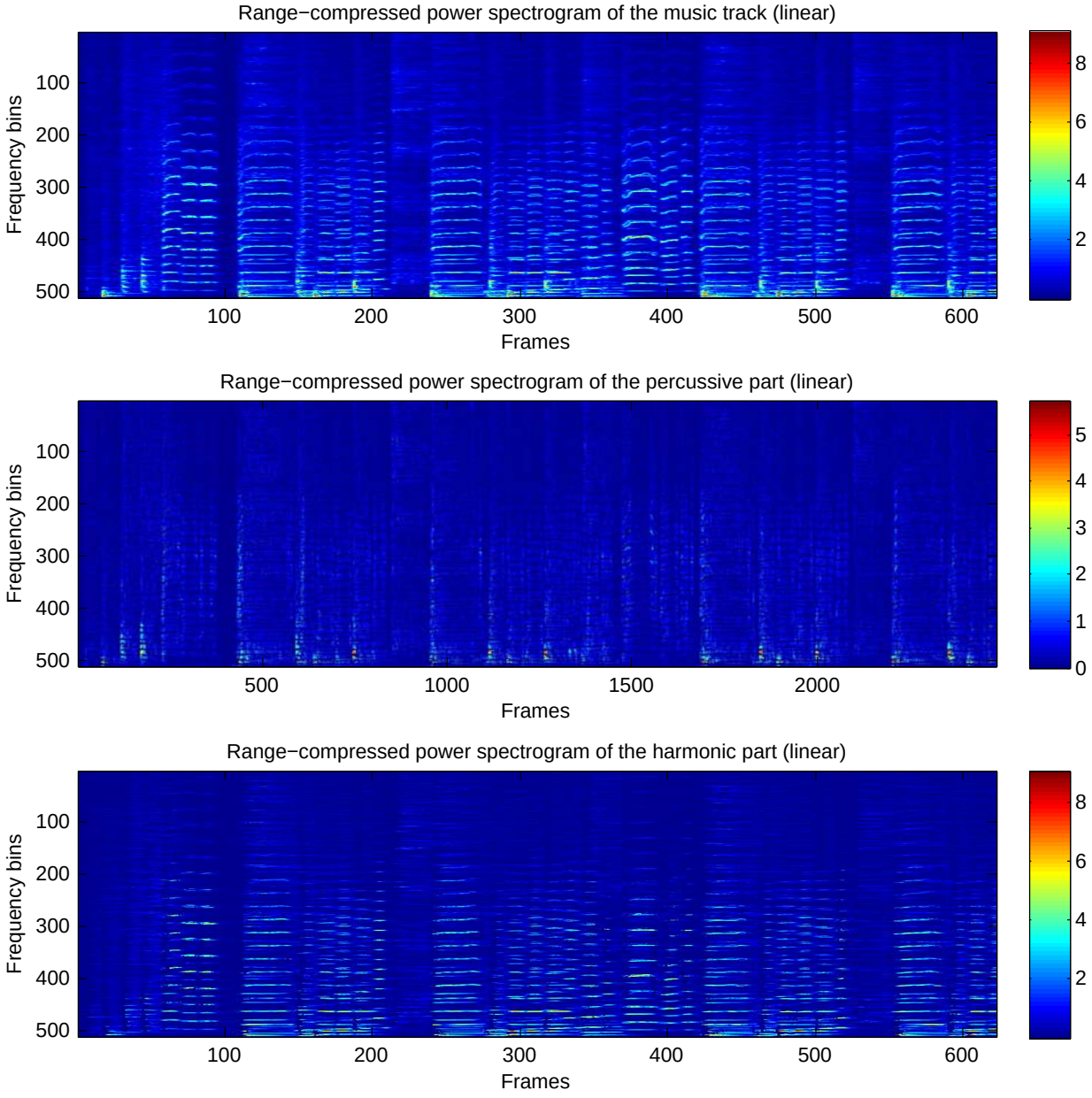


FIGURE 11 – Résultat de l'algorithme CDS pour 200 itérations. En haut, le spectrogramme de puissance comprimé du morceau calculé avec fenêtrage de Hanning de $60ms$. Au milieu la partie percussive extraite : $alpha = 0.3$, recouvrement de 93.75%. En bas la partie harmonique extraite : $alpha = 0.7$, recouvrement de 75%.

3 Méthodes du type *Match and adapt*

Ce type de méthode permet de reconstruire et d'extraire en même temps le son de chaque élément constitutif de la batterie d'un morceau de musique.

Pour cela l'algorithme nécessite une détection d'onsets. Ceci permet de ne pas avoir à balayer tout le morceau pour chercher les sons percussifs, en effet ceux-ci peuvent être contenus dans des intervalles de temps recouvrant les onsets.

Il reste alors à comparer le signal de musique à ces instants avec des signaux références ou "templates" (la forme d'onde en temporel [22] ou la densité spectrale de puissance dans le domaine spectral [23]) pour chaque élément de la batterie. Par exemple, si l'on possède un template de caisse claire on va chercher les zones à cheval sur un onset qui présentent la plus grande similarité. C'est l'étape *Match*.

Sur la durée du morceau une moyenne est effectuée et si l'on suppose que le son de caisse claire varie peu au cours du morceau (en pratique les caractéristiques seront légèrement variables, le son dépend de l'intensité et de la position de la frappe de la baguette sur la peau [24]) on obtiendra le son original propre à l'enregistrement. C'est l'étape *Adapt*.

Après affinage du template de référence on peut soustraire le son percussif au signal audio. On réalise au final une extraction, une transcription et une resynthèse de la partie percussive.

En pratique, la comparaison n'est pas évidente car le signal audio peut présenter un grand nombre d'instruments jouant simultanément. Ainsi, un son de cymbale pouvant être considéré comme un bruit large bande risque de passer inaperçu lors des comparaisons. De plus, afin d'effectuer un comparaison correcte il faut ajuster l'amplitude du template de référence au niveau du signal audio à chaque onset. Nous ne détaillons pas ici les moyens algorithmiques mis en oeuvre pour la comparaison, le lecteur intéressé pourra se référencer aux deux articles cités.

Troisième partie
Algorithmes développés

1 Modèle d'enveloppe et de trajectoire percussive

1.1 Des fonctions d'activations temporelles obtenues par NMF vers un modèle de trajectoire moyenne percussive

L'algorithme NMF fait l'hypothèse que le spectrogramme peut être factorisé (approximativement) en spectres stationnaires modulés temporellement par des fonctions d'activations. Lorsque l'on factorise le spectrogramme d'un seul son percussif on obtient un spectre moyen et une fonction de base temporelle moyenne. Ces deux courbes caractérisent approximativement le son étudié. Nous observons *figure 12* les fonctions de bases temporelles de différents éléments de batterie obtenues par NMF.

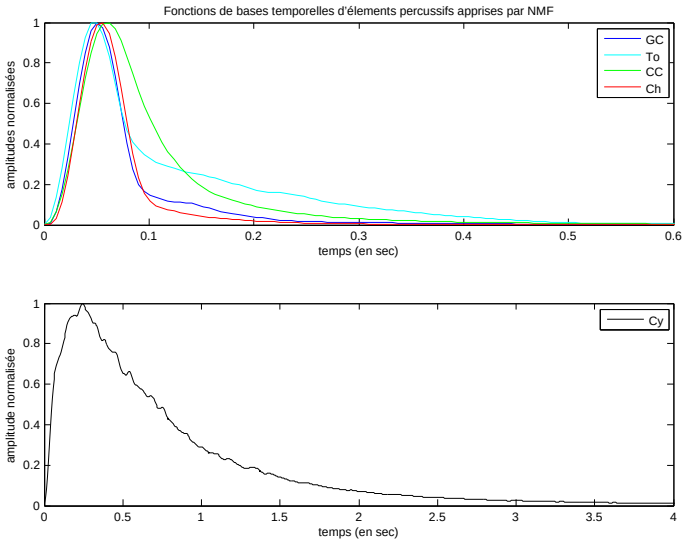


FIGURE 12 – Fonctions de bases temporelles de différents éléments de batterie (Grosse-Caisse, Tom, Caisse-Claire, Charleston fermé, Cymbale Crash) apprises par NMF. Spectrogrammes calculés avec fenêtre de Hanning de 90ms et avance de 1/16.

Dans un signal polyphonique le mélange des composantes fréquentielles rend difficile la distinction et l'extraction des instruments en étudiant les spectres de chaque trame du spectrogramme (cf. les méthodes *match and adapt* présentées partie II.3.). Cependant, si l'on regarde dans la direction de l'axe temporel du spectrogramme, on peut espérer classer les bins fréquentiels en fonction de leur trajectoire dans le temps. L'algorithme développé lors de ce stage se base sur cette idée, qu'un bin sera défini plutôt "percussif" ou "harmonique" selon sa trajectoire temporelle.

Dans le domaine temporel, selon le modèle ADSR [8] de l'enveloppe temporelle d'un son (illustré *figure 13*), une enveloppe percussive sera caractérisée par une quasi

8. Attack Decay Sustain Release

absence de *sustain* (temps très court, voire nul et amplitude faible) et d'un *decay-release* rapide, si la réverbération n'est pas trop importante [9].

Remarque : nous veillerons à parler d'**enveloppe percussive** pour l'enveloppe temporelle d'un son de percussion et de **trajectoire percussive** pour un bin fréquentiel du spectrogramme classé percussif.

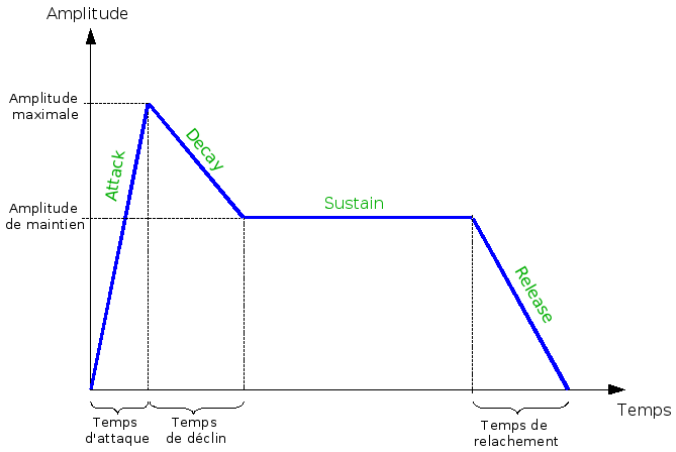


FIGURE 13 – Modèle ADSR d'enveloppe de la forme d'onde temporelle d'un son.

Cette considération se retrouve dans le plan temps-fréquence. On présente *figure 14* les spectrogrammes avec normalisation en amplitude du maximum pour chaque bin fréquentiel de divers éléments de batterie et d'un note de piano Do3.

Pour la grosse-caisse et le charleston fermé, cette hypothèse de trajectoire percussive semble se vérifier en moyenne (lissage des oscillations) pour l'ensemble des bins. Pour la caisse-claire et le tom, on note la présence de partiels comportant un sustain important par rapport aux autres bins. Enfin, pour la cymbale, le spectrogramme comporte de nombreuses résonances chaotiques et la décroissance globale du son est lente comparée aux autres éléments percussifs.

On constate aussi pour le piano qu'en dehors des partiels, certains bins possèdent une trajectoire percussive. Ceci résulte de la frappe du marteau sur les cordes. Si ces bins (de faible amplitude par rapport aux bins harmoniques) ne sont pas recouverts par d'autres instruments, ils seront extraits avec les percussions de batterie. De manière générale, tous les instruments excités par percussion feront l'objet de cette séparation partie percussive / harmonique.

Au final, de la même façon que l'algorithme *Complementary Diffusion Spectrogram* présenté dans l'état de l'art, un masque binaire sera construit et appliqué sur le spectrogramme afin de séparer les deux parties. Pour la resynthèse la matrice des phases sera appliquée sur les deux spectrogrammes.

La séparation ne sera pas optimale, mais selon les applications envisagées elle pourra être suffisante. Pour une écoute "simple", ou un remixage, le phénomène de masquage de l'oreille interne permettra de combler les vides du spectrogramme et d'obtenir un son satisfaisant. Pour une application de détection multipitch, si les

9. Voir les formes d'ondes pour divers éléments de batterie présentées en I.2.2. *figure 3*

partiels ne sont pas accidentellement extraits dans la partie percussive on aura réalisé un débruitage efficace.

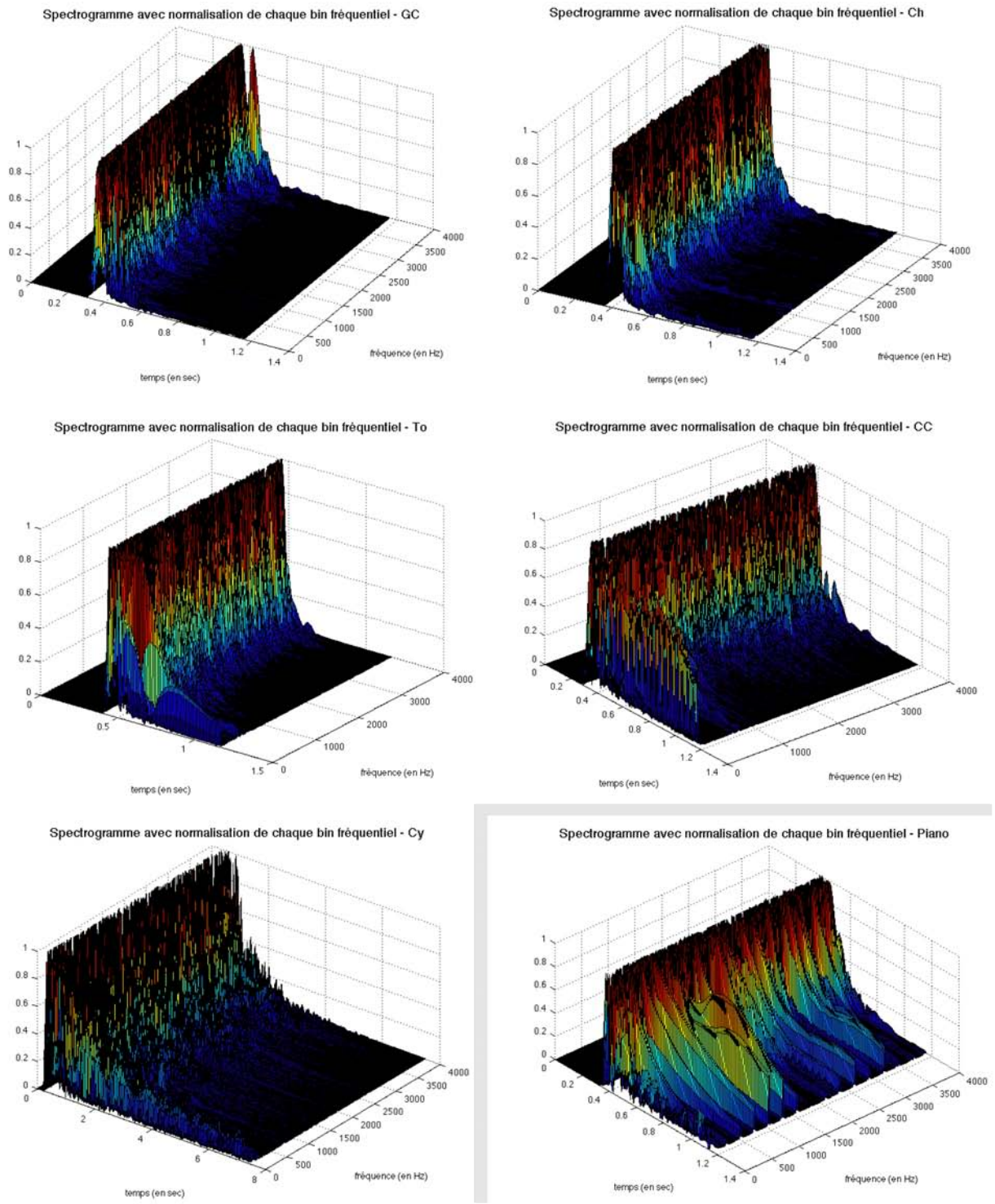


FIGURE 14 – Spectrogrammes (échelle linéaire) avec normalisation en amplitude du maximum pour chaque bin fréquentiel de différents éléments de batterie (Grosse-Caisse, Charleston Fermé, Tom, Caisse-claire, Cymbale Crash) et d'une note de piano Do3. Sons échantillonés à 8kHz.

1.2 Modèle paramétrique de l'enveloppe temporelle du son produit par la percussion d'une peau

Comme remarqué précédemment, l'enveloppe temporelle d'un son percussif se limite à une phase d'attaque et une phase de décroissance. Nous recherchons ici un modèle paramétrique simple pour la décrire.

- Pour étudier la partie décroissante de l'enveloppe temporelle nous utilisons l'expression mathématique du champ de pression acoustique résultant de la frappe d'une timbale [10] [25]. La timbale est modélisée par une membrane circulaire homogène uniformément tendue sur une cavité et mise en vibration par un impact. En considérant les pertes viscoélastiques et le couplage de la peau avec l'air dans l'équation de propagation, on peut montrer que le déplacement transversal de la membrane résulte de la contribution de chaque mode de vibration dont l'amplitude est modulée par une courbe exponentielle décroissante en fonction du temps $e^{-\alpha t}$ (correspond pour une onde à des pulsations propres complexes de la forme $\omega = 2\pi f_0 + j\alpha$).
  En se plaçant en champ lointain et en considérant qu'un écran plan rigide empêche le court-circuit acoustique entre les deux faces de la membrane il est possible de déterminer une expression simple de la pression captée par le microphone lors de l'enregistrement. Le champ de pression complexe obtenu est de la forme suivante :

$$P(r, \theta, \phi, \omega, t) = -\frac{\omega^2 \rho}{2\pi r} \cdot e^{-jkr} \cdot W(k, \theta, \phi, \omega, t)$$

  Avec, r, θ, ϕ les coordonnées sphériques, ω la pulsation, k le vecteur d'onde, ρ la masse volumique de l'air et W la transformée de Fourier **spatiale** du déplacement transversal de la membrane. Le terme multiplicatif $e^{-\alpha t}$ ne dépendant pas des coordonnées spatiales, il reste présent dans W et correspond de même à la décroissance de l'amplitude du champs de pression en fonction du temps. En supposant que le microphone transforme linéairement la pression acoustique en tension électrique alors ce modèle de décroissance du son peut être utilisé.

- La modélisation physique du transitoire d'attaque est complexe, nous considérons simplement une croissance linéaire rapide (quelques ms).

Ainsi le modèle d'enveloppe normalisée se présente sous la forme :

$$env(t) = \begin{cases} t/t_a & \text{pour } \ t \leq t_a \\ e^{-\alpha(t-t_a)} & \text{pour } \ t > t_a \end{cases}$$

Le paramètre t_a correspond à l'instant de passage de l'attaque à la décroissance. Celui-ci varie environ entre 2 et 6 ms selon le type de peau étudiée. α est le coefficient d'amortissement.

Nous paramétrons *figure 15* les enveloppes de différents éléments de la batterie. Les enveloppes théoriques ne collent pas parfaitement aux enveloppes réelles (notamment pour le tom). Ceci vient du modèle qui considère un coefficient d'amortissement

10. Instrument de la famille des "peaux" comportant la particularité de contribuer à la partie rythmique et tonale d'un morceau de musique. Les sons qu'il produit possèdent une hauteur bien définie.

indépendant de la fréquence. Nous avons vu dans la partie précédente sur les spectrogrammes "normalisés" qu'en dehors de la trajectoire moyenne des bins des sons de grosse-caisse et du charleston fermé cette hypothèse n'est pas correcte.

Bien que le modèle d'enveloppe temporelle soit construit à partir de l'étude des vibrations et du rayonnement des peaux, on remarque que celui-ci colle relativement bien au son de charleston fermé et peut approximer celui des cymbales. On peut donc le généraliser à l'ensemble des éléments percussifs de la batterie.

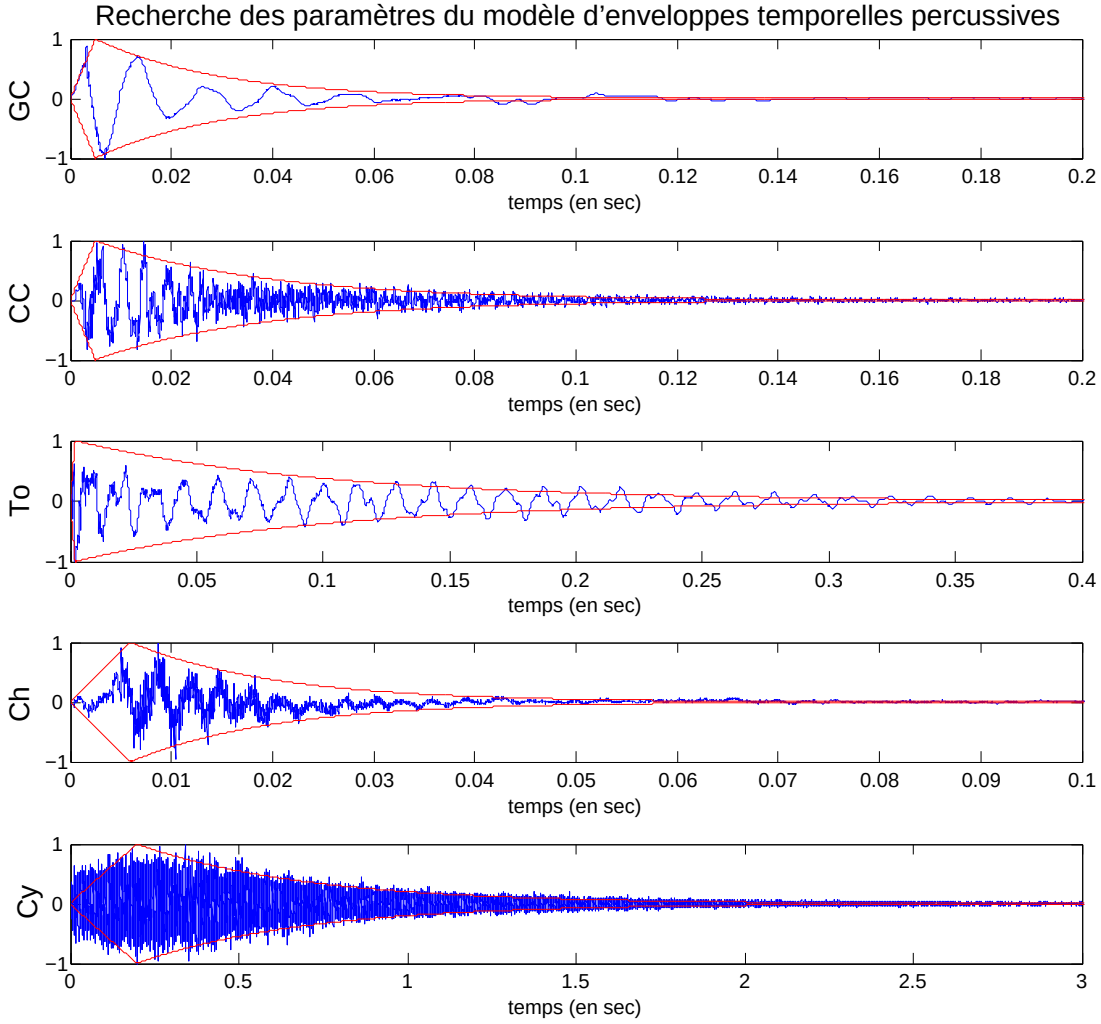


FIGURE 15 – Recherche de t_a et α pour différents éléments de la batterie. Grosse-Caisse : $t_a = 5ms$ et $\alpha = 40$ / Caisse-Claire : $t_a = 5ms$ et $\alpha = 30$ / Tom : $t_a = 2ms$ et $\alpha = 10$ / Charleston fermé : $t_a = 6ms$ et $\alpha = 70$ / Cymbale Crash : $t_a = 0.2s$ et $\alpha = 2$.

1.3 Modèle paramétrique de la trajectoire temporelle moyenne d'un bin percussif

L'enveloppe paramétrique définie correspond à l'enveloppe temporelle du son. Pour étudier la trajectoire d'un bin dans le spectrogramme, il faut tenir compte de l'effet de la transformée de Fourier à court terme sur cette enveloppe.

Prenons une exponentielle complexe modulée temporellement par une enveloppe percussive :

$$x(t) = env(t) \cdot e^{j2\pi f_0 t}$$

La TFCT correspond au calcul de la TF du signal fenêtré soit

$$X[\tau, f] = TF[x(t) \cdot w(t - \tau)]$$

Si la fenêtre est symétrique alors $w(t - \tau) = w(\tau - t)$.
On peut écrire

$$X[\tau, f] = \int_{-\infty}^{\infty} e^{-j2\pi(f - f_0)t} \cdot env(t) \cdot w(\tau - t) \cdot dt$$

Le calcul peut être développé pour différents types, tailles et avances de fenêtres. Intéressons nous à l'évolution temporelle de la fréquence $f = f_0$.

$$X[\tau, f = f_0] = \int_{-\infty}^{\infty} env(t) \cdot w(\tau - t) \cdot dt$$

Soit,

$$X[\tau, f = f_0] = (env * w)(\tau)$$

Pour la fréquence modulée, l'amplitude suit une trajectoire correspondant à la convolution de l'enveloppe temporelle par la fenêtre d'analyse. Pour d'autres fréquences, le calcul est plus complexe mais on constate *figure 17* que les bins suivent des trajectoires quasi-identiques en moyenne.

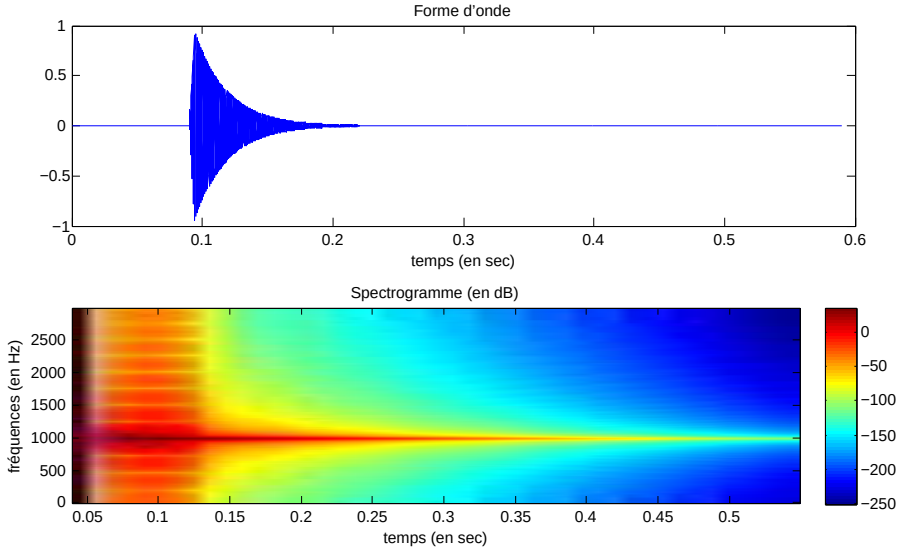


FIGURE 16 – Spectrogramme d'une sinusoïde de fréquence 1kHz modulée par une enveloppe percussive ($t_a = 4ms$ et $\alpha = 40$) et échantillonée à 3kHz. Fenêtrage de Haning de durée 90 ms et d'avance 1/8.

Le transitoire d'attaque étant généralement beaucoup plus court que la décroissance et la fenêtre d'analyse, l'influence du paramètre t_a sera minime. Nous jouerons uniquement sur le coefficient d'amortissement α pour générer une famille de trajectoires temporelles de bin percussifs (cf. *figure 18*) avant d'effectuer des comparaisons sur le spectrogramme traité.

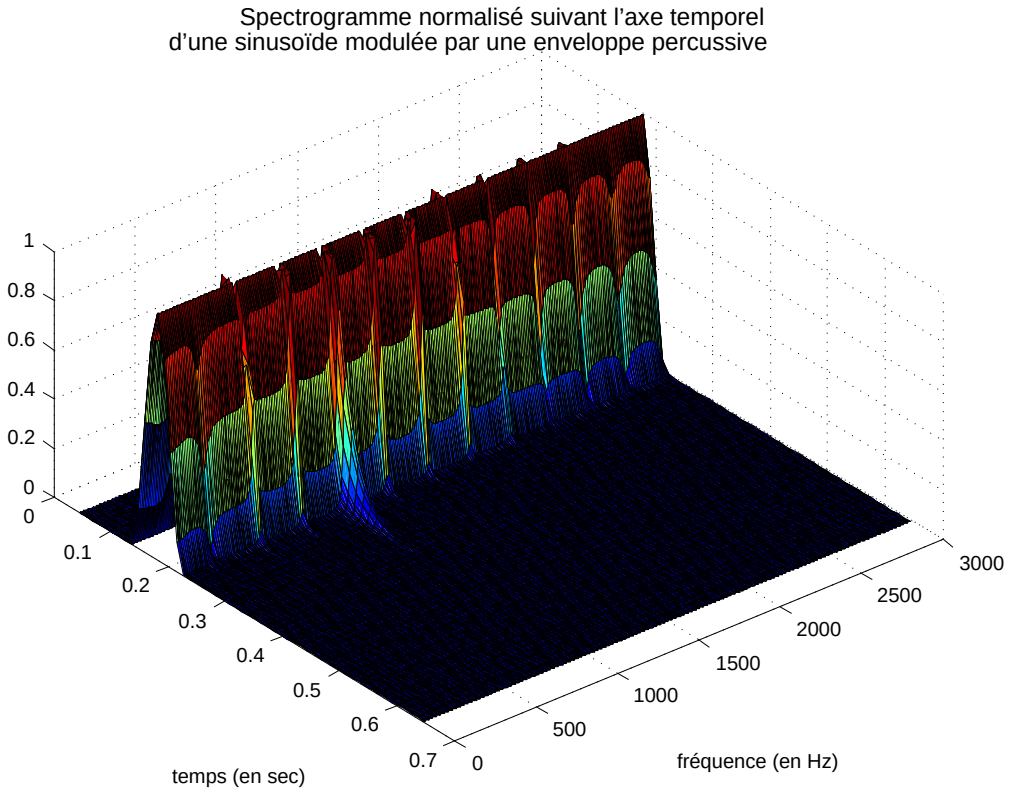


FIGURE 17 – Spectrogramme (échelle linéaire) avec normalisation en amplitude du maximum pour chaque bin, de la sinusoïde modulée représentée *figure 16*.

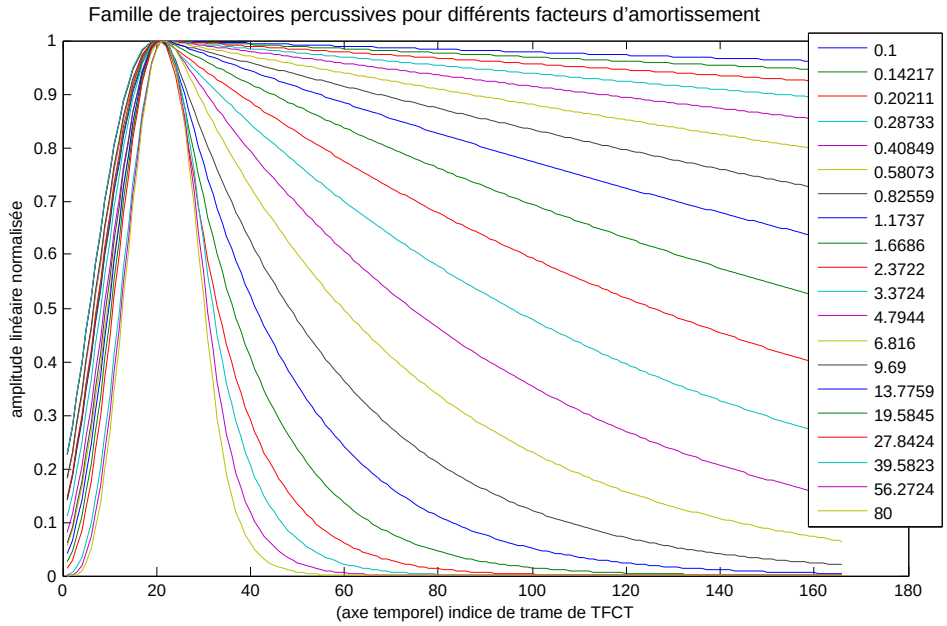


FIGURE 18 – Famille de trajectoires percussives générées pour $t_a = 4ms$, α variant logarithmiquement entre 0.1 et 80 et pour un fenêtrage de Hanning de longueur 90ms et d'avance 1/32.

2 Comparaison de trajectoires

2.1 Utilisation de la détection d'onset pour la définition des zones de comparaisons

Pour définir le début des zones de comparaison entre les bins du spectrogramme et les trajectoires de référence générées via notre modèle, nous utilisons la détection d'onset développée à l'IRCAM par Axel Roebel [2] [3].

Les trajectoires de référence sont calculées pour une enveloppe temporelle de 400 ms. La durée de comparaison correspond alors au minimum entre la durée de la trajectoire de référence et la durée séparant deux onsets. En général, la durée entre deux onsets est plus courtes que 400 ms (par exemple, 1 croche toutes les 0,4 secondes correspond à un tempo de 75 bpm). Si la durée d'extraction est trop courte on risque d'entendre une réverbération de batterie "éloignée" dans la partie harmonique. Ceci peut arriver si les bins percussifs ne sont pas réaffectés à la batterie à l'onset suivant. Cependant, mis à part les cymbales qui possèdent des résonances trop longues pour être totalement prises en compte, la durée entre 2 onsets suffit généralement pour extraire correctement les percussions de peaux et de charleston fermé. Une illustration de ces zones de comparaison est proposée *figure 19*.

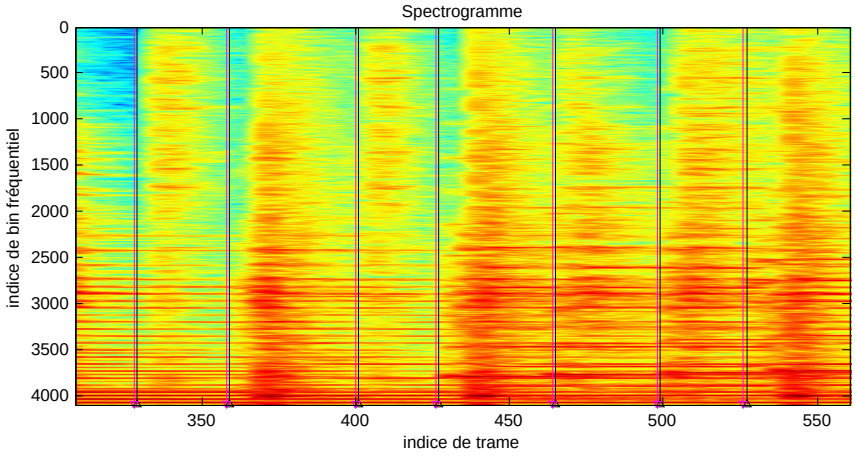


FIGURE 19 – Visualisation des zones de comparaison sur le spectrogramme (hanning, 90ms, 1/16) d'un morceau de musique. Sont représentés, en noir les instants d'onsets et en violet les fins de comparaisons (toujours définies par l'apparition d'un nouvel onset dans cet exemple).

2.2 Mesure de distance entre deux trajectoires

Pour comparer la trajectoire d'un bin à une trajectoire percussive de référence nous introduisons la mesure d'une distance. Pour chaque comparaison les amplitudes sont normalisées. Nous visualisons *figure 20* la comparaison de la trajectoire d'un bin avec une trajectoire de référence calculée pour $t_a = 4ms$ et $\alpha = 30$.

Pour étudier l'évolution moyenne du bin, nous proposons de calculer l'intégrale de la courbe de différence *trajectoire_bin - trajectoire_ref*. Ainsi, pour une trajectoire de bin oscillant autour de la trajectoire de référence la distance sera nulle. Si elle

est globalement inclue dans la trajectoire de référence, la distance sera négative. Au contraire, dans le cas d'un bin correspondant à un partiel, la distance sera positive.

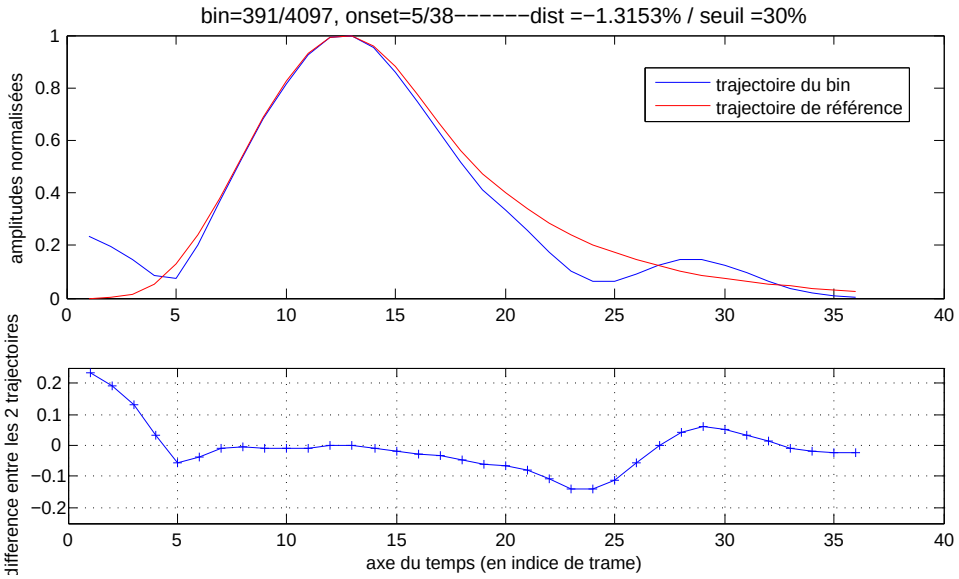


FIGURE 20 – Comparaison de la trajectoire d'un bin avec une trajectoire de référence ($t_a = 4ms$ et $\alpha = 30$). (Spectrogramme : hanning, 90ms, 1/16)

Afin d'obtenir une distance normalisée significative pour n'importe quel bin traité, nous divisons la distance mesurée par la distance la plus grande possible (cas ou le bin reste à '1' tout le temps, cf. *figure 21*).

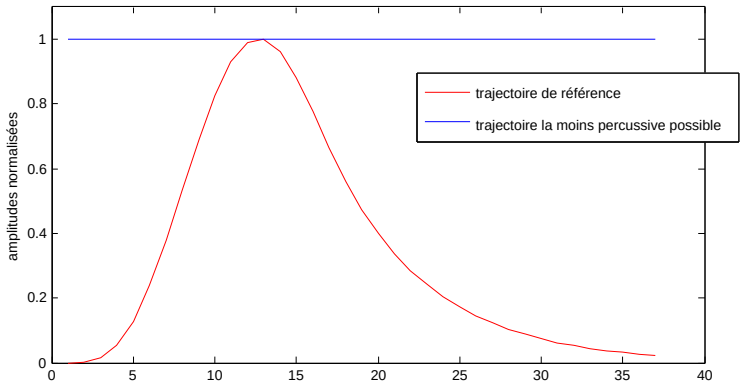


FIGURE 21 – Distance pour la normalisation correspondant à la comparaison d'un bin "le moins percussif" possible. (Spectrogramme : hanning, 90ms, 1/16)

Au final la distance pour un bin de longueur N se calcule de la façon suivante :

$$d = \frac{\sum_{n=1}^{N} trajectoire\_bin[n] - trajectoire\_ref[n]}{\sum_{n=1}^{N} trajectoire\_bin[n] - \mathbb{1}_{[0,N]}[n]}$$

2.3 Précision du calage temporel avant la mesure de distance

La mesure de distance entre la trajectoire d'un bin et une trajectoire de référence est fonction de leur position relative. Deux méthodes ont été testées pour essayer d'ajuster au mieux celle-ci :

- Il est judicieux de penser à une détection de la position du maximum de l'intercorrélation entre les deux signaux (ce qui correspond à la position pour laquelle les deux courbes présentent une ressemblance maximale), cependant comme la comparaison est effectuée ici pour l'ensemble des bins correspondant à chaque onset avec une ou plusieurs enveloppes de références (selon l'algorithme), le temps de calcul augmente énormément (même si celles-ci sont effectuées par multiplications dans le domaine de Fourier).

- La solution simple qui a été retenue pour l'algorithme 1 est d'ajuster le maximum de la trajectoire de référence sur le maximum de la trajectoire du bin. Cette méthode est approximative, surtout si l'avance de la fenêtre d'analyse de la TFCT est grande (on choisira en général 1/16).

3 Algorithme 1 : Extraction rapide mais limité

3.1 Limitation aux éléments grosse-caisse, caisse-claire, charleston fermé

Dans la partie III.1.2. nous avons remarqué que les enveloppes des sons de grosse-caisse, caisse-claire et charleston fermé possédaient les plus forts coefficient d'amortissement (respectivement 40, 30 et 70). Pour séparer ces trois éléments, il "suffit" *a priori* de générer une trajectoire correspondant à la décroissance la plus lente (i.e. la caisse-claire, $\alpha = 30$) et de considérer tous les bins ayant une trajectoire inclue dans cette enveloppe de référence comme appartenant à la partie percussive. Pour déterminer si un bin évolue de façon percussive ou non en moyenne, on place simplement un seuil sur la distance mesurée lors de la comparaison avec la trajectoire de référence. Normalement, les bins correspondant à la grosse caisse et au charleston fermé devraient être totalement inclus dans la trajectoire de référence et donc avoir une distance négative. Le seuil sera alors fixé supérieur à zéro, par exemple 30% (voir les mesures de seuil optimal partie IV.).

3.2 Prise en compte des effets dégradant la partie harmonique

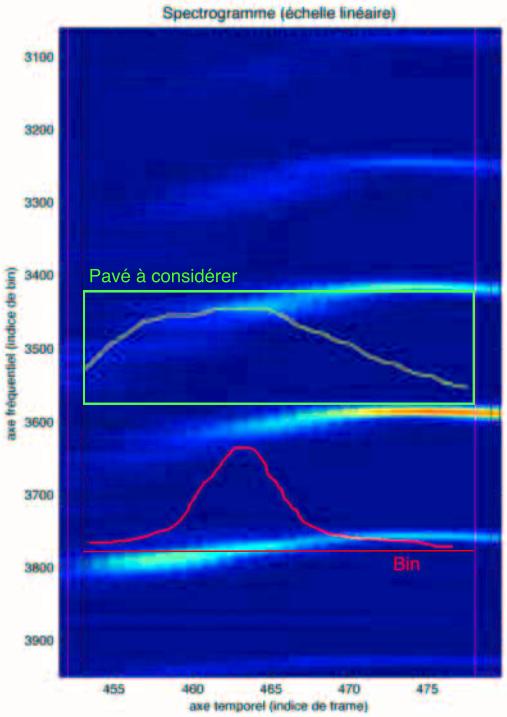
A ce niveau de l'algorithme on peut remarquer par écoute des deux signaux de sortie des erreurs de séparation.

Comme on l'avait prévu partie III.1.1. en visualisant la figure *figure 14*, les partiels associés à la caisse claire et le partiel de grosse caisse ne sont pas extraits de la partie harmonique. Ceux-ci se retrouvent dans le résiduel (piste de batterie du

multipiste à laquelle on soustrait la piste de batterie extraite).

Dans l'autre sens, on retrouve dans la partie percussive des sons de la partie harmonique. Ceux-ci traduisent les limitations de notre modèle de discrimination entre partie percussive et harmonique. En testant sur plusieurs morceaux de musique nous avons recensés certains points où le modèle d'enveloppe n'est pas suffisant et nous présentons des améliorations apportées à l'algorithme.

– Le fait de regarder l'évolution d'un bin le long de l'axe temporel limite la séparation à des signaux dont les fréquences ne sont pas modulées dans le temps [11]. Ainsi, tous les glissandi et vibratos sont considérés comme appartenant aux percussions.



Pour les partiels de la partie harmonique il faudrait en théorie faire un suivit de pic et regarder leurs trajectoires dans le temps. En pratique ceci est complexe, nous choisissons plutôt d'effectuer un post-traitement qui suppose que si un ensemble de bin voisins est considéré comme percussifs, alors la trajectoire moyennée sur le pavé doit aussi avoir une allure percussive. La *figure* de gauche illustre ce phénomène sur une partie du spectrogramme (en échelle linéaire) présentant du chant.
Cette méthode permet de réduire l'apparition de la voix dans la partie percussive (voir les mesures partie IV.2.2.).

– Un autre problème se pose avec les partiels à fréquence constante, lié à la modulation d'amplitude de l'enveloppe temporelle. Les bins correspondant à l'attaque du son possèdent une allure percussive (zone orangée au départ du partiel présenté *figure 22*) mais ils ne doivent pas être extraits avec la partie percussive. Nous proposons un méthode (qui n'a pas encore été évaluée) permettant de tenir compte de ce phénomène.

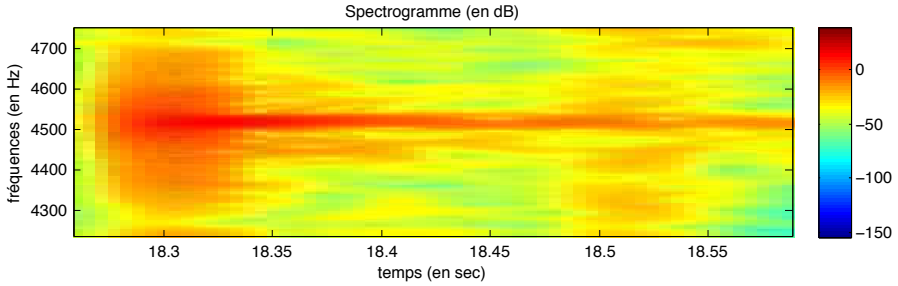


FIGURE 22 – Zoom sur un partiel de piano. Spectrogramme : hanning, 90ms, 1/16

11. problème identique à l'algorithme Complementary Diffusion Spectrogram

En visualisant l'évolution de la distance entre les trajectoires des bins et la trajectoire de référence le long de l'axe fréquentiel pour un onset donné, on se rend compte que les partiels correspondent à des pics. Plutôt que d'appliquer le seuil sur la mesure de distance de chaque bin, nous proposons d'appliquer un seuil sur les maxima locaux. Si un maxima est détecté comme non percussif, alors nous extrayons le pic complet (compris entre deux minima locaux) et le considérons comme appartenant à la partie harmonique. Une illustration est proposée *figure 23*.

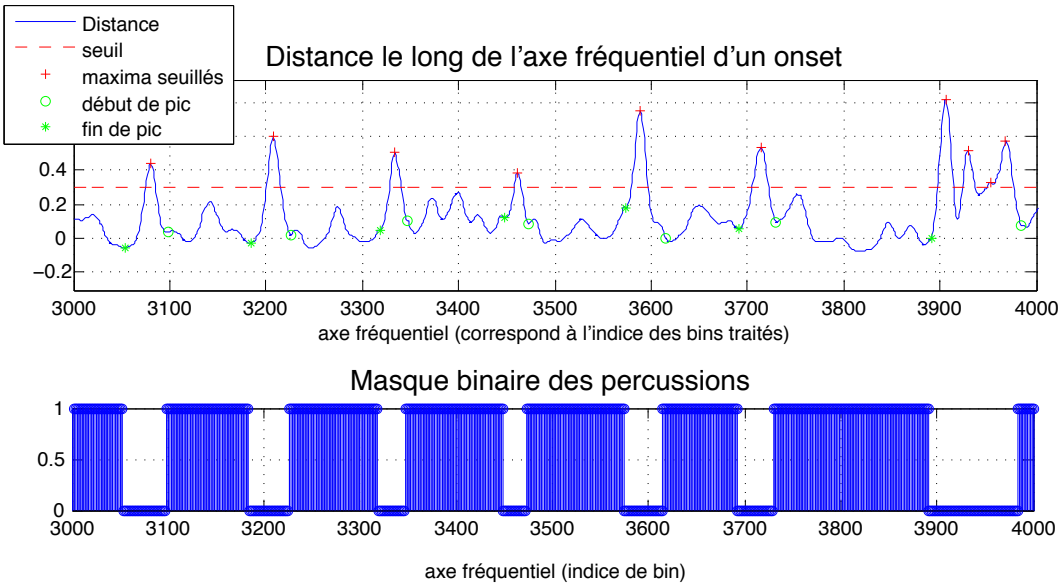


FIGURE 23 – Application du seuil sur les maxima de la courbe de distance

– Enfin, tous les instruments dont on étouffe immédiatement le son après l'attaque comportent une enveloppe temporelle percussive et donc sont extraits avec la batterie. Ceci se retrouve par exemple dans le cas de guitare jouée en Palm Mute [12]. Cette limitation ne pourra pas être contournée dans notre algorithme.

4 Algorithme 2 : Prise en compte de la mesure du niveau de bruit pour l'extraction des toms et cymbales (développement en cours)

4.1 Recherche du alpha optimal pour chaque bin

Le coefficient d'amortissement α est une caractéristique de chaque élément percussif. Ainsi, pour pouvoir espérer extraire entièrement la batterie, il semble nécessaire de prendre en compte plusieurs trajectoires percussives de référence (voir la famille

12. Le bord de la main repose sur les cordes au niveau du chevalet pour étouffer la vibration des cordes.

de trajectoires affichée partie III.1.3. *figure 18*) . Contrairement à l'algorithme précédent où nous prenions une unique trajectoire de référence et nous extrayions les bins ayant des trajectoires ressemblantes (jusqu'à un certain degré fixé par le seuil), ici nous allons chercher pour chaque bin la trajectoire de référence ("fittant" au mieux) tel que le paramètre α minimise la distance introduite partie III.2.2. .

Nous générons 60 trajectoires pour α variant entre 90 en 0.1 de manière logarithmique afin de balayer tout le plan (amplitude,temps). Ici, seule la décroissance des trajectoires nous intéresse nous calculerons donc la distance uniquement sur cette zone et non plus sur toute la trajectoire.

4.2 Utilisation du spectrogramme de niveau de bruit

Dans le développement de l'algorithme 1 nous avons remarqué que les bins voisins sur l'axe fréquentiel possédaient des trajectoires variables et nous cherchions en moyenne celles qui ressemblaient le plus à la trajectoire de référence. Pour obtenir une certaine régularité dans l'évolution des bins voisins nous proposons d'introduire le spectrogramme de niveau de bruit.

Celui-ci est calculé grâce à l'algorithme *Adaptative Noise Level Estimation* développé à l'IRCAM par Chungsin Yeh [26]. Pour chaque trame d'analyse nous obtenons l'évolution du niveau de bruit du spectre (cf. *figure 24*), et nous construisons le spectrogramme correspondant (cf. *figure 25*). Entre deux trames, les variations du niveau de bruit seront beaucoup plus faibles que les variations d'un unique bin et celles-ci devraient traduire son évolution moyenne si le bin n'appartient pas à un pic sinusoïdal (cf. *figure 27*).

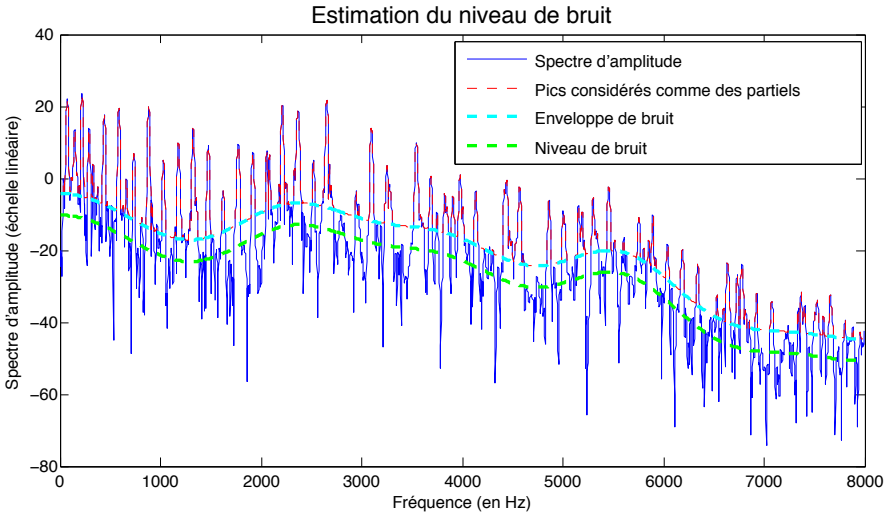


FIGURE 24 – Estimation du niveau de bruit d'un spectre.

4.3 Deux approches en cours de développement

Si nous considérons que la batterie produit essentiellement des sons appartenant au bruit (les partiels ne sont pas pris en compte), les trajectoires des bins du spectrogramme doivent suivre en moyenne les trajectoires des bins du spectrogramme de niveau de bruit (calculé pour les mêmes paramètres d'analyse).

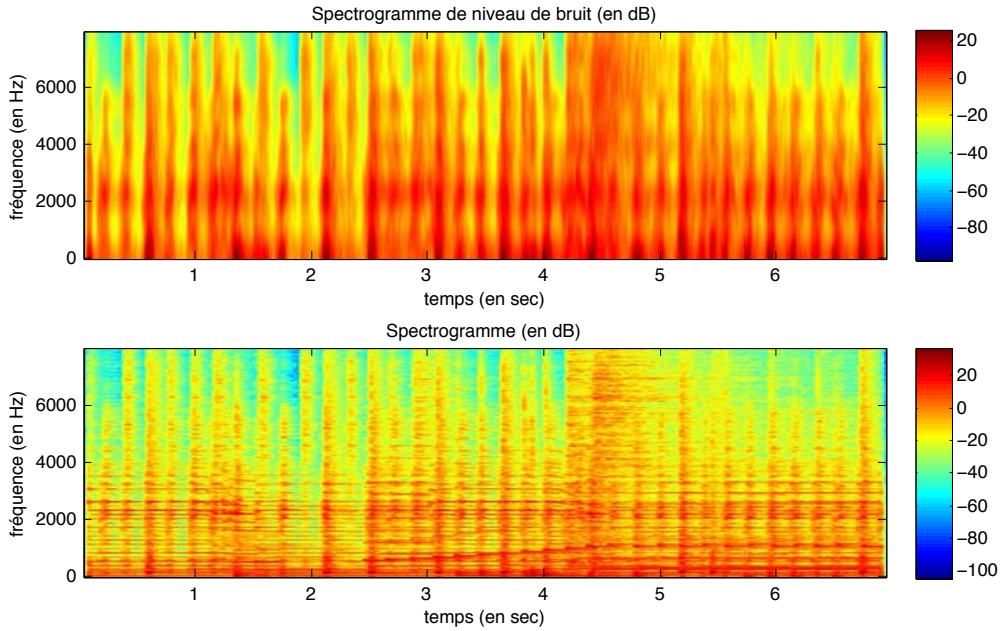


FIGURE 25 – Spectrogramme de niveau de bruit et spectrogramme d'un morceau de musique (hanning, 90ms, avance 1/16).

Ainsi, nous proposons deux approches pour mesurer, pour un bin donné, la similarité entre ses trajectoires dans les deux spectrogrammes :

– Une première démarche, consiste à chercher le paramètre α correspondant à ses deux trajectoires, soit α_{signal} et α_{bruit} et à les comparer. Pour un bin défini plutôt percussif, la variation de α entre les deux trajectoires devrait être relativement faible. On pourra alors appliquer un seuil sur cette différence. Nous illustrons *figure 26* l'évolution de α_{signal} et α_{bruit} le long de l'axe fréquentiel pour un onset.

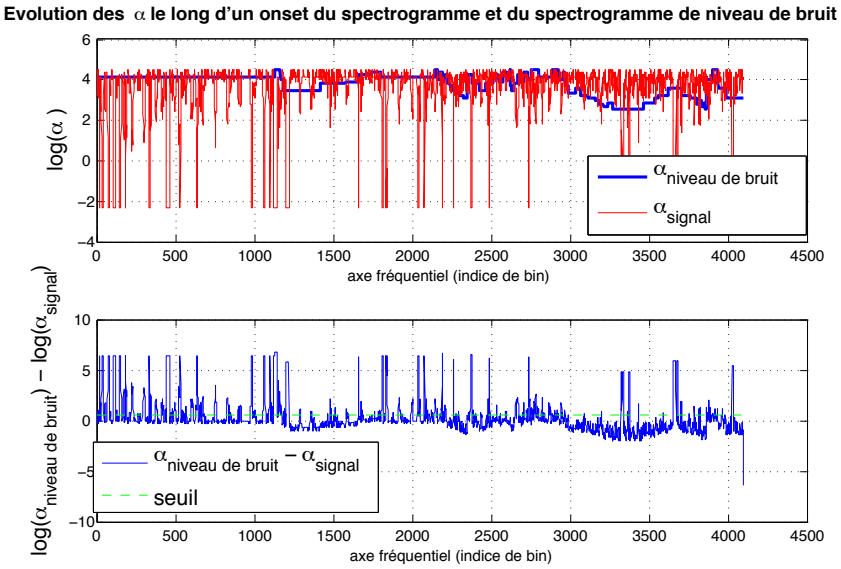


FIGURE 26 – Comparaison de l'évolution de α le long de l'axe fréquentiel pour un onset donné dans le spectrogramme de niveau de bruit et le spectrogramme.

Cependant, l'estimation de α semble peu robuste. Entre deux trajectoires très similaires, celui-ci peut varier de manière importante. De plus, elle est fortement influencée par la position relative des trajectoires comparées (cf. problème de calage présenté partie III.2.3.).

– Pour limiter le nombre d'estimations de paramètres α, la seconde méthode proposée consiste à uniquement estimer α_{bruit} et à comparer la trajectoire de référence ainsi retenue à la trajectoire du bin dans le spectrogramme. On revient ainsi vers l'approche proposée dans l'algorithme 1, sauf qu'ici, pour chaque comparaison, la trajectoire de référence n'est plus fixée arbitrairement de manière à extraire la grosse-caisse, le charleston fermé et la caisse-claire, mais est adaptée à chaque élément en fonction de l'évolution de son niveau de bruit. Comme précédemment on pourra appliquer un seuil sur la distance calculée et reprendre les méthodes de traitement de pavé de bins (correspondant aux modulations de fréquences) et d'attaques de notes.

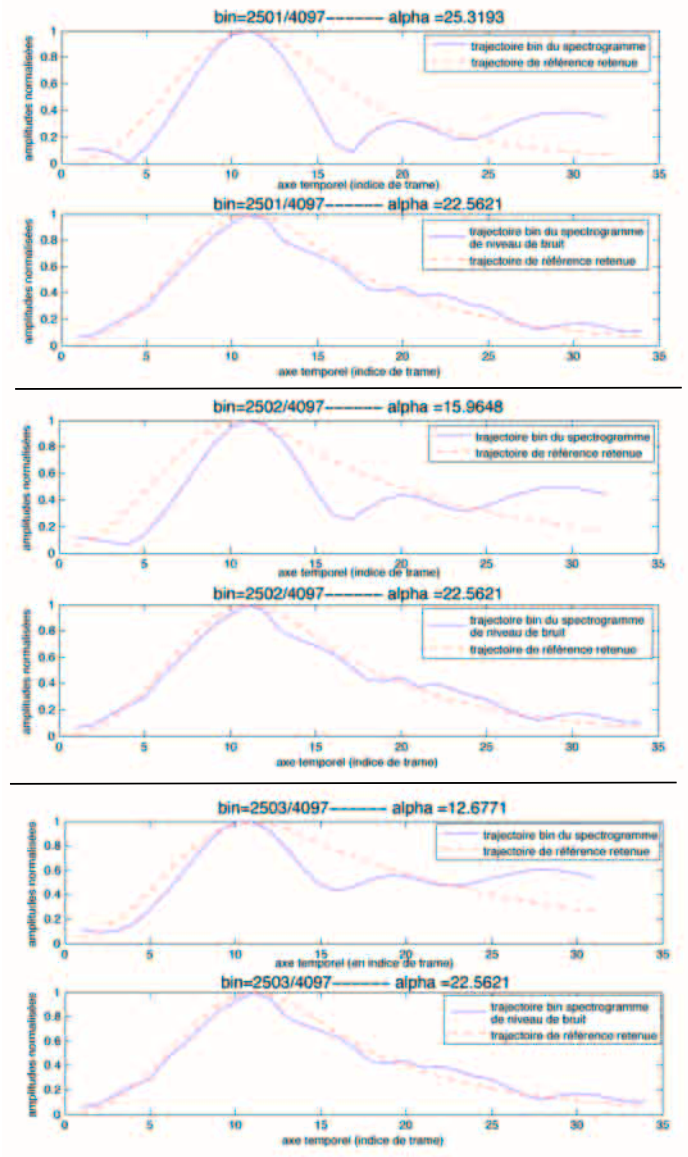


FIGURE 27 – Evolution de 3 bins consécutifs du spectrogramme et du spectrogramme de niveau de bruit.

Quatrième partie

Mesures de performances et comparaison avec certains algorithmes de l'état de l'art

1 Présentation du corpus

Pour comparer les performances de notre algorithme (pour l'instant seule la première version est étudiée) nous travaillons sur un corpus composé de 6 extraits de morceaux de musique mono échantillonnés à 16 kHz sous forme de multipistes :

1. Marvin Gaye - I Heard It Through The Grapevine : Soul - 11 sec
   Basse électrique / Guitare électrique clean / Orgue électrique / Ensemble de cordes / Voix / Choeurs / Batterie acoustique / Congas
2. Stevie Wonder - Superstition : Funk - 9 sec
   Basse électrique distordue / Clavinet (8 pistes) / Cuivres / 2 Voix / Batterie acoustique
3. Vieux Farka Toure - Ana : Blues / World - 13 sec
   Basse électrique / Guitare électrique clean / Guitare acoustique / 2 Voix / Batterie acoustique + percussions
4. Kismet - TV On : Rock - 6 sec
   Basse électrique/ Guitare électrique clean / Guitare électrique crunch / Piano / Batterie acoustique
5. Berlin - Roads : Ballade Pop - 14 sec
   Guitare Acoustique / Basse électrique / Synthétiseur / Piano / Voix / Batterie Acoustique
6. Fort Minor - Remember The Name : Rap US - 11 sec
   Basse synthétiseur / Ensemble de cordes / Voix / Boite à rythme

2 Mesures de performances

2.1 Définition du Rapport Signal sur Résiduel

Pour quantifier la qualité de la séparation réalisée nous calculons les rapports signal sur résiduel (RSR) en dB des parties percussives et harmoniques. Celui-ci mesure le rapport en dB de l'énergie d'une source sur l'énergie de l'erreur d'estimation de cette source. Plus cette quantité est élevée plus la séparation est bonne.

$$RSR_{db}(harmo) = 10\log_{10}\frac{\sum_{n=1}^{N} s_{harmo}[n]^2}{\sum_{n=1}^{N}(s_{harmo}[n]-\hat{s}_{harmo}[n])^2}$$

$$RSR_{db}(percu) = 10\log_{10}\frac{\sum_{n=1}^{N} s_{percu}[n]^2}{\sum_{n=1}^{N}(s_{percu}[n] - \hat{s}_{percu}[n])^2}$$

Remarque : En général (pas le cas de la NMF) pour les algorithmes de séparation en 2 sources, le résiduel sera identique pour les 2 sources si la chaine de traitement n'introduit pas de bruit. En effet, on a $s = s_{harmo} + s_{percu}$ et $\hat{s}_{harmo} = s - \hat{s}_{percu}$. En estimant une source, par soustraction avec le signal original on obtient la seconde source. Ainsi $\hat{s}_{harmo} - s_{harmo} = \hat{s}_{percu} - s_{percu}$.

2.2 Evaluation des performances de l'algorithme 1

Pour étudier l'influence de la valeur du seuil appliqué sur la distance dans notre algorithme, nous réalisons des mesures de RSR pour un seuil variant de 0 à 35 avec un pas de 5 (rappel : la distance peut-être négative). De même, nous regardons l'influence du traitement des pavés (contenant des modulations de fréquences) présentée partie III.3.2. Les paramètres "internes" à l'algorithme choisis sont les suivants : fenêtre de hanning de 90 ms avec avance de 1/8, trajectoire de référence définie par $t_a = 4ms$ et $\alpha = 30$. Nous visualisons les résultats sur la *figure 28*.

Nous vérifions que le traitement des pavés améliore les performances de l'algorithme (courbe bleue en dessous de la courbe rouge). D'une manière générale le rapport résiduel sur bruit de la partie harmonique est bien meilleur que celui de la partie percussive. Ceci vient du fait que la partie harmonique possède plus d'énergie que la partie percussive dans le mixage global du morceau de musique (pour les 2 RSR l'énergie du résiduel est la même, ils suivent donc les mêmes variations).

On remarque qu'il n'existe pas de seuil optimal pour l'ensemble des séparations (de l'extrait 1 à l'extrait 6 le maximum de RSR est obtenu pour un seuil de 30, 15, 0, 10, 25, 5). En écoutant les résultats de séparations on se rend compte que le critère de RSR ne correspond pas forcement à un critère d'écoute. Le seuil optimal donné par le RSR ne serait sûrement pas le même que celui donné par un test perceptif. Réaliser ce test sur un ensemble de sujets pourrait être intéressant pour évaluer qualitativement la pertinence de notre séparation [13].

2.3 Comparaison avec d'autres algorithmes de séparation de partie percussive

Nous comparons ici les mesures de RSR avec deux algorithmes présentés dans l'état de l'art en utilisant les paramètres détaillés dans les articles :

– Algorithme *Complementary Diffusion Spectrogram* (présenté partie II.2.2.)
Paramètres : fenêtre de hanning de 63ms avec avance de 1/2. $\alpha = 0.3$, $\gamma = 0.3$, 50 itérations.

13. Je trouve personnellement qu'un seuil de 30 en moyenne donne les meilleurs résultats d'écoute pour la partie harmonique.

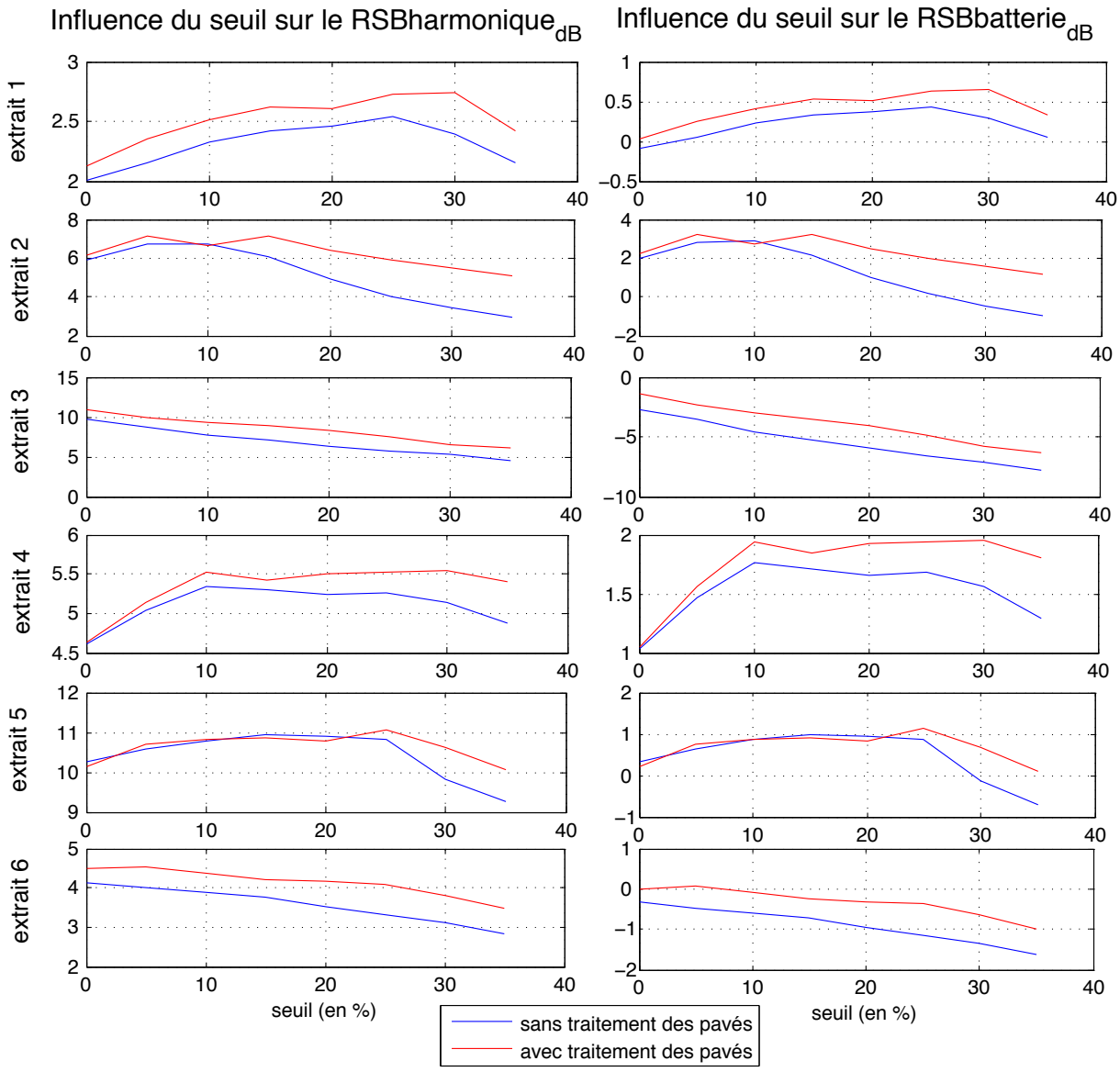


FIGURE 28 – Influence de la valeur du seuil fixé sur la distance et du traitement des pavés de bins sur le rapport signal sur résiduel pour la partie harmonique et la partie percussive.

– Algorithme *NMF+SVM* (présenté partie II.1.3.2.)
Paramètres : fenêtre racine carrée de hanning de 40ms avec avance de 1/2. Décomposition en 20 sources, classification à l'oreille.

Nous visualisons *figure 29* pour l'ensemble des morceaux du corpus la valeur des RSR harmoniques et percussifs en dB obtenus par application des trois algorithmes. L'algorithme CDS présente les résultats les plus faibles sur l'ensemble des morceaux. Pour la partie harmonique notre algorithme possède des résultats comparables à ceux obtenus par NMF+SVM. Pour la partie percussive les résultats sont moins bons.

Cependant, comme nous l'avons mentionné dans la partie précédente, la mesure du RSR ne semble pas tout à fait objective pour la mesure de la qualité de séparation. En effet, la NMF malgré ces bons résultats de RSR présente une qualité de son assez

mauvaise. Ceci est lié à la factorisation en 20 sources du spectrogramme initial qui dégrade fortement le signal. Nous présentons *figure 30* la mesure des Rapport Signal sur Bruit de la chaîne de traitement globale sur le signal original. Celui-ci est en moyenne de l'ordre de 12dB alors que pour les deux autres algorithmes comparés il est supérieur à 300dB. En effet, pour notre algorithme et l'algorithme CDS les deux parties séparées résultent de l'application de deux masques binaires complémentaires sur le spectrogramme. Le seul bruit de la chaîne vient du calcul de la FFT et de son inversion, celui-ci est donc extrêmement faible.

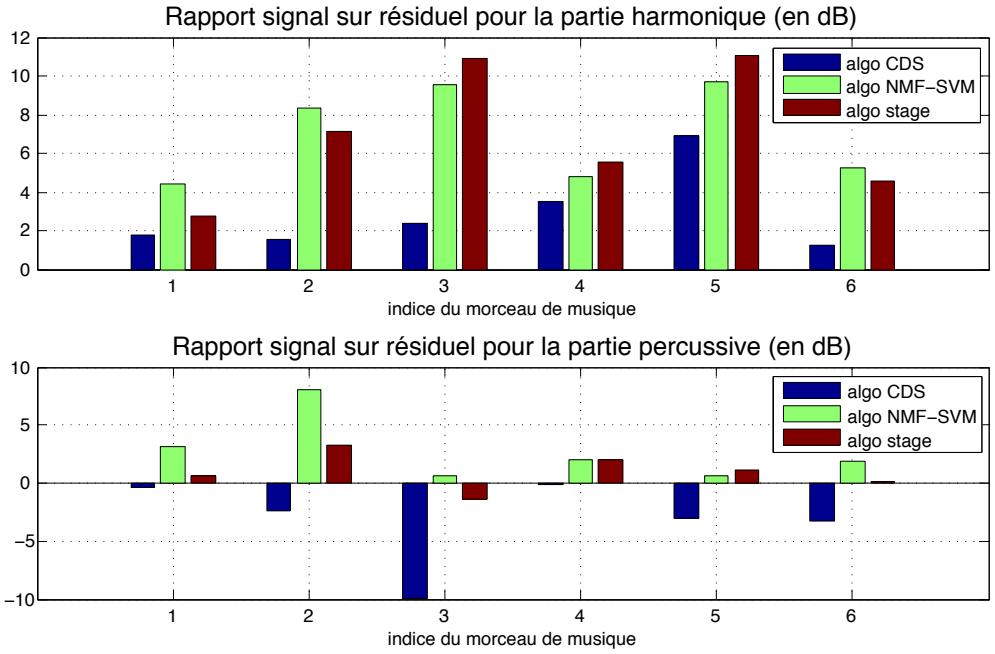


FIGURE 29 – Comparaison des rapports signal sur résiduel pour la partie harmonique et la partie percussives de différents algorithmes de séparation de partie percussive testés sur les 6 extraits de musique du corpus.

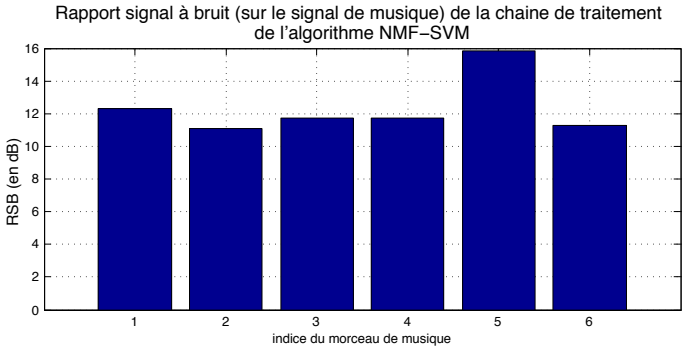


FIGURE 30 – Rapport signal sur bruit (en dB) du signal musical après reconstruction, mesuré sur les différent morceaux du corpus pour l'algorithme NMF-SVM

Pour tenter de faire le lien avec une étude perceptive il pourrait être intéressant de calculer les RSR dans différentes bandes de fréquence. Il est étonnant que l'algorithme NMF-SVM donne de si bons résultats de RSR pour la partie harmonique alors que celui-ci effectue une moyenne sur toutes les modulations de fréquence (vibratos de

voix, ...). La majorité de l'énergie étant répartie dans les basses fréquences, il est possible qu'une erreur sur un vibrato à 3 kHz (par exemple, centrée sur la zone de plus grande sensibilité de l'oreille) soit fortement perçue à l'écoute mais qu'elle joue faiblement dans le calcul du RSR.

3 Evaluations plus pertinentes de séparation de sources (futur travail)

Il serait intéressant d'effectuer des mesures de performances plus pertinentes pour comparer ces algorithmes. Pour cela, de nombreux articles traitant la mesure de performance de séparation de sources sont disponibles, ainsi qu'une toolbox $MATLAB^{TM}$ (BSS Eval). Ceci sera testé durant le mois de prolongation de stage.

Cinquième partie

Conclusion et perspectives

Cette nouvelle approche de travail basée sur l'idée de trajectoire percussive moyenne a montré des résultats intéressants pour la séparation de piste percussive. Elle est encore en cours de développement pour prendre en compte l'ensemble des éléments de la batterie et améliorer la robustesse du modèle discriminatif entre percussions et instruments harmoniques.

Une combinaison avec l'approche *match and adapt* qui travaille sur les spectres de trames permettrait de prendre en compte les deux directions (temporelle et fréquentielle) du spectrogramme. On pourrait ainsi imaginer inclure un modèle d'enveloppe spectrale pour chaque élément percussif. Ceci se rapprocherait de l'idée de séparation par NMF (modèle de spectre multiplié par un modèle de trajectoire) mais avec l'avantage de ne pas endommager le spectrogramme en contraignant sa factorisation en un ensemble de sources stationnaires (problématique pour la partie harmonique, probablement moins pour la partie percussive).

Enfin, une évaluation plus significative de la séparation de partie percussive sera menée. On essayera d'introduire des mesures quantitatives traduisant les hypothèses effectuées pour la séparation, et des jugements plus qualitatifs avec par exemple des tests perceptifs menés sur un ensemble sujets.

Je tiens à remercier mes encadrants de stage pour m'avoir permis de travailler sur ce sujet et pour leur disponibilité.

A Schéma de principe de l'algorithme 1

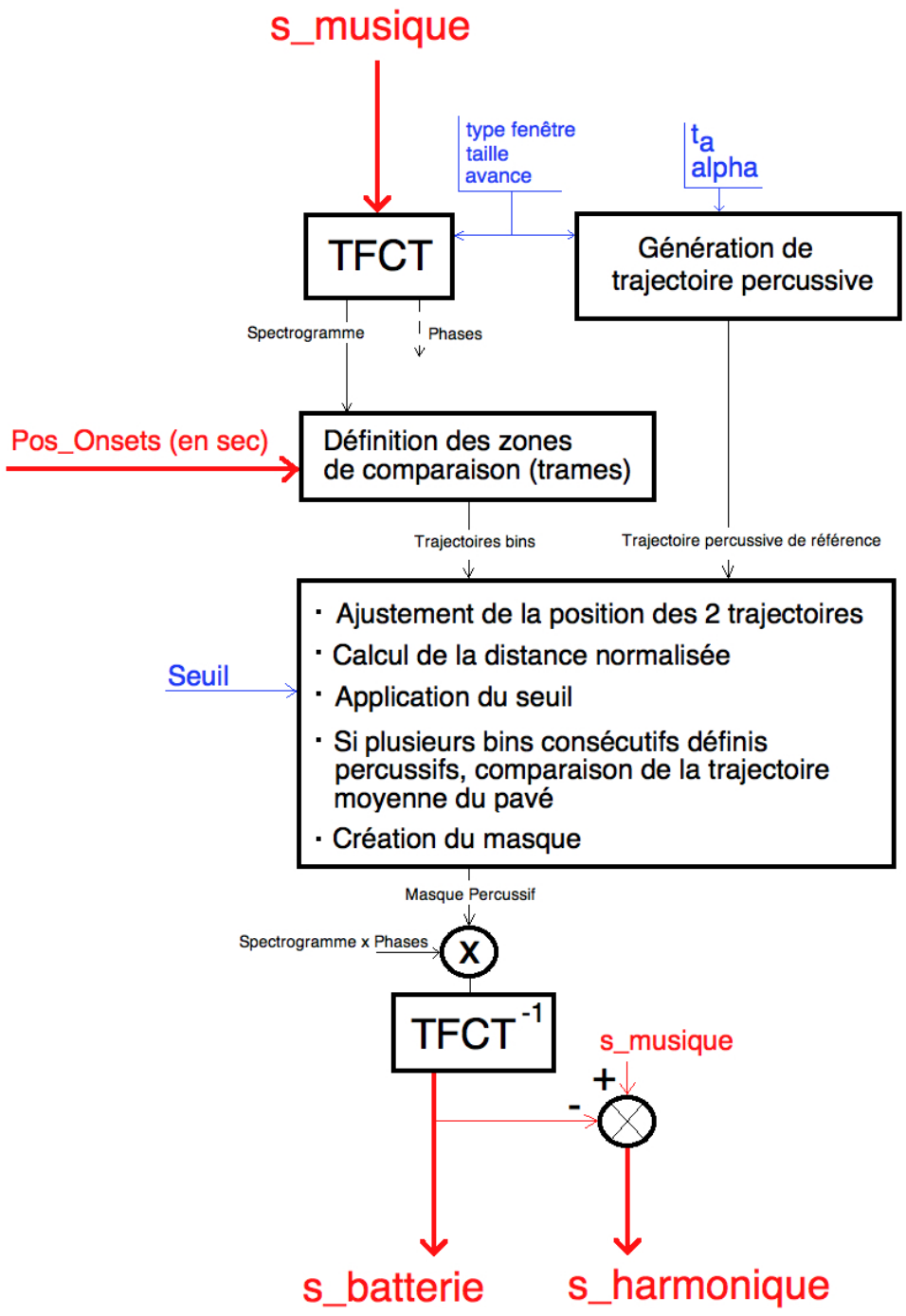


FIGURE 31 – Schéma de principe de l'algorithme d'extraction de partie percussive détaillé partie III.3.

B Schéma de principe de l'algorithme 2

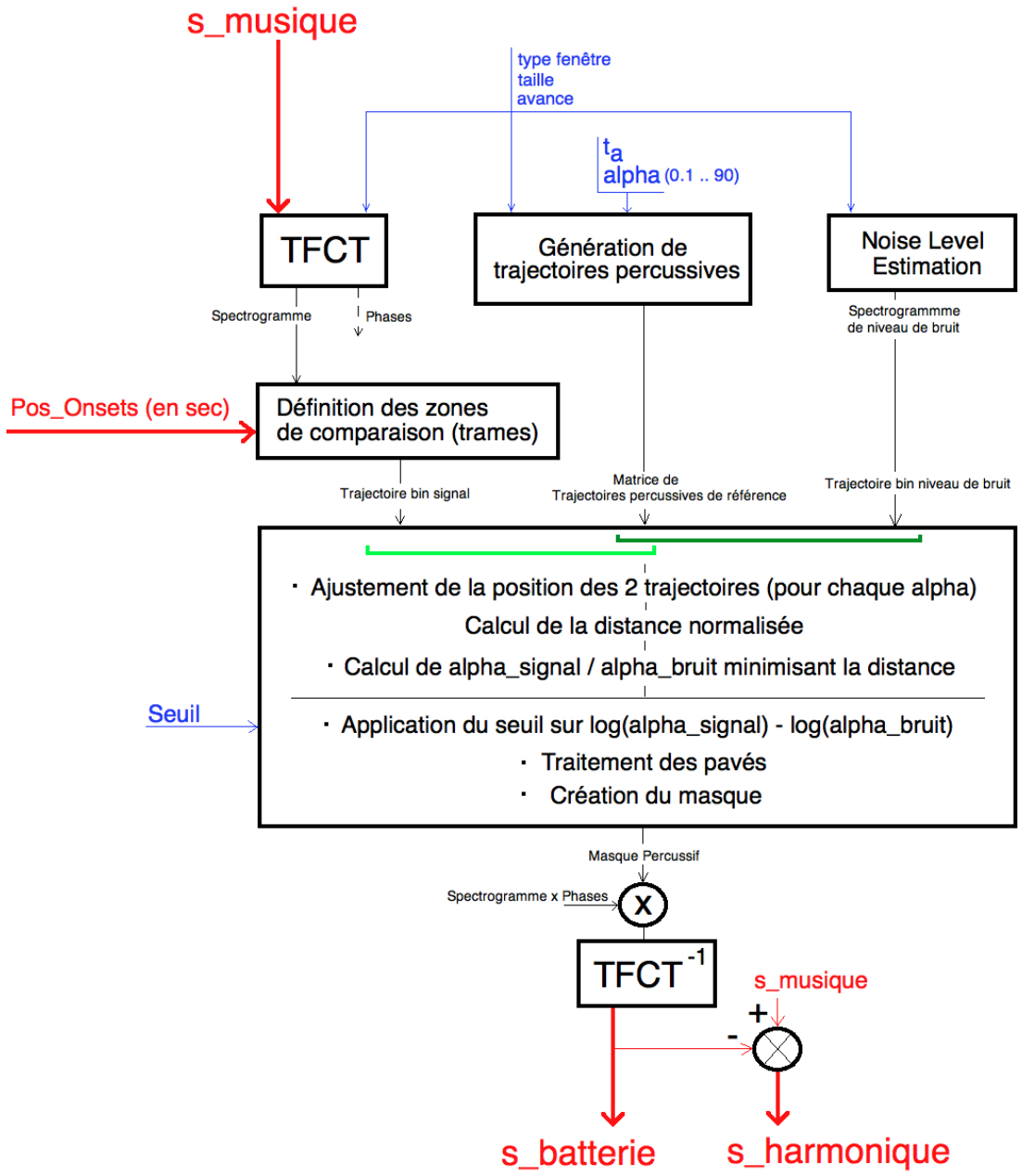


FIGURE 32 – Schéma de principe de l'algorithme d'extraction de partie percussive détaillé partie III.4.

Bibliographie

[1] Chunghsin Yeh, Axel Röbel, and Xavier Rodet. Multiple fundamental frequency estimation of polyphonic music signals. *ICASSP*, 2005.

[2] Axel Röbel. A new approach to transient processing in the phase vocoder. *Digital Audio Effects*, 2003.

[3] Axel Röbel. Onset detection in polyphonic signals by means of transient peak classification. *ISMIR*, 2005.

[4] Olivier Gillet and Gael Richard. Transcription and separation of drum signals from polyphonic music. *Transactions on Audio, Speech and Language Processing*, 2008.

[5] Thomas D. Rossing. Acoustics of percussion instruments : Recent progress. *Acoust. Sci. and Tech.*, 2001.

[6] Christian Uhle, Christian Dittmar, and Thomas Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. *4th Internation Symposium on Independent Component Analysis and Blind Signal Separation*, 2003.

[7] Dan Barry, Derry Fitzgerald, Eugene Coyle, and Bob Lawlor. Drum source separation using percussive feature detection and spectral modulation. *ISSC*, 2005.

[8] Nobutaka Ono, Kenichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka, and Shigeki Sagayama. Separation of a monaural audio signal into harmonic-percussive components by complementary diffusion on spectrogram. *EUSIPCO*, 2008.

[9] Olivier Gillet and Gael Richard. Enst-drums : an extensive audio-visual database for drum signals processinge. *In Proceeding of the 7th International Symposium on Music Information Retrieval ISMIR*, 2006.

[10] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database : Popular, classical, and jazz music databases. *International Conference on Music Information Retrieval ISMIR*, 2002.

[11] Jouni Paulus. *Signal processing method for drum transcription and music structure analysis*. PhD thesis, 2009.

[12] Derry FitzGerald, Bob Lawlor, and Eugene Coyle. Prior subspace analysis for drum transcription. *Audio Engineering Society 114th convention*, 2003.

[13] Derry FitzGerald, Eugene Coyle, and Bob Lawlor. Sub-band independent subspace analysis for drum transcription. *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, 2002.

[14] Marko Helén and Tuomas Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. *European Signal Processing Conference*, 2005.

[15] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances In Neural Information Processing Systems*, 2001.

[16] Jouni Paulus and Tuomas Virtanen. Drum transcription with non-negative spectrogram. *European Signal Processing Conference*, 2005.

[17] Arnaud Dessein. Incremental multi-source recognition with non-negative matrix factorization. Master's thesis, ATIAM, 2009.

[18] Paris Smaragdis. Non-negative matrix factor deconvolution ; extracation of multiple sound sources from monophonic inputs. *International Congress on Independent Component Analysis and Blind Signal Separation*, 2004.

[19] Roland Badeau, Rémy Boyer, and Bertrand David. Eds parametric modeling and tracking of audio signals. *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, 2002.

[20] Bertrand David, Gael Richard, and Roland Badeau. An eds modelling tool for tracking and modifying musical signals. *Proceedings of the Stockholm Music Acoustics Conference*, 2003.

[21] Olivier Gillet and Gael Richard. Extraction and remixing of drum tracks from polyphonic music signals. *Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[22] Aymeric Zils, François Pachet, Olivier Delerue, and Fabien Gouyon. Automatic extraction of drum tracks from polyphonic music signals. *International Conference on Web Delivering of Music*, 2002.

[23] Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G. Okuno. Automatic drum sound description for real-world music using template adaptation and matching methods. *ISMIR*, 2004.

[24] Andreas Wagner. Analysis of drumbeats - interaction between drummer, drumstick and instrument. Master's thesis, 2006.

[25] *Acoustique des instruments de musique*, chapter 14. 2008.

[26]

[27] Derry FitzGerald. *Automatic drum transcription and source separation.* PhD thesis, 2004.

[28] James J. Clark. Advanced programming techniques for modular synthetizers. Technical report.

[29] Eric Battenberg. Improvements to percussive component extraction using non-negative matrix factorization and support vector machines. 2009.

[30] Olivier Gillet and Gael Richard. Drum track transcription of polyphonic music using noise subspace projection. *Internation Conference on Music Information Retrieval ISMIR*, 2005.

[31] Olivier Gillet and Gael Richard. Automatic transcription of drum loops. *International Conference on Acoustics, Speech, and Signal Processing*, 2004.