MUSAÏCING MULTI-SOURCES PAR FACTORISATION DE MATRICES NON-NÉGATIVES

HADRIEN FOROUGHMAND

Sous la supervision de : Geoffroy Peeters

Rapport de Stage

Master 2 ATIAM
IRCAM, UPMC, Telecom Paristech

Février 2017 - Juillet 2017

"And the invention of transformations of certain figures has become the most important in musical composition.".

— Karlheinz Stockhausen

"Prendre des p'tits bouts d'trucs et puis les assembler ensemble".

— Stupeflip

RÉSUMÉ

Dans ce rapport de stage, nous décrivons la nouvelle méthode de synthèse dite *concaténative* que nous avons développée. Une synthèse *concaténative* vise à produire un son *cible* (par exemple de la parole, un son environnemental ou un morceau de musique) par concaténation temporelle de courts extraits audio (samples) issus d'une collection de sons *sources*. Lorsque le son *cible* est de la musique, cette méthode est souvent appelée musaïcing (mosaïque audio). A chaque instant, l'extrait audio choisi pour la concaténation doit posséder des propriétés acoustiques similaires au morceau *cible*. Ceci de manière à reproduire au mieux l'évolution temporelle des propriétés acoustiques du morceau *cible*. Ceci est habituellement effectué en recherchant des extraits dont les descripteurs audio ont des valeurs similaires à ceux du morceau *cible*. L'essentiel des méthodes développées jusqu'à présent utilisent un échantillon unique à chaque instant, il s'agit de musaïcing mono-source. Notre objectif est de développer un algorithme de musaïcing multi-sources. Notre méthode s'inscrit dans la continuité de celles de Driedger [8] et Burred [3] par l'utilisation de méthodes de factorisation en matrices non-négatives (NMF). Nous étendons ces travaux de plusieurs manières : utilisation de la transformée à Q-constant, algorithme itératif de reconstruction de la phase pour une transformée à Q-constant, extension de la NMF à la déconvolution temporelle (NMF-D), à la déconvolution temps-fréquence (NMF-2D), utilisation d'un algorithme de séparation harmonique/percussif et décomposition sur des bases respectives. Notre algorithme procède en deux étapes : 1) création de la collection de *bases sources* par application de l'algorithme NMF/NMF-D ou NMF-2D sur une collection de morceaux de musique, 2) apprentissage des meilleures activations temporelles afin de reconstruire le morceau cible à partir de ces *bases sources*. Comme dans Driedger, des contraintes sont appliquées à l'algorithme afin de conserver non-seulement le timbre des extraits choisis mais aussi d'obtenir finalement une reproduction fidèle permettant d'identifier le morceau *cible*. Afin de permettre une comparaison des différentes variantes de l'algorithme (NMF, NMF-D, NMF-2D) nous avons effectué un test perceptif en ligne dont nous discutons ici les résultats.

ABSTRACT

This report describes a new synthesis method called *concatenative* that we have developped. A *concatenative* synthesis aims to product a *target* sound (like speech, environmental sound or an audio track) by temporal concatenation of shorts audio samples from a collection of *source* sounds. When target sound is music, we often called this method musaicing (audio mosaicing). At every moment, the chosen audio exerpt for concatenation must have similar acoustic properties like that of the target track. this in order to reproduce as best as possible the temporal evolution of acoustic propreties of the target sound. Most of methods developped until now use a single sample at every moment, this is mono-source musaicing. Our goal is to develop a multi-source musaicing algorithm. Our method is in line with that of Driedger [8] and Burred [3] by the use of non-negative matrix factorisation methods (NMF). We expand those works in many ways : using a constant-Q transform, iterative algorithm of phase reconstruction for a constant-Q transform, extension of NMF to temporal deconvolution (NMF-D), to time/frequency deconvolution (NMF-2D), using of an harmonic/percussive source separation algorithm on respective basis. Our algorithm proceeds in two steps : 1) establishing of the *source basis* collection by applying the NMF/NMF-D or NMF-2D algorithm on a audio track collection, 2) learning of the best temporal activations in order to rebuild the target track from those *source basis*. In order to allow a comparison of the different variants of the algorithm (NMF, NMF-D, NMF-2D) we carried out an online perceptual test of wich we discuss here the results.

REMERCIEMENTS

Je tiens tout d'abord à remercier particulièrement mon encadrant Geoffroy Peeters pour son accueil mais aussi pour sa disponibilité, son soutient et sa confiance ainsi que pour ses précieux conseils tout au long de ce stage.

Je tient également à remercier tous les membres de l'équipe analyse/synthèse, Axel Roebel pour son accueil dans l'équipe, Nicolas Obin pour son aide précieuse lors de ma candidature au Master ATIAM mais aussi les doctorants Alice, Gabriel et Guillaume pour leur bonne humeur et leurs conseils avisés. Je remercie Luc Ardaillon pour son aide et son expertise sur le questionnaire perceptif.

Merci à mes collègues de bureau, Dominique et Rémi pour avoir répondu à mes interrogations sur le domaine tout au long de ces 5 mois et merci à Frédéric Cornu qui m'a généreusement apporté son aide afin d'optimiser certains programmes.

Je souhaite également remercier tout mes camarades du Master ATIAM, ceux présents à l'IRCAM, Lou, Andréa, Tristan, Pierre, Eugénia, Victor et Marc pour nos discussions passionnantes du midi et les autres pour leur présence tout au long de cette riche année et un merci particulier à Mathieu pour sa présence joyeuse et son amitié sans faille tout au long de nos études supérieures.

Un grand merci à tous mes amis pour leur jovialité permanente et à Adrien qui par sa patience, sa sérénité et son humour m'a permis de garder confiance.

Je remercie particulièrement ma mère pour son soutien sans faille et ma sœur Roxane pour sa curiosité communicative. Enfin je remercie sincèrement Adèle qui par son amour indéfectible, sa douceur et son soutien permanent m'a permis de toujours continuer à avancer.

TABLE DES MATIÈRES

1 INTRODUCTION 1
1.1 Présentation globale du projet . . . 1
1.1.1 Présentation du domaine du "Music Information Retrieval" . 1
1.1.2 Contexte . . . 1
1.2 Etat de l'art . . . 2
1.2.1 Synthèse concaténative en temps réel avec CataRT . . . 3
1.2.2 Musaïcing par NMF de Driedger . . . 4
1.2.3 Limitations de l'approche du musaïcing de Driedger et motivations pour le développement d'une nouvelle méthode . . 7
1.2.4 Synthèse croisée par NMF de Burred . . . 8
1.3 Plan . . . 9
2 REPRÉSENTATION EN ENTRÉE DE LA NMF ET RECONSTRUCTION DU SIGNAL AUDIO 11
2.1 Transformée de Fourier à Court Terme . . . 11
2.2 Algorithme de Griffin & Lim, synthèse du module de la TFCT avec estimation de la phase du signal . . . 11
2.3 Transformée à Q-constant . . . 12
2.4 Adaptation de l'algorithme de Griffin & Lim à partir du module de la CQT . . . 14
3 MUSAÏCING PAR FACTORISATION EN MATRICES NON-NÉGATIVES DÉCONVOLUTIVE 16
3.1 La NMFD . . . 16
3.2 Musaïcing par NMFD . . . 17
3.3 Musaïcing par NMFD contrainte . . . 19
3.3.1 Contrainte sur la répétition successive des bases W^{τ} . . . 19
3.3.2 Contrainte sur la superposition des bases relatives à chaque enregistrement source . . . 20
3.4 Limitation de la méthode . . . 20
4 MUSAÏCING PAR FACTORISATION EN MATRICES NON-NÉGATIVES 2D DÉCONVOLUTIVES 22
4.1 La NMF2D . . . 22
4.2 Musaïcing par NMF2D . . . 23
4.2.1 Illustration . . . 25
4.3 Musaïcing par NMF2D contrainte . . . 25
4.3.1 Contrainte sur la répétition successive des bases W^{τ} . . . 26
4.3.2 Contrainte sur la superposition des bases relatives à chaque enregistrement source . . . 26
4.3.3 Illustration . . . 26
5 MUSAÏCING POST-SÉPARATION HARMONIQUE/PERCUSSIVE 28
5.1 Décomposition HPSS de Filtzgerald . . . 28
5.1.1 Illustration . . . 29
5.2 Algorithme de musaïcing post HPSS . . . 29
6 RÉSULTATS 32
6.1 Implémentation des méthodes . . . 32
6.1.1 Musaïcing de Driedger . . . 32

6.1.2 Musaïcing par NMFD 32
6.1.3 Musaïcing par NMF2D 33
6.1.4 Musaïcing post HPSS 33
6.2 Tests d'écoute 33
6.2.1 Matériaux sonores 34
6.2.2 Questionnaire 34
6.2.3 Musaicïng post-HPSS 35
6.3 Résultats 35
6.3.1 Participation 35
6.3.2 Analyse des résultats 35
7 CONCLUSION 41
7.1 Discussion 41
7.2 Perspectives 42

BIBLIOGRAPHIE 43

TABLE DES FIGURES

FIGURE 1 "Van Gogh Portrait", Mosaïcing de Robert Silver. 2
FIGURE 2 Schéma global du musaïcing de Driedger *tel que publié dans [8]*. 4
FIGURE 3 Schéma simplifié de la factorisation en matrices non-négatives 5
FIGURE 4 Schéma simplifié des différentes contraintes appliquées à la matrice d'activations lors de son apprentissage par NMF *tel que publié dans [8]*. 7
FIGURE 5 Schéma explicatif de la méthode de synthèse croisée de Bured. 9
FIGURE 6 Algorithme de Griffin & Lim. 13
FIGURE 7 Comparatif entre deux représentations temps/fréquence du son. 14
FIGURE 8 Algorithme de Griffin & Lim adapté au module de la CQT. 15
FIGURE 9 Modèle de factorisation par NMFD *tel que publié dans [21]*. . 17
FIGURE 10 Schéma explicatif du musaicing par NMFD. 19
FIGURE 11 Exemple de musaicing par NMFD. 20
FIGURE 12 Exemple de musaïcing par NMFD contrainte. 21
FIGURE 13 Modèle de factorisation par NMF2D *tel que publié dans [16]*. . 23
FIGURE 14 Schéma explicatif du musaïcing par NMF2D. 24
FIGURE 15 Exemple de musaïcing par NMF2D. 25
FIGURE 16 Exemple de musaïcing par NMF2D contrainte. 27
FIGURE 17 HPSS de Filtgerald. 30
FIGURE 18 Musaicing mixte post-HPSS. 31
FIGURE 19 Graphique des réponses à la question 1 : Comment jugez-vous la qualité du son ? 37
FIGURE 20 Graphique des réponses à la question 2 : La structure temporelle et harmonique du son cible est-elle reconnaissable ? 38
FIGURE 21 Graphique des réponses à la question 3 : Le(s) son(s) source(s) sont ils identifiables(s) ? 39
FIGURE 22 Graphique des réponses à la question 4 : Comment jugez vous l'intérêt créatif de l'exemple sonore ? 40
FIGURE 23 Schéma comparatif entre les méthodes de l'état de l'art et les nôtres. 42

ACRONYMES

KL Divergence de Kullback-Leibler

NMF Factorisation de matrices non-négatives - Non-negative Matrix Factorisation

NMFD Déconvolution de facteurs de matrices non-négatives - Non-negative Matrix Factor Deconvolution

NMF2D Déconvolution de facteurs de matrices non-négatives 2D - 2D Non-negative Matrix Factor Deconvolution

TFCT Transformée de Fourier à Court Terme

FFT Transformée de Fourier Rapide - Fast Fourier Transform

CQT Transformée à Q-Constant - Constant-Q Transform

HPSS Séparation de Sources Harmonique/Percussive - Harmonic/Percussive Source Separation

MFCC Mel-Frequency Cepstral Coeficients

GL Algorithme de re-synthèse du signal et d'estimation de sa phase de Griffin & Lim à partir de la TFCT d'amplitude

1 INTRODUCTION

Cette partie est consacrée à l'introduction des différents concepts entourant le projet de stage. Pour cela nous posons dans un premier temps le contexte global du stage. Nous commençons par une brève présentation du domaine du "Musical Information Retrieval" (MIR) puis nous enchainons sur les motivations principales qui ont engendrées les recherches effectuées au cours du stage. Dans un second temps nous parlons des différentes techniques de musaïcing qui ont pu être développées par différents chercheurs et nous nous concentrons sur la méthode de Driedger sur laquelle nous nous sommes focalisé. Nous concluons cette introduction par le plan de ce rapport.

1.1 PRÉSENTATION GLOBALE DU PROJET

1.1.1 *Présentation du domaine du "Music Information Retrieval"*

Ce stage s'est déroulé au sein de l'équipe analyse/synthèse de l'IRCAM, plus précisément dans la sous-équipe MIR. Les objectifs de cette équipe dirigée par Geoffroy Peeters sont orientés vers la recherche autour du "Music Information Retrieval" (littéralement "Récupération d'Informations Musicales").

Comme son nom l'indique, le MIR vise à extraire des informations à partir de la musique (un morceau ou bien une base de donnée musicale pouvant compter plusieurs milliers de morceaux). L'extraction se fait principalement via des algorithmes de traitement du signal audio mais on peut aussi relier d'autres domaines comme la musicologie et l'apprentissage machine ou encore des combinaisons de plusieurs de ces disciplines.

Une fois ces informations quantifiables obtenues, les applications les utilisant sont multiples. Le traitement des données peut aussi se faire grâce à des méthodes d'apprentissage machine. On peut ainsi citer les algorithmes de recommandation musicale, de transcription automatique, de classification par genre ou par époque ou encore de génération et de synthèse de nouveaux sons. Notre projet fait partie intégrante de cette dernière catégorie à l'aspect plus créatif.

1.1.2 *Contexte*

1.1.2.1 *Motivation*

Dans la musique contemporaine, avec le développement de la musique électronique entre autre, il est commun pour les artistes de composer leurs morceaux à partir de courts extraits musicaux existants. Ainsi des artistes tel que John Oswald (particulièrement dans l'album « Plexure »), concatènent manuellement des sample issues de différentes œuvres afin de donner lieu à un nouveau type de création musicale. Cette technique peut cependant être limitée par le fait que la sélection se fait à partir d'un très grand nombre de morceaux. Cette sélection est ainsi une tâche particulièrement laborieuse et ce malgré l'aide apportée par les logiciels d'aide à la composition.

1.1.2.2 *Techniques existantes*

Cette contrainte a donc mené les chercheurs en traitement du signal audio à développer des techniques de concaténation automatique d'extraits musicaux à partir d'une collection de morceaux. Le « musical mosaïcing » (« musaïcing », [25]) désigne ces techniques et, comme son nom l'indique, est directement inspiré des travaux dans le domaine du traitement de l'image. Le terme de « mosaïcing » désigne la construction d'une image à partir d'une multitude d'autres images. L'artiste Robert Silver [20] est un pionnier de cette méthode comme le montre l'œuvre 1 [1].

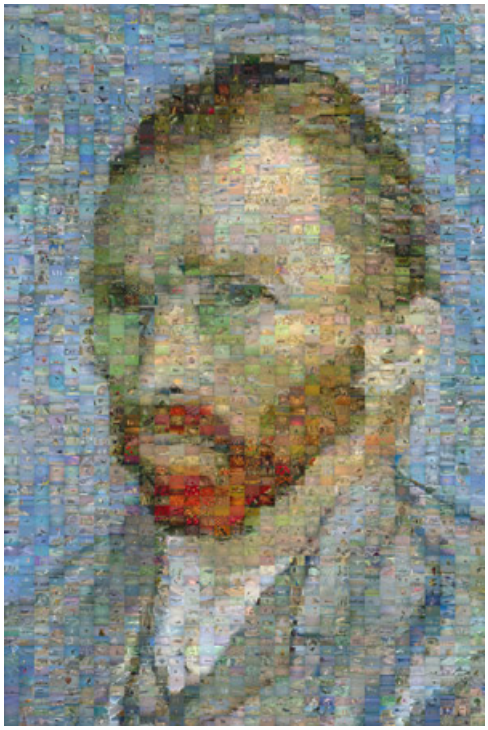

FIGURE 1 – "Van Gogh Portrait", Mosaïcing de Robert Silver.

Les différents travaux réalisés dans cette optique sont variés de part leurs applications précises et les techniques utilisées [4, 5, 8, 11, 18, 19, 25]. Ce stage s'inscrit donc dans le contexte du musaïcing. Ici, l'objectif est de reproduire un morceau « cible » par concaténation temporelle de courts extraits sonores issus d'une collection de morceaux. Ces morceaux « sources » sont automatiquement sélectionnés pour leurs propriétés acoustiques similaires à celles du morceau « cible » à chaque instant.

Driedger et al. 2015 [8] propose un musaïcing utilisant la factorisation en matrices non-négatives contraintes afin de reproduire un morceau cible à partir de bruits ambiants constituant la collection de sources. C'est cette méthode que nous étudierons et étendrons pendant ce stage.

1.2 ETAT DE L'ART

Dans cette partie nous présentons les différents travaux apparentés au travail que nous avons réalisé durant ce stage. Nous commençons par présenter la syn-

1. Oeuvre de Robert Silver que l'on retrouve sur son site internet : http://www.photomosaic.com/portfolio.html.

thèse concaténative CataRT de Schwarz & al. [19] qui constitue la référence en matière de synthèse concaténative par descripteurs audio. Nous présentons ensuite les deux méthodes existantes de synthèse concaténative utilisant la décompositon en matrices non-négatives : celle de Driedger [8] et celle de Burred [3] (bien que celle-ci soit plutôt une méthode de synthèse croisée que concaténative).

1.2.1 *Synthèse concaténative en temps réel avec CataRT*

La méthode de synthèse concaténative de Schwarz & al. est décrite dans l'article [19]. L'article présente le système CataRT permettant de réaliser différentes applications interactives basées sur la synthèse concaténative.

La méthode est appliquée sur une grande base de donnée d'extraits sonores segmentée en "unités". L'objectif vise à re-synthétiser un morceau (ou phrase) cible à partir de ce corpus d'extraits sonores. Pour se faire, des descripteurs audio sont extraits des unités sources ce qui permet d'obtenir des caractéristiques afin d'opérer à une sélection précise. Les unités ainsi sélectionnées sont ensuite adaptées afin de correspondre au mieux aux caractéristiques du morceau cible pour finalement les concaténer. Les applications résultantes de cette méthode sont nombreuses. La démarche permet une synthèse sonore dite granulaire applicable à des instruments, de la voix ou encore à des sons ambiants. On peut alors parler d'une démarche artistique visant la création sonore en temps réel.
On note donc une analogie entre ces travaux et le musaïcing (ce dernier est par ailleurs cité au sein de l'article en tant qu'application du système). La différence réside dans les outils et le modèle mis en place pour la méthode.

1. Tout d'abord une phase d'analyse est effectuée. Elle permet d'extraire un ensemble de descripteurs du corpus source. Ces descripteurs sont détaillés dans la thèse de Schwarz [18]. Une unité est alors caractérisée par une matrice de scalaire représentant l'ensemble des valeurs de descripteurs de l'unité au cours du temps (comme la moyenne, la variance, le minimum, le maximum,...) mais aussi des méta-données telles que la durée du morceau, son origine, son temps de départ et de fin.
2. Les données sont ainsi stockées dans une matrice dont les colonnes correspondent aux différents descripteurs et les lignes aux unités. On peut parler d'un espace de descripteur ou chaque unité est représentable par un point précis de l'espace.
3. La sélection des unités dans cet espace se fait par le calcul de la distance de Mahalanobis [6] entre un point x de l'espace et chaque unité. Afin d'affiner la selection, un algorithme de k-plus-proches-voisins [7] est employé ici.
4. Une fois les unités les plus proches sélectionnées, elles sont concaténées puis synthétisées. La synthèse granulaire est employée ici, plusieurs modifications applicables à l'unité choisie sont possibles (comme la transposition ou les modifications de l'intensité sonore).
5. Enfin une interface intuitive permet de représenter l'espace des descripteurs en deux dimensions. La souris est utilisée afin d'optimiser la sélection des unités dans cet espace. Différentes options sont de plus incorporées à l'interface afin d'augmenter la dimension créative de la synthèse. Par exemple, la sélection de certains "modes" permet de choisir l'unité la plus proche qui n'as pas encore été mise à contribution.

L'article présente pour finir différents types d'utilisations de l'application détaillés dans l'article. On peut citer la synthèse granulaire exploratoire, la synthèse contrôlée par le geste ou le signal audio (qui s'approche particulièrement du musaïcing) ou encore la synthèse de parole expressive.

Ces travaux nous permettent de souligner l'étendue créative et applicative de la synthèse concaténative dans laquelle s'inscrit notre projet autour du musaïcing.

1.2.2 *Musaïcing par NMF de Driedger*

Dans son article [8], Driedger décris un algorithme en deux étapes permettant de reconstituer un morceau cible (ex : « Let it be » - The Beatles) avec des enregistrements sources (ex : bourdonnement d'abeilles) préalablement transposées afin de couvrir l'ensemble de la gamme tempérée. Les enregistrements sources sont choisis pour leurs propriétés relativement stationnaires.

La Figure 2 présente l'architecture globale de cet algorithme.

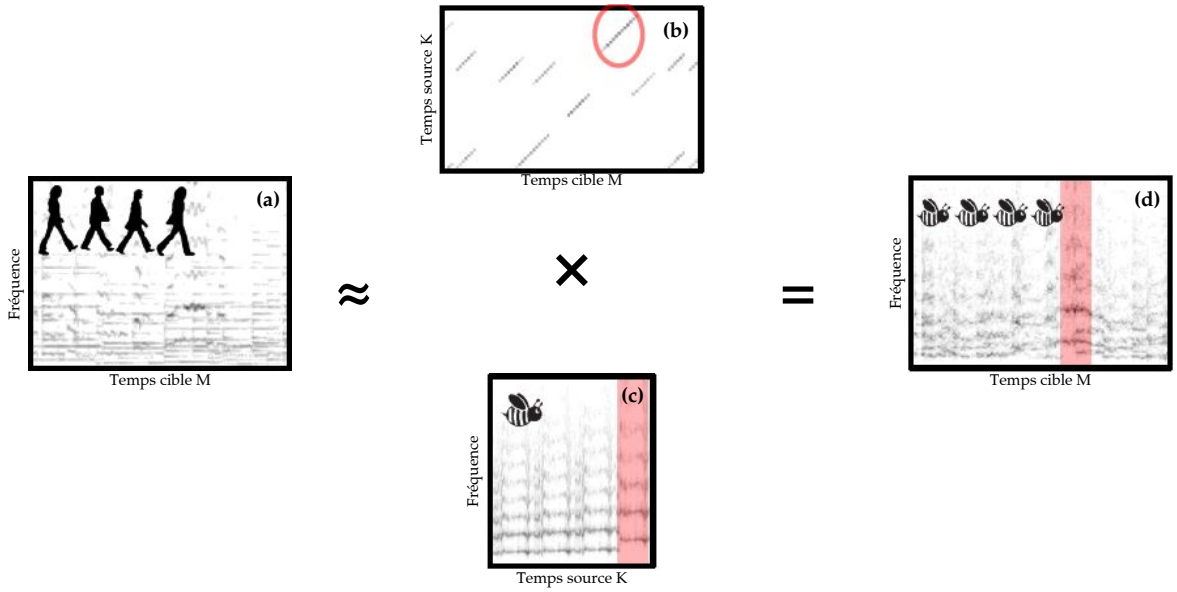

FIGURE 2 – Schéma global du musaïcing de Driedger *tel que publié dans [8]*.
(a) : Matrice Cible (ici spectrogramme de "Let it be" - The Beatles), **(b)** : Matrice d'activations apprises par NMF contrainte, **(c)** : Matrice source (ici enregistrement de bourdonnements d'abeilles transposés sur la gamme chromatique), **(d)** : Matrice de musaïcing résultante ("Let it Bee").

1.2.2.1 *Factorisation en matrices non-négatives*

La NMF est une méthode de décomposition matricielle. Elle permet d'obtenir une représentation par parties à coefficients non-négatifs d'un objet. Elle a d'abord été utilisée en traitement de l'image [13] et est devenue très populaire en traitement du signal audio pour des tâches de séparation de sources ou de transcription. Elle est dans ce cas appliquée à la matrice non-négative du spectrogramme d'amplitude.

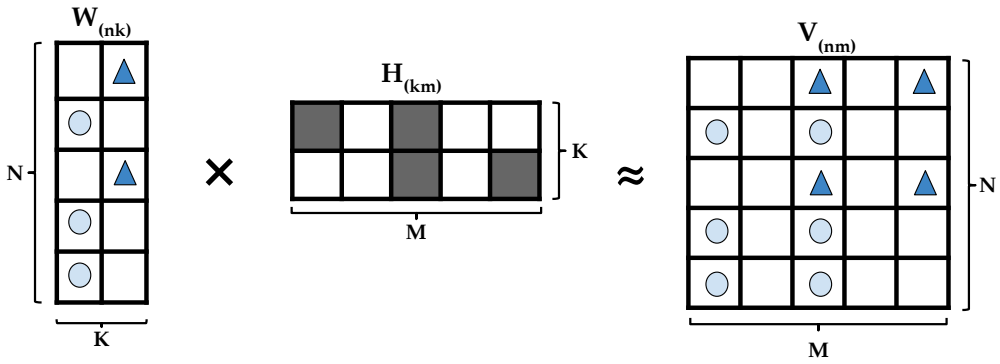

FIGURE 3 – Schéma simplifié de la factorisation en matrices non-négatives

Considérant une matrice, ici un spectrogramme d'amplitude d'un signal audio obtenue par transformée de Fourier à court terme[2], $V_{N\times M}$ aux coefficients positifs ou nuls, N représente le nombre de fréquences et M le nombre de trames temporelles[3], la NMF approxime V de la manière suivante :

$$V_{nm} \approx \hat{V}_{nm} = W_{nk}H_{km} \tag{1}$$

avec $W_{N\times K}$ la matrice non-négative de bases (les atomes spectraux) et $H_{K\times M}$ la matrice d'activations temporelles. K représente le nombre de bases (d'atomes spectraux). $N, M, K \in \mathbb{N}$.

L'approximation de V par WH est obtenue par minimisation d'une distance ou divergence (fonction de coût). Parmi les plus utilisées, on peut citer la distance euclidienne, la divergence d'Itakura-Saito ou la divergence de Kullback-Leibler (KL) que nous utiliserons ici :

$$D(V||WH) = \sum_{nm} V_{nm}\log(\frac{V_{nm}}{(WH)_{nm}}) - V_{nm} + (WH)_{nm} \tag{2}$$

1.2.2.2 *Utilisation de la NMF dans le musaïcing de Driedger*

Dans la méthode du musaïcing de Driedger seule la matrice d'activations H est estimée. La matrice de base W est fixée et est prise égale à la matrice du spectrogramme d'un signal source. Les atomes spectraux sont donc directement des trames du spectrogramme du signal source. Si nous notons x_C le signal cible (le morceau des Beatles dans notre exemple) et x_S le signal source (les bourdonnements d'abeilles transposés), on calcule le spectrogramme d'amplitude de ces deux signaux par TFCT et l'on pose $V = |X_C|$ et $W = |X_S|$[4]. Pour estimer la matrice d'activations, H est d'abord initialisée aléatoirement. Ensuite, pour un nombre d'itérations L_{it} fixé, la matrice H est apprise itérativement suivant la formule de mise à jour suivante :

$$H_{km} \leftarrow H_{km}\frac{\sum_n W_{nk}V_{nm}/(WH)_{nm}}{\sum_n W_{nk}} \tag{3}$$

2. TFCT, la méthode est décrite plus précisément au Chapitre 2
3. Ces notations seront celles utilisées tout au long de ce rapport
4. Les indices C et S désignent respectivement cible et source

Le spectrogramme d'amplitude de musaïcing reconstruit s'écrit : $|X_{mus}| = |X_S|.H$.

Dans une reconstruction classique par NMF, la phase du signal d'origine est appliquée au signal final. Cependant dans le cas du musaïcing c'est un nouveau signal qui doit être obtenu malgré la conservation de la structure temporelle du morceau d'origine. Dans le but de synthétiser le signal x_{mus}, un algorithme itératif de reconstruction de phase proposé par Griffin & Lim [10] est appliqué au spectrogramme d'amplitude $|X_{mus}|$. Cet algorithme est décrit au Chapitre 2.

1.2.2.3 *Rajout de contraintes lors de l'apprentissage NMF dans le musaïcing de Driedger*

Driedger note dans [8] que l'application telle quelle de l'algorithme NMF (tel qu'expliqué ci-dessus) conduit à un son du musaïcing non "fluide" à l'écoute. Il en décrit les trois causes principales et propose trois contraintes à appliquer lors de l'apprentissage de H afin d'en diminuer les effets.

Soit λ_{it} un coefficient permettant d'atténuer les éléments contraints de la matrice H sans les annuler. λ_{it} dépend de l'indice d'itération $l \in [1 : L_{it} - 1]$ et du nombre d'itérations tel que $\lambda_{it} = \frac{l+1}{L_{it}}$.

1. La répétition successive d'atomes spectraux (trames temporelles de la source) identiques entraine un effet de "bégaiement". Il propose de calculer une matrice de "limitation des répétitions" R d'atome spectral tel que :

$$R_{km} = \begin{cases} H_{km} & \text{si } H_{km} = \mu^r_{km} \\ H_{km}(1-\lambda_{it}) & \text{sinon} \end{cases} \tag{4}$$

   avec μ^r_{km} la valeur maximale de H dans un voisinage horizontal $[m-r; m+r]$.

2. La superposition de trop nombreuses activations sur chaque trame spectrale engendre des artéfacts d'annulation de phase. Il propose de calculer une matrice de limitation polyphonique P :

$$P_{km} = \begin{cases} R_{km} & \text{si } k \in \Omega^p_m \\ R_{km}(1-\lambda_{it}) & \text{sinon} \end{cases} \tag{5}$$

   avec p le nombre d'activations dans une colonne de la matrice d'activation (autrement dis le degré de polyphonie désiré dans le musaïcing) et où Ω^p_m contient les indices des p-plus grands éléments de la $m^{ième}$colonne de R.

3. La perte de continuité temporelle de la source engendre un rendu saccadé où les caractéristiques acoustiques deviennent difficilement reconnaissables. Il propose le calcul d'une matrice d'amélioration de la continuité C calculée comme la convolution de la matrice avec une matrice à noyau diagonal tel que :

$$C_{km} = \sum_{i=-c}^{c} P_{(k+i)(m+i)} \tag{6}$$

   avec c qui définit la longueur du noyau. On parle de « sparcité » des diagonales (de l'anglais « sparse », clairsemé) ou de parcimonie.

La formule de mise à jour de H (3) devient alors :

$$H_{km} \leftarrow C_{km} \frac{\sum_n W_{nk} V_{nm} / (WC)_{nm}}{\sum_n W_{nk}} \tag{7}$$

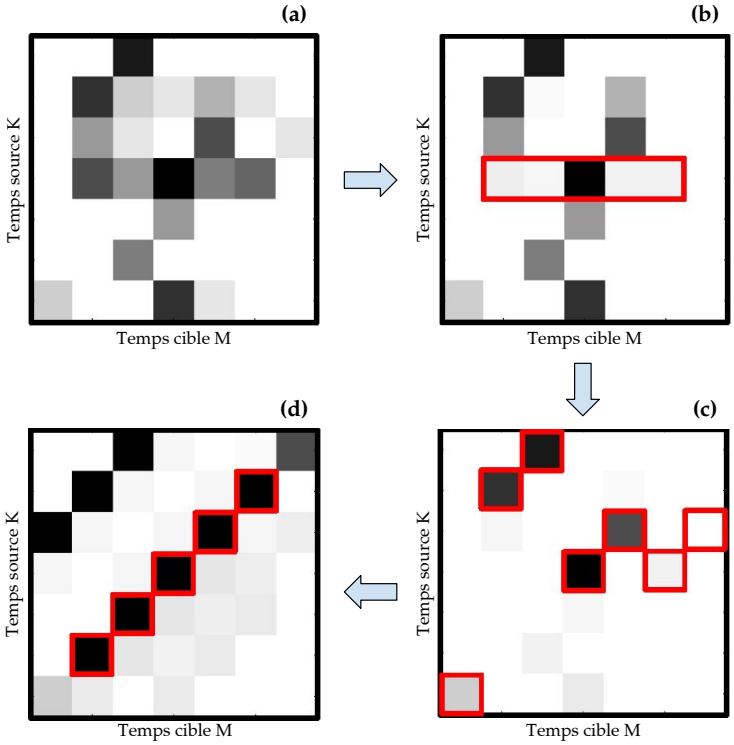

FIGURE 4 – Schéma simplifié des différentes contraintes appliquées à la matrice d'activations lors de son apprentissage par NMF *tel que publié dans [8].*
(a) : matrice d'activations H_{km} **(b)** : matrice contrainte par la limitation des répétitions successives des trames temporelles R_{km} avec $r = 2$, **(c)** : matrice contrainte par un degré de polyphonie P_{km} avec $p = 1$, **(d)** : matrice contrainte de parcimonie diagonale C_{km} avec $c = 2$.

La Figure 4 décrit les différentes contraintes appliquées à la matrice d'activation lors de son apprentissage par NMF.

1.2.3 *Limitations de l'approche du musaïcing de Driedger et motivations pour le développement d'une nouvelle méthode*

Bien qu'intéressante, la méthode de Driedger possède plusieurs limitations :

1. La re-synthèse du signal cible V_{nm} est effectuée en utilisant directement les tranches temporelles du spectrogramme d'un son source. Ces tranches sont considérées comme les bases W_{nk}. Le nombre de source est limité à une seule. Nous souhaitons étendre la méthode en permettant l'utilisation de bases correspondant à des sources séparées (donc résultant de la décomposition préalable d'un signal) et en utilisant plusieurs sources.

2. La méthode de Driedger reposant sur l'utilisation de la NMF simple ne permet pas la prise en compte du déroulement temporel des sources. De ce fait, Driedger rajoute de contraintes favorisant l'activation temporelle de base contiguës. Nous souhaitons prendre en compte le déroulement temporelle par l'utilisation de la NMF Déconvolutive à la place de la NMF simple.
3. Afin de permettre la réduction du nombre de bases nécessaires (dans la méthode de Driegler, les sources doivent être transposée afin de couvrir l'ensemble des notes de la gamme chromatique, conduisant à un nombre de bases particulièrement élevé) et d'étendre leur utilisation, nous étudierons la NMF Déconvolutive 2D (en temps et en log-fréquences).
4. Afin de garantir une invariance en fréquence de nos bases, notre méthode de musaïcing fonctionne sur la transformée à Q-constant qui s'avère indispensable pour l'utilisation de la NMF2D et ainsi pour la transposition automatique des bases lors de la reconstruction du signal de musaïcing. Nous proposons une nouvelle méthode de re-synthèse du signal audio correspondant au module de la transformée à Q-constant. Cette méthode est inspirée de l'algorithme de Griffin & Lim. La CQT ainsi que la méthode de reconstruction sont détaillés au Chapitre 2.

Ce sont ces limitations qui nous ont conduit à étendre le musaïcing de Driedger.

1.2.4 *Synthèse croisée par NMF de Burred*

Dans [3] Burred propose une méthode de synthèse croisée qui s'apparente non seulement à la méthode de Driedger mais aussi à notre méthode. Basée sur la factorisation en matrices non-négatives, la méthode de synthèse croisée de Burred consiste en un échange des bases entre une matrice cible et une matrice source. De cette façon on obtient après resynthèse (Griffin & Lim) un signal possédant la structure de la cible tout en ayant conservé le timbre de la source.

En détaillant la méthode on peut la décrire en plusieurs étapes :

1. Calcul des matrices cible X_C et source X_S par TFCT à partir des signaux respectifs (x_C et x_S) ;
2. Apprentissage de la matrice de bases W_C et W_S et d'activations H_C et H_S par NMF (le choix du nombre d'atomes K_C et K_S dépend de l'application de l'algorithme ainsi que de la nature des sons sources et cibles) ;
3. Extraction de descripteurs audio de type "Mel Frequency Cepstral Coeficient" (MFCC) [14] à partir des matrices W_C et W_S afin d'obtenir une description compacte de l'enveloppe spectrale de chacune des bases ;
4. Calcul d'une matrice de similarité entre les MFCCs calculés à l'étape précédente en vue de ne conserver que les atomes à dynamique similaire ;
5. Multiplication entre les atomes ainsi obtenus appelés W_{S+C} et la matrice d'activation des sources H_S (selon le papier, l'expérience peut être réalisée en conservant la matrice d'activation de la cible H_C) ;
6. Inversion de la matrice résultante X_{S+C} par l'algorithme de Griffin & Lim, synthèse du signal hybride.

La Figure 5 nous montre le schéma explicatif de la méthode de synthèse croisée afin de marquer les différences avec celle de Driedger et la notre.

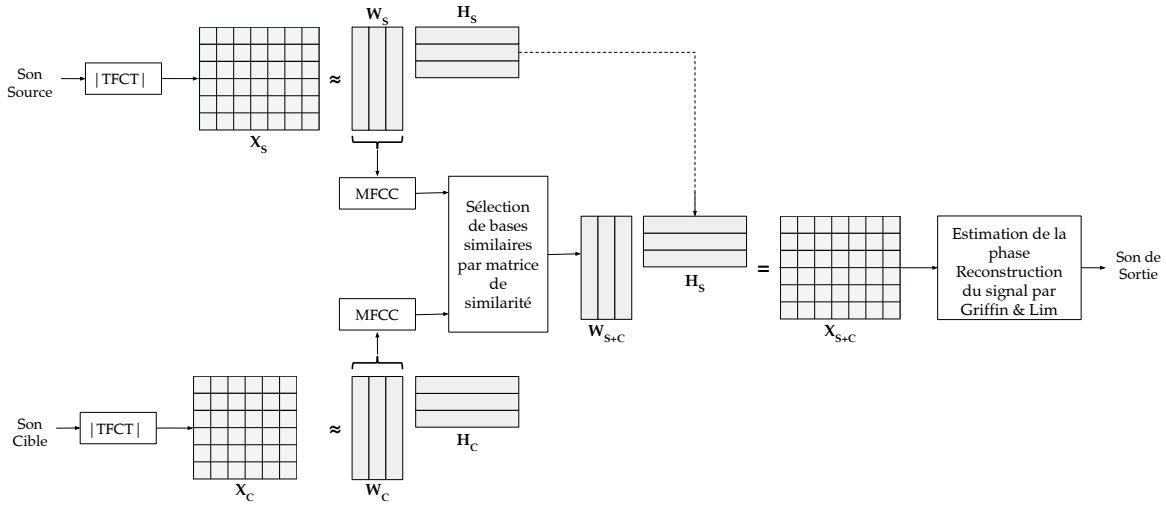

FIGURE 5 – Schéma explicatif de la méthode de synthèse croisée de Bured.

On retrouve ainsi plusieurs différences entre la méthode de Burred, celle de Driedger et la notre.

— Pour ce qui est de la méthode de Driedger nous avons pu étudier Section 2 que l'unique apprentissage par NMF était celui des activations. Le matériau d'entrée est le spectrogramme de l'enregistrement source.
— La méthode de Burred en revanche apprend les bases et activations de la cible et de la source séparément puis les mélanges. Le signal re-synthétisé résultant peut être par exemple la synthèse croisée entre un extrait de notes de piano (cible) et le son résonnant d'une cloche (source). Cela résulte en un effet de piano résonnant (ou à la substitution du timbre des notes du piano au profit de notes caractérisées par un timbre de cloche).
— Enfin nos méthodes se basent sur un apprentissage des bases et activations de plusieurs sources en premier lieu. L'apprentissage des activations de la cible en tenant compte des bases sources pré-apprises fixé est ensuite opéré. Ces opérations sont décrites dans les parties correspondantes tout au long de ce rapport.

1.3 PLAN

Dans la suite de ce rapport, nous allons décrire une à une les différentes contributions apportées à cet état de l'art.

— Dans le chapitre 2, nous présentons les différentes représentations matricielles du son ainsi que les algorithmes nous permettant de re-synthétiser le signal à partir de celles-ci . Nous rappelons les avantages de la transformée à Q-constant (CQT) comme représentation temps/fréquence. Nous présentons notre adaptation de l'algorithme de Griffin-Lim (originellement appliqué au module de la TFCT) à la CQT.
— Dans le chapitre 3, nous présentons notre méthode de musaïcing étendue à la NMF déconvolutive . Nous décrivons son fonctionnement et les modifications que nous lui avons appliqué afin de l'utiliser dans notre contexte.

— Dans le chapitre 4, nous présentons notre extension à la NMF déconvolutive 2D. Cette méthode permet d'étendre la NMF et la NMFD afin d'obtenir une invariance fréquentielle et ainsi de réduire le nombre de bases nécessaires. C'est naturellement ce qui nous a poussé à l'utiliser pour le musaïcing.
— Dans le chapitre 5, afin d'étendre la dimension créative du musaïcing, nous utilisons un algorithme de séparation harmonique/percussive sur le signal source en amont de notre musaïcing par NMFD et par NMF2D. Cette opération permet d'isoler les deux parties et de pouvoir les synthétiser indépendamment en vue de cibler de manière plus précise les caractéristiques acoustiques des sources contribuant à la reconstruction du morceau cible.
— Dans le chapitre 6, afin de quantifier nos résultats, nous présentons les résultats d'un test d'écoute que nous avons mis en place. Nous avons demandé à un panel d'auditeurs de juger de la pertinence de nos différents algorithmes.

Enfin dans le chapitre 7, nous concluons et nous parlons des perspectives de recherche futures.

2

REPRÉSENTATION EN ENTRÉE DE LA NMF ET RECONSTRUCTION DU SIGNAL AUDIO

Dans ce chapitre nous étudions les deux représentations utilisées en entrée des algorithmes NMF (le module de la TFCT et de la transformée à Q-constant) ainsi que les méthodes de reconstruction du signal audio utilisant l'algorithme de Griffin & Lim pour la reconstruction itérative du spectre de phase à partir de celui d'amplitude.

2.1 TRANSFORMÉE DE FOURIER À COURT TERME

Il est difficile de traiter les informations directement à partir de la forme d'onde d'un signal audio. Aussi, la Transformée de Fourier à Court Terme (TFCT) est couramment utilisée pour représenter son contenu [23].

Le module de la TFCT étant non-négatif est souvent utilisé en entrée des méthodes de factorisations en matrices non-négatives.

La TFCT s'inspire de la décomposition du son par l'oreille humaine : elle décompose le signal sur un banc de filtres. Les signaux audio sont non-stationnaires, leurs propriétés varient au cours du temps. La TFCT repose sur l'application de la transformée de Fourier discrète (TFD) sur de courts extraits sonores dès lors considérés stationnaires localement.

La TFD est appliquée à chaque portion de signal centrée autour d'un échantillon n :

$$X(k,n) = \sum_{m=0}^{N-1} x(m)w(n-m)e^{-j2\pi\frac{k}{N}m} \tag{8}$$

avec m le numéro d'échantillon, k les fréquences discrètes ($\forall k \in [0, N]$) et w la fenêtre de pondération dont le type et la longueur temporelle (égale pour chaque fréquence) sont les deux paramètres qui déterminent les caractéristiques spectrales obtenues. Plus la fenêtre est longue, plus la précision fréquentielle est grande. A l'inverse, plus la fenêtre est courte, plus la précision temporelle précise.

Afin d'éviter le repliement spectral, on fixe la fréquence d'échantillonnage Fe deux fois supérieur à la fréquence maximale présente dans le signal. On parle de fréquence de Nyquist $f_{Nyquist} = \frac{Fe}{2} > f_{max}$.

La figure 7 **(a)** illustre la représentation du module de la TFCT.

2.2 ALGORITHME DE GRIFFIN & LIM, SYNTHÈSE DU MODULE DE LA TFCT AVEC ESTIMATION DE LA PHASE DU SIGNAL

Dans le musaïcing par NMF de Driedger, le signal audio est reconstruit par TFCT inverse du spectrogramme complexe (méthode d'addition/recouvrement ou d'overlapp-add, OLA [1]). L'algorithme de NMF produit une matrice $\hat{V}_{n,m}$ représentant le module du spectrogramme. Afin de reconstruire la phase corres-

pondante nous utilisons l'algorithme proposé par Griffin & Lim. Celui-ci permet d'estimer itérativement la phase d'un signal :

— Soit le spectogramme d'amplitude $|Y(mS, \omega)|$ (dans lequel $S \in \mathbb{N}$ représente la période d'échantillonnage de $Y(n, \omega)$ dans la variable n et m le nombre de trames temporelles) et un signal aléatoire $x^i(n) = x_{rand}(n)$ qui nous sert d'initialisation à l'algorithme.

— On calcule le spectrogramme $X^i(mS, \omega) = \text{TFCT}\{x^i(n)\}$. On estime alors un spectrogramme complexe avec $|Y(mS, \omega)|$ en tant qu'amplitude et $\Delta X^i(mS, \omega)$ en tant que phase :

$$\hat{X}^i = |Y(mS, \omega)|.e^{i\Delta X^i(mS,\omega)} \tag{9}$$

— A partir de l'estimation du spectrogramme, on applique la méthode d'overlapp-add (OLA) modifiée. Alors que l'OLA classique s'exprime :

$$x(n) = \frac{\sum_m y_w(mS, n)}{\sum_m w(mS - n)} \tag{10}$$

dans lequel w est la fenêtre d'analyse ; l'OLA de Griffin & Lim s'exprime :

$$x^{i+1}(n) = \frac{\sum_m w(mS - n)\hat{X}^i(mS, \omega)}{\sum_m w^2(mS - n)} \tag{11}$$

— Le nombre d'itération est déterminé par l'erreur quadratique moyenne entre $|X^i_\omega(mS, \omega)|$ et $|Y(mS, \omega)|$. On appelle cette opération filtrage de Wiener. L'objectif d'un filtre de Wiener est de calculer une estimation statistique d'un signal inconnu (ici le signal re-synthétisé) à l'aide d'un signal associé en tant qu'entrée (ici un signal aléatoire) et filtrant ce signal connu pour produire l'estimation en tant que sortie. Puisque dans le cas du musaïcing, nous ne cherchons pas à reproduire exactement le signal cible, le signal final devra conserver une certaine divergence avec le signal cible. Pour cela un nombre d'itération est fixé par l'utilisateur.

En appliquant la méthode à un signal sans l'opération de musaïcing, on retrouve sensiblement le même son à l'écoute après 20 à 30 itérations.

Pour plus de clarté, la figure 6 schématise l'algorithme.

2.3 TRANSFORMÉE À Q-CONSTANT

En audio musical les fréquences sont logarithmiquement espacées (selon la gamme tempérée) : les fréquences des notes adjacentes sont plus rapprochées en basse fréquence et plus espacées en haute fréquences.

La TFCT utilise une longueur temporelle de fenêtre d'analyse identique pour toutes les fréquences, donc une résolution fréquentielle identique pour toutes les fréquences. Celle-ci peut être insuffisante pour distinguer les fréquences de notes adjacentes en basses fréquence et trop importante en haute fréquence.
L'utilisation de la transformée à Q-constant (CQT) [2] permet de résoudre ce problème par l'utilisation de longueurs temporelles variables.

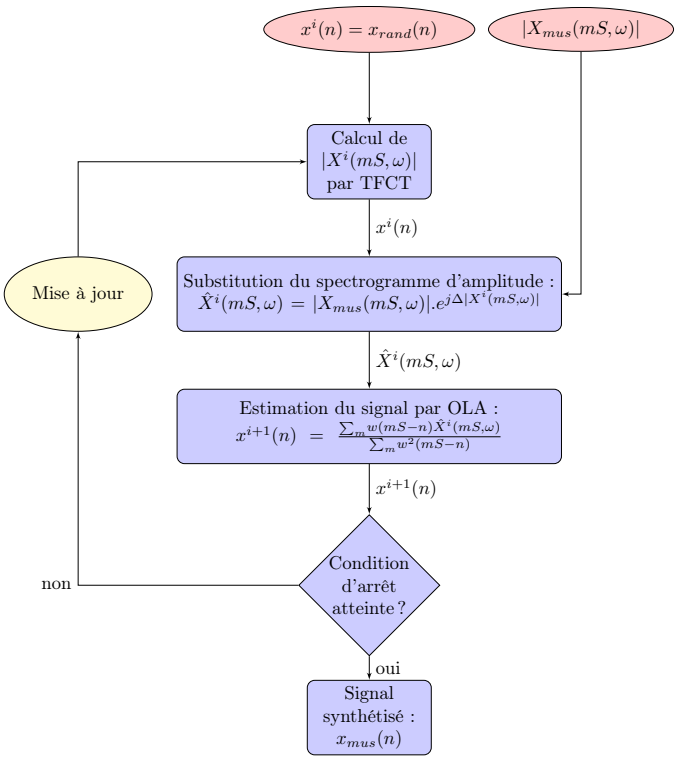

FIGURE 6 – Algorithme de Griffin & Lim.

La longueur fréquentielle de la fenêtre d'analyse est calculée en fonction de la fréquence f_k considérée.

Soit le facteur $Q = \frac{f_k}{f_{k+1}-f_k}$, constant en fréquence tel que : $Q = \frac{f_k}{B\omega} = \frac{f_k.L_k}{C\omega}$, avec $B\omega = \frac{C\omega}{L}$ la résolution fréquentielle (ou largeur de bande à -3dB) et $C\omega$ le facteur caractéristique de la fenêtre.

La longueur de la fenêtre pour chaque fréquence peut être calculée telle que : $L_k = \frac{Q.C\omega}{f_k}$.

Dans notre méthode NMF2D, nous utilisons la CQT afin de permettre l'obtention de bases (au sens de la NMF) invariantes par transposition. Par extension, nous l'utilisons également pour la NMFD afin de comparer les résultats obtenus à ceux obtenus en utilisant la TFCT.

La figure 7 compare les deux méthodes de représentation temps/fréquence TFCT et transformée à Q-constant. Le signal d'exemple possède une fréquence d'échantillonnage $Fe = 44.1$kHz.

Les paramètres de la TFCT sont les suivants :

— Nombre de points de la FFT : $N_{fft} = 4096$;

— Fenêtre de Hanning de taille : $N_w = 4096$;

— Pas d'avancement : $N_{hop} = 1024$.

Pour la CQT on utilise une résolution de 24 fréquences par octave. La musique est généralement représentée entre 0 à 8kHz, on applique une légère marge pour la bande fréquentielle de calcul de la CQT. Elle est calculée entre 20Hz et 9kHz. L'échelle des fréquences est logarithmique.

Par gain de temps, nous avons utilisé la Toolbox Matlab de Schörkhuber et al. [17] pour le calcul de la CQT d'un signal. Cette Toolbox est complète et propose

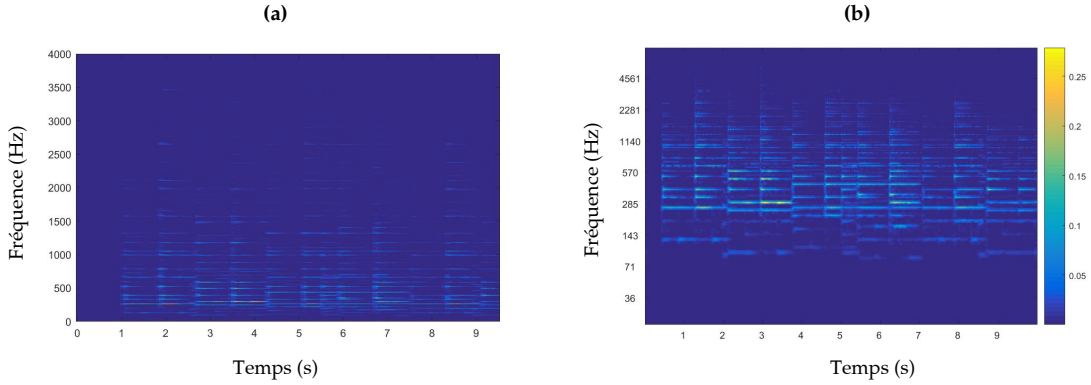

FIGURE 7 – Comparatif entre deux représentations temps/fréquence du son. L'exemple utilisé est les 10 premières secondes de "Let it be" des Beatles. **(a)** : $|\text{TFCT}|$; **(b)** : $|\text{CQT}|$.

une fonction d'inversion de la CQT. L'exemple d'utilisation fourni par la Toolbox est un script de transposition automatique à partir de la CQT d'un signal puis de sa CQT inverse.

Cet exemple légèrement modifié nous a permis de transposer certain de nos enregistrements sources préalablement à la méthode de Driedger ou au musaïcing par NMFD Section 3.

2.4 ADAPTATION DE L'ALGORITHME DE GRIFFIN & LIM À PARTIR DU MODULE DE LA CQT

Comme dans le cas de la TFCT, lorsque la NMF est appliquée au module de la transformée à Q-constant, son résultat est également un module. Afin de resynthétiser le signal audio, la phase de la transformée à Q-constant doit donc être reconstruite.

Pour cela nous nous sommes inspiré de l'algorithme de Griffin & Lim tel qu'appliqué à la TFCT [10] et l'avons modifié.

Ainsi nous choisissons d'utiliser 24 bins par octave afin d'avoir une précision deux fois supérieure à l'échelle chromatique (quart de tons). La bande fréquentielle est fixée de telle sorte à couvrir toutes les fréquences de notes.

La figure 8 montre l'adaptation de l'algorithme de Griffin & Lim au module de la CQT. On peut notifier la ressemblance avec l'algorithme appliqué à la TFCT.

Nous nous sommes ici servi de la fonction de CQT inverse fournie dans la Toolbox [17]. Lors de l'utilisation de la méthode ainsi adaptée, nous obtenons un signal final légèrement bruité. En effet un sifflement aigu est perceptible à l'écoute du signal. L'hypothèse est que la méthode reste une estimation et que l'utilisation de la CQT limite le calcul sur la bande fréquentielle choisie ce qui se traduit par des artefacts audio. Un filtrage passe-bande est appliqué entre les fréquences minimum et maximum de calcul de la CQT pour palier à ce problème.

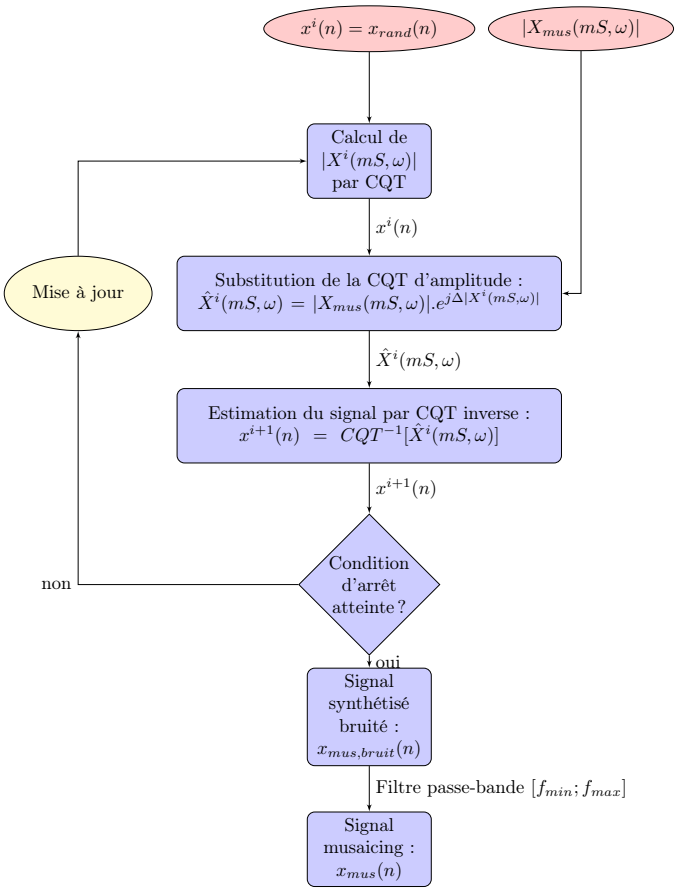

FIGURE 8 – Algorithme de Griffin & Lim adapté au module de la CQT.

3

MUSAÏCING PAR FACTORISATION EN MATRICES NON-NÉGATIVES DÉCONVOLUTIVE (NMFD)

Le modèle NMF présenté dans le chapitre précédent ne permet pas la prise en compte de la position relative de chaque spectre. Ceci entraine une perte d'information temporelle que Driedger compense par l'utilisation de contraintes appliquées lors de la l'apprentissage de la matrice d'activations (spécifiquement la contrainte de parcimonie des diagonales).

Dans son article [8], pour palier à cela, Driedger propose comme travaux futures d'étudier la déconvolution de facteurs de matrices non-négatives (NMF Déconvolutive ou NMFD) à la place de la NMF.

Le premier objectif de ce stage, est donc l'étude de la NMFD pour le musaïcing. Notre objectif est de remplacer la contrainte de parcimonie diagonale (qui assure la conservation de la temporalité des bases).

3.1 LA NMFD

Dans [21], Smaragdis introduit une version étendue de la NMF à la déconvolution (NMFD), invariante par translation temporelle. Elle est appliquée à la séparation de sources sonores. La NMFD permet donc de factoriser des motifs temps/-fréquence selon le modèle :

$$V = \hat{V} \approx \Lambda = \sum_{\tau=0}^{T-1} W^{\tau} . \overset{\tau\rightarrow}{H} \tag{12}$$

dans lequel $V_{M\times N}$ est la matrice à approximer, $W^{\tau}_{M\times K\times T}$ le tenseur des bases (chacune des K bases est une matrice de dimension $(M \times T)$) et $H_{K\times N}$ celle des activations pondérées.

L'opérateur $\overset{\tau\rightarrow}{(.)}$ indique un décalage vers la droite (de $\tau \in \mathbb{N}$) de tous les éléments de la matrice tout en mettant à zéro les éléments de gauche afin de conserver la taille original de la matrice modifiée. On pose ensuite $V = \Lambda$ et on définit la fonction de coût de Kullback-Leibler modifiée :

$$D(V\|\Lambda) = \| \, V \otimes \log(\frac{V}{\Lambda}) - V + \Lambda \, \| \tag{13}$$

Les mises à jour de la matrice d'activation et des matrices de base deviennent :

$$\begin{cases} H \leftarrow H \otimes \frac{(W^{\tau})^{T} . [\frac{\overset{\leftarrow\tau}{V}}{\Lambda}]}{(W^{\tau})^{T} . 1} \\ W^{\tau} \leftarrow W^{\tau} \otimes \frac{\frac{V}{\Lambda} . \overset{\tau\rightarrow}{H}^{T}}{1 . \overset{\tau\rightarrow}{H}} \end{cases} \tag{14}$$

A chaque itération, on met à jour la matrice H et chaque matrice W^{τ}.

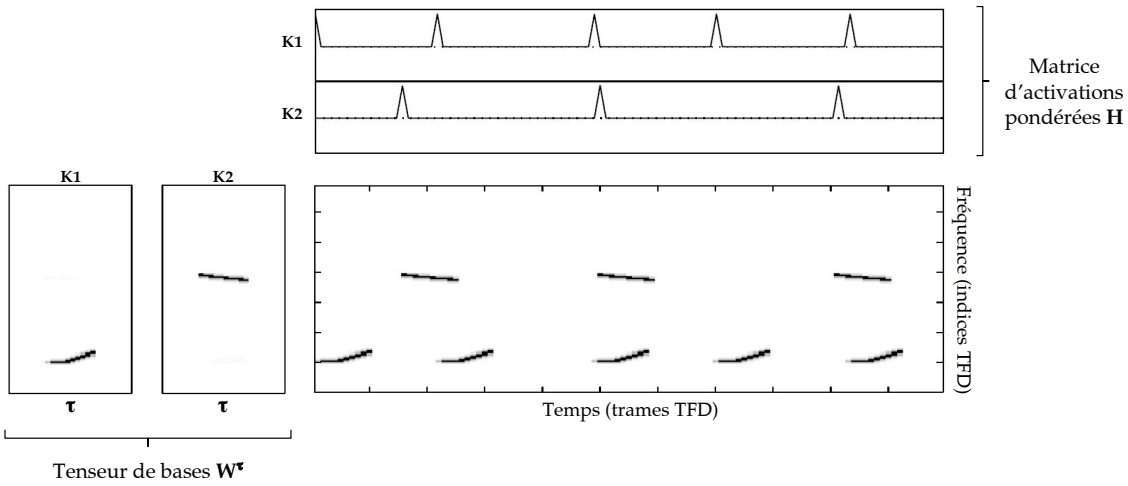

FIGURE 9 – Modèle de factorisation par NMFD *tel que publié dans [21]*.
Décomposition invariante par translation temporelle d'un spectrogramme. Les motifs de base évoluent au cours du temps. Ici $K = 2$, $T = 7$ trames temporelles.

Comme pour la NMF, H contient les activations pondérées. W^τ contient les bases étendues dans les deux dimensions de l'entrée (temps et fréquence) sous la forme de K-spectrogramme(s) (un par base). Autrement dit chaque base apprise évolue temporellement sur une courte durée. Cette représentation est donc particulièrement adéquate pour représenter les évènements musicaux dont les propriétés spectrales évoluent au cours du temps.

Nous illustrons la NMFD à la figure 9 : les motifs de bases W^τ et les fonction d'activations H sont appris à partir de la matrice V en . Notons que la matrice d'activations H peut être aussi bien représentée par K fonctions d'activations (une fonction pour chaque base).

3.2 MUSAÏCING PAR NMFD

Par rapport à la NMF, la NMF Déconvolutive apporte l'invariance temporelle des atomes spectraux W. Les bases W ne sont plus des vecteurs en fréquence mais des segments de spectrogramme de taille $(N \times T)$. Cette caractéristique de la NMFD s'inscrit parfaitement dans le cadre du musaïcing puisque l'on cherche à reproduire le morceau cible à partir de courts extraits sonores.

Nous décrivons ci-dessous l'algorithme pour le musaïcing-par-NMFD. Il est important de noter que dans notre musaïcing le signal cible n'est pas décomposé sur les bases correspondant à une seule source, mais sur les bases correspondants à N_S sources différentes (N_S enregistrements audio différents).

1. Pour chaque enregistrements (la cible et les N_S sources), calcul du module de la TFCT afin d'obtenir les matrices cible X_C et sources $X_S\{n_{src}\}$ avec $n_{src} \in [1, N_S]$.
2. Pré-apprentissage des bases par NMFD à partir des matrices sources. Afin de couvrir toutes les notes de la gamme chromatique, les sources sont préalablement transposées. Nous appliquons la NMFD sur chaque enregistrement source $X_S\{n_{src}\}$ séparément et ne gardons que les K_S-bases $W_S^\tau\{n_{src}\}$ résul-

tantes de sa décomposition.

$$X_S\{n_{src}\} = \hat{X}_S\{n_{src}\} \approx \sum_{\tau=0}^{T-1} W_S^{\tau}\{n_{src}\}.\overset{\tau\rightarrow}{H}_S\{n_{src}\} \tag{15}$$

On concatène ensuite toutes les bases obtenues (quelque soit leur source) pour obtenir un tenseur de bases fixées W_S^{τ} de taille $[N \times K \times T]$ avec $K = K_S.N_S$.

3. Alors que dans la NMFD classique, nous décomposerions la matrice cible telle que :

$$X_C = \hat{X}_C \approx \sum_{\tau=0}^{T-1} W_C^{\tau}.\overset{\tau\rightarrow}{H}_C \tag{16}$$

Dans notre algorithme de musaïcing, nous remplaçons les bases W_C par celles pré-apprises sur les sources W_S :

$$X_{mus} = \sum_{\tau=0}^{T-1} W_S^{\tau}.\overset{\tau\rightarrow}{H}_C \tag{17}$$

4. Finalement, le signal audio temporel est reconstruit par l'utilisation de l'algorithme de Griffin & Lim.

Le pré-apprentissage des bases W^{τ} par NMFD permet une pré-sélection de celles-ci au sein d'un ou de plusieurs enregistrements sources. Ainsi c'est à nous de choisir la taille des tenseurs de bases T afin de garantir une conservation du timbre des sources lors de l'écoute du signal de musaïcing. On choisit aussi le nombre de bases K.

Prenons un enregistrement source préalablement transposé. A partir de celui-ci nous apprenons K-base de longueur T de telle sorte à pouvoir minimiser la fonction de coût et ainsi reconstruire le signal source par NMFD le plus fidèlement possible. K est fixé à un multiple de 12 pour chaque enregistrement afin de pouvoir pré-apprendre un nombre de bases relatif au nombre de demi-tons présent sur l'échelle chromatique et ainsi permettre la sélection automatique de base couvrant cette échelle. Cette étape préliminaire permet d'extraire les caractéristiques les plus pertinentes du signal source utilisé.

On parle enfin de multi-source car il nous est possible de pré-apprendre des bases sur différents enregistrements source. Ces bases extraites, une simple concaténation permet de les utiliser en tant que W^{τ} fixé lors de la reconstruction du signal de musaïcing.

Le déroulement global de la méthode est indiqué sur la figure 10.

La figure 11 illustre l'application de la méthode avec la même cible qu'utilisée précédemment (le morceau "Let it be") et la même source que précédemment (un enregistrement de bourdonnement d'abeilles transposé sur l'ensemble de la gamme chromatique). 12 bases sont ensuite pré-apprises sur le morceau source par NMFD. La figure indique les spectrogrammes obtenus par TFCT des bases W^{τ}, des activations H et du morceau de musaicing reconstruit X_{mus}. Sur le spectrogramme de X_{mus}, on peut observer que l'on conserve la structure temporelle du morceau cible. Pour nos figures d'exemples nous choisissons d'utiliser un seul enregistrement source par soucis d'affichage. La figure illustre cependant le fait que plusieurs bases pré-apprises sur des enregistrements variés peuvent être utilisées.

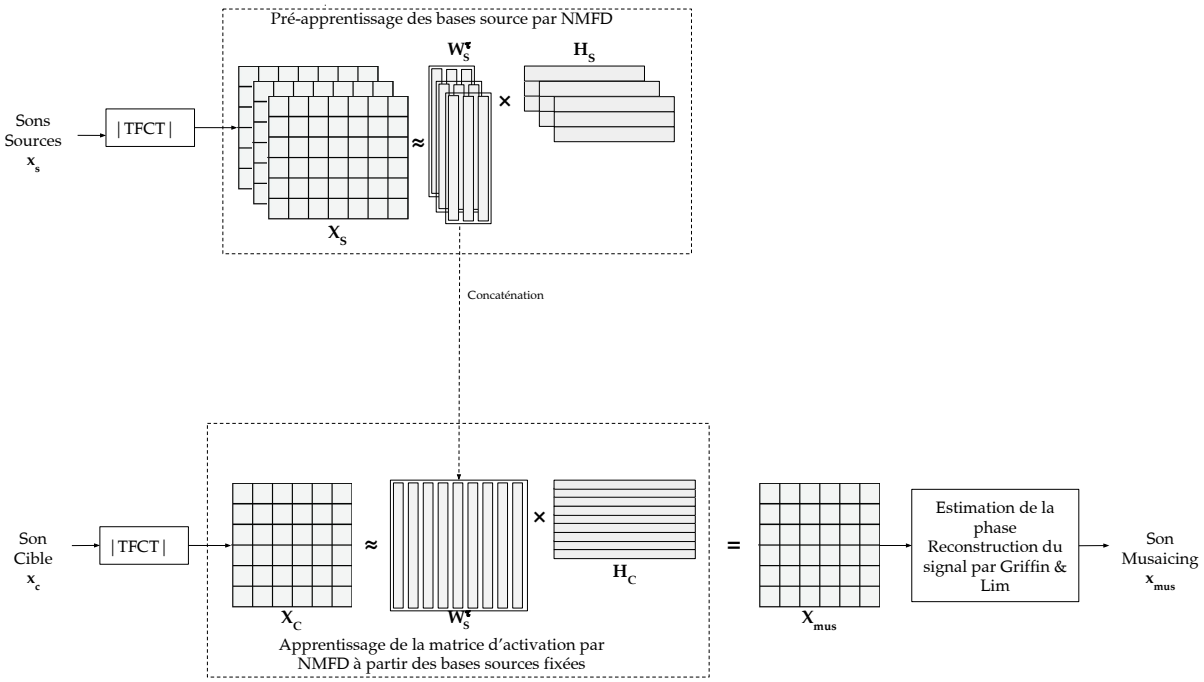

FIGURE 10 – Schéma explicatif du musaicing par NMFD.

L'algorithme précédent a été expliqué dans le cas d'une représentation temps/-fréquence de type TFCT. Nous l'avons également appliqué dans le cas d'une représentation de type transformée à Q-constant. Les résultats obtenus sont cependant assez similaires mais avec un temps de calcul deux à trois fois plus long. Nous avons donc décidé de n'utiliser la transformée à Q-constant que pour notre méthode de musaïcing par NMF2D (chapitre suivant).

3.3 MUSAÏCING PAR NMFD CONTRAINTE

Le musaïcing par NMFD reproduit de manière fidèle la structure du son cible. Cependant la superposition temporelle des bases d'une même source voir de plusieurs sources (dans le cas du multi-source) nuit à la reconnaissance des caractéristiques acoustiques de ces sources. Il est donc indispensable de contraindre l'algorithme d'apprentissage afin de favoriser la parcimonie des activations.

A l'image du musaïcing de Driedger nous appliquons deux contraintes à la matrice d'activation. Ces contraintes permettent une meilleure préservation du timbre des sources dans le résultat sonore finale du musaïcing.

3.3.1 *Contrainte sur la répétition successive des bases W^τ*

La première contrainte, appliquée lors de l'apprentissage de la matrice d'activation, permet de limiter le nombre de répétition successive d'activation d'une même base. Elle est identique à celle utilisée par Driedger (voir partie 1.2.2.3). Sa formulation se trouve pour rappel à l'équation 4.

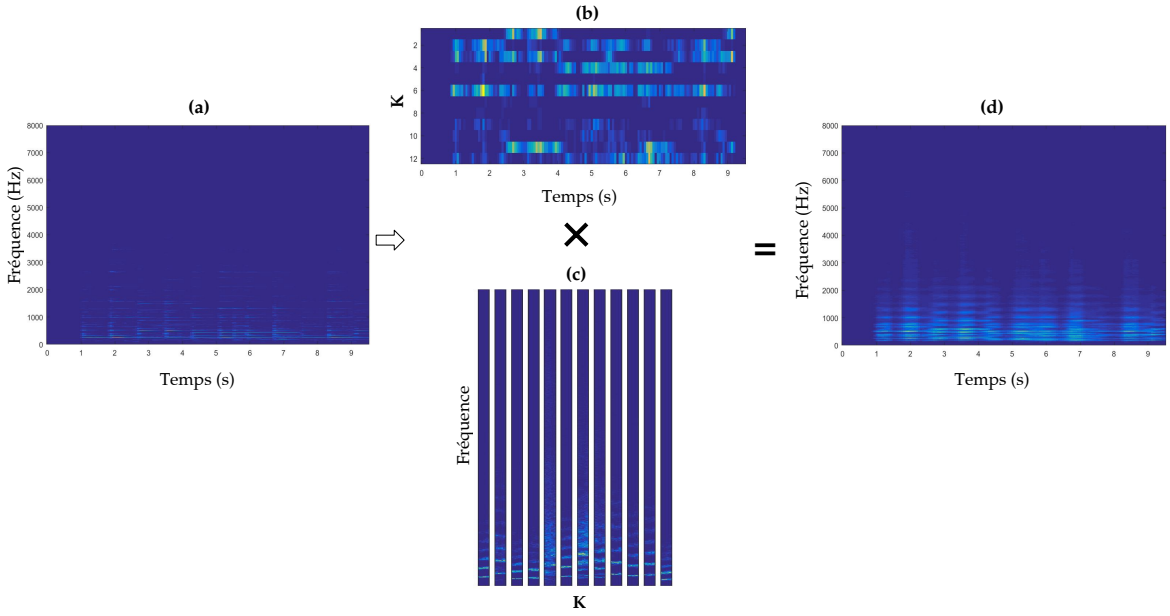

FIGURE 11 – Exemple de musaicing par NMFD.
(a) : matrice cible, "Let it be" - The Beatles ; **(b)** : tenseur de bases (pré-apprise par NMFD sur la matrice source) W^{τ} avec $K = 12$; **(c)** :matrice d'activation H ; **(d)** : matrice de musaicing reconstruite X_{mus}. Les matrices sont calculées par TFCT.

3.3.2 *Contrainte sur la superposition des bases relatives à chaque enregistrement source*

Dans le cas du musaïcing multi-sources, nous implémentons également une fonction de contrainte sur le nombre de sources activées simultanément. Notre objectif est d'éviter une superposition trop importantes de sources différentes ce qui entraine un effet de "cacophonie".

Pour cela, nous calculons à chaque itération et à chaque trame la contribution énergétique de chaque source (somme des activations correspondant aux différentes bases d'une même source). Nous conservons uniquement dans la matrice d'activations H, les activations des bases provenant de source ayant la plus grande contribution énergétique.

3.4 LIMITATION DE LA MÉTHODE

La méthode NMFD s'avère intéressante dans le cadre du musaicing de part ses propriété d'invariance par translation temporelle.

Cependant dans l'algorithme NMFD, il est toujours nécessaire de transposer (manuellement) les enregistrements sources afin de représenter l'ensemble des notes pouvant potentiellement être présentes dans l'enregistrement cible. Cela nous a encouragé à étendre cette méthode à l'invariance par translation non seulement temporelle mais également fréquentielle 4.

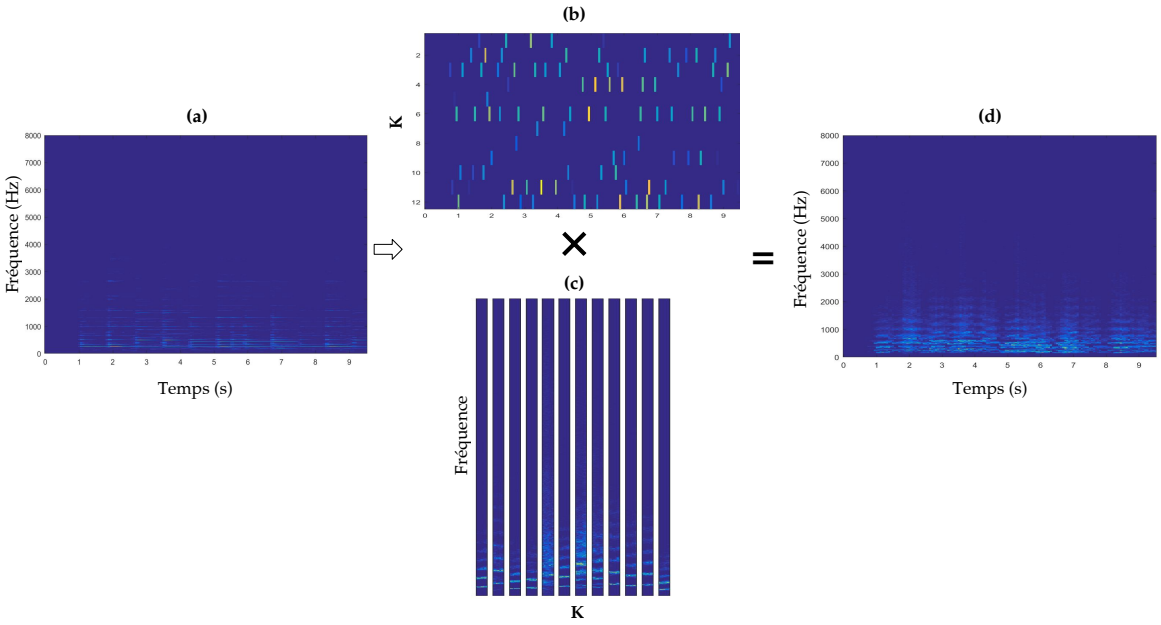

FIGURE 12 – Exemple de musaïcing par NMFD contrainte.
(a) : matrice cible, "Let it be" - The Beatles ; **(b)** : tenseur de bases (pré-apprise par NMFD sur la matrice source) W^{τ} avec $K = 12$; **(c)** :matrice d'activation contrainte H avec $r = T/2$ (T est la longueur temporelle des matrices de bases) ; **(d)** : matrice de musaïcing reconstruite X_{mus} Les matrices sont calculées par TFCT.

4

MUSAÏCING PAR FACTORISATION EN MATRICES NON-NÉGATIVES 2D DÉCONVOLUTIVES (NMF2D)

La NMFD de Smaragdis propose de reconstruire le signal par superposition de bases (atomes temps/fréquence) translatées à travers le temps. Elle repose sur l'invariance temporelle des bases.

La NMF2D, étudiée dans ce chapitre, étend cette translation des bases également à l'axe des fréquences. Pour cela une dimension fréquentielle est rajoutée aux activations H. On parle d'invariance temporelle et fréquentielle. Cette invariance fréquentielle nous permet de réduire le nombre de bases à apprendre pour une source ; puisque cette base pourra être translatée en log-fréquence afin de couvrir les différentes hauteurs de note présentes dans la cible.

Pour l'étude et l'implémentation de cette méthode, nous nous sommes basé sur l'article de Schmidt & Mørup [16]. Les auteurs appliquent la NMF2D à des morceaux musicaux complexes afin d'en séparer de manière aveugle leurs différentes sources. L'exemple d'une séparation d'un mélange piano/trompette est présenté dans l'article.

4.1 LA NMF2D

La factorisation en matrices non-négatives 2D déconvolutive permet de factoriser un spectrogramme exprimé en log-fréquence (comme celui de la transformée à Q-constant, voir la partie 2) en utilisant un modèle permettant de représenter la structure temporelle et le changement de hauteur induit lorsqu'un instrument joue différentes notes.

En partant du modèle de NMF et de NMFD, le modèle de NMF2D peut s'écrire ainsi :

$$V = \hat{V} \approx \Lambda = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} \overset{\downarrow\phi}{W^{\tau}} \overset{\rightarrow\tau}{H^{\phi}} \tag{18}$$

dans lequel $V_{M\times N}$ est la matrice à approximer, $W^{\tau}_{M\times K\times T}$ le tenseur de motifs temps/fréquence et $H^{\phi}_{K\times N\times P}$ le tenseur d'activations possédant une dimension P relative à la translation fréquentielle.

$\overset{\downarrow\phi}{(.)}$ indique un décalage (de $\phi \in \mathbb{N}$) des fréquences vers le bas de tous les éléments de la matrice en mettant à zéro les éléments du haut afin de conserver la taille originale de la matrice ainsi modifiée. On introduit alors une invariance de translation fréquentielle en plus de l'invariance de translation temporelle. En posant $\phi = \{0\}$, on retrouve le modèle de NMFD décrit par Smaragdis (voir le chapitre 3). Chaque élément de Λ peut s'écrire :

$$\Lambda = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} \sum_{k=0}^{K-1} W^{\tau}_{i-\phi,k} H^{\phi}_{k,j-\tau} \tag{19}$$

avec K le nombre de bases, i (fréquence) et j (temps) respectivement les indices de ligne et de colonne de la matrice. Dans le cas d'une fonction de coût de type

divergence de Kullback-Leibler (eq. 13), les règles de mise à jour de W^{τ}et de H^{ϕ} sont :

$$\begin{cases} W^{\tau} \leftarrow W^{\tau} . \frac{\sum_{\phi} (\frac{\overset{\uparrow\phi}{V}}{\Lambda}) \overset{\rightarrow\tau^T}{H^{\phi}}}{\sum_{\phi} 1 . \overset{\rightarrow\tau^T}{H^{\phi}}} \\ H^{\phi} \leftarrow H^{\phi} . \frac{\sum_{\tau} \overset{\downarrow\phi^T}{W^{\tau}}}{\sum_{\tau} \overset{\downarrow\phi^T}{W^{\tau}} . 1} \end{cases} \tag{20}$$

Ce modèle factorise une matrice cible V en un tenseur d'atomes W^{τ} convolué à un tenseur d'activations H^{ϕ}. Afin de permettre la translation fréquentielle d'un motif spectral harmonique (dans le sens de modification de la fréquence fondamentale) on utilise la transformée à Q-constant (CQT). Cette fois il n'est pas nécessaire de choisir K multiple des demi-tons. La transposition automatique relative à la NMF2D assure de couvrir l'ensemble du contenu fréquentiel de la cible.

La figure 13 fournit une illustration de l'application de la NMF2D. On peut observer que lors d'une activation sur les matrice H correspondant à un motif spectral W, l'intégralité du motif est translaté.

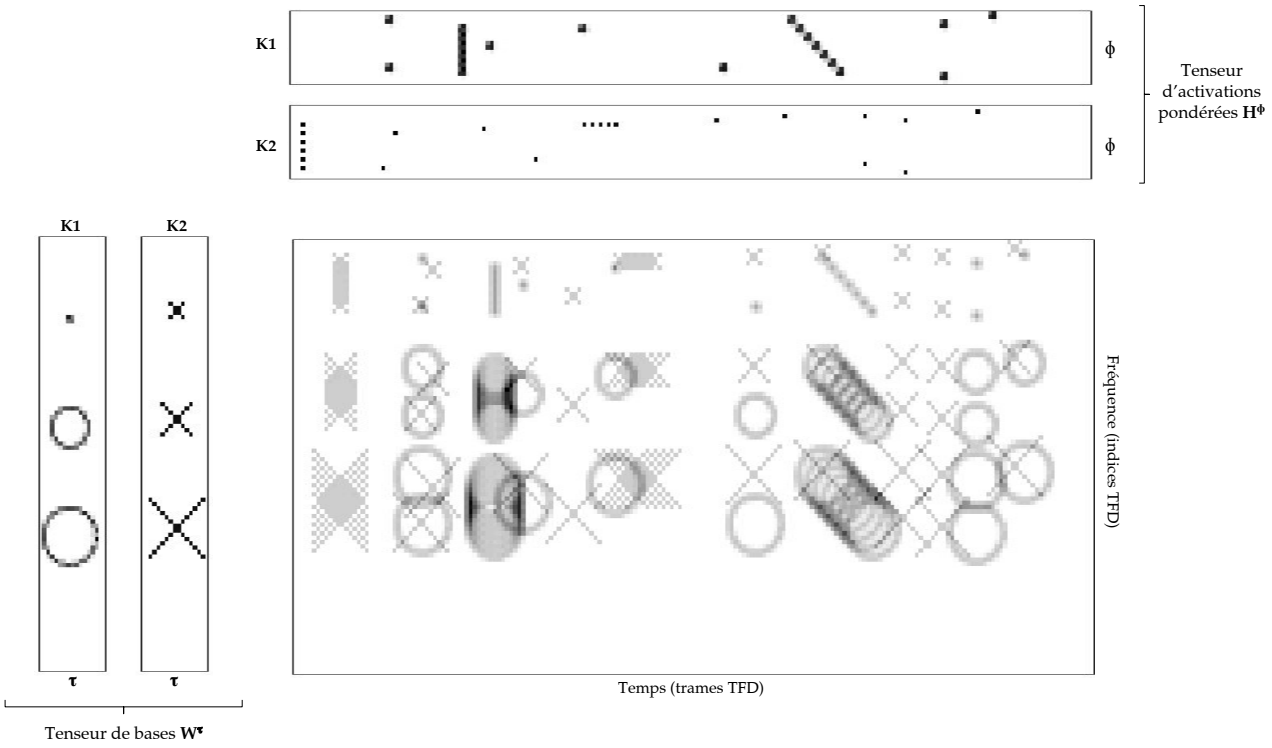

FIGURE 13 – Modèle de factorisation par NMF2D *tel que publié dans [16].*

4.2 MUSAÏCING PAR NMF2D

Le musaïcing par NMFD reste encore limité du fait que les sources doivent être préalablement transposées. C'est dans l'optique de remédier à cette contrainte que l'utilisation de la NMF2D est théoriquement plus intéressante pour le musaïcing.

Nous décrivons ci-dessous l'algorithme pour le musaïcing-par-NMF2D.

1. On calcule le module de la CQT de l'enregistrement cible et des enregistrements source afin d'obtenir les matrices X_C et $X_S\{n_{src}\}$ avec $n_{src} \in [1, N_S]$.
2. Pour chaque enregistrement source n_{src}, on pré-apprend les bases W_S^τ à l'aide de l'algorithme NMF2D. Cette fois la transposition préalable des enregistrements n'est pas nécessaire.

$$X_S = \hat{X}_S \approx \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} W_S^{\tau} \overset{\downarrow\phi}{H_S^{\phi}} \overset{\rightarrow\tau}{} \tag{21}$$

   On concatène ensuite toutes les bases obtenues (quelque soit leur source) pour obtenir un tenseur de bases unique.
3. On applique ensuite la NMF2D à la matrice cible en remplaçant les bases par le tenseur de base des sources obtenu à l'étape précédente. On obtient donc la matrice de musaïcing :

$$X_{mus} = \sum_{\tau=0}^{T-1} \sum_{\phi=0}^{P-1} W_S^{\tau} \overset{\downarrow\phi}{H_C^{\phi}} \overset{\rightarrow\tau}{} \tag{22}$$

4. Enfin on reconstruit le signal de musaïcing x_{mus} grâce à l'algorithme de Griffin & Lim adapté au module de la CQT (voir section 2.4).

Les étapes de la méthode sont représentées figure 14.

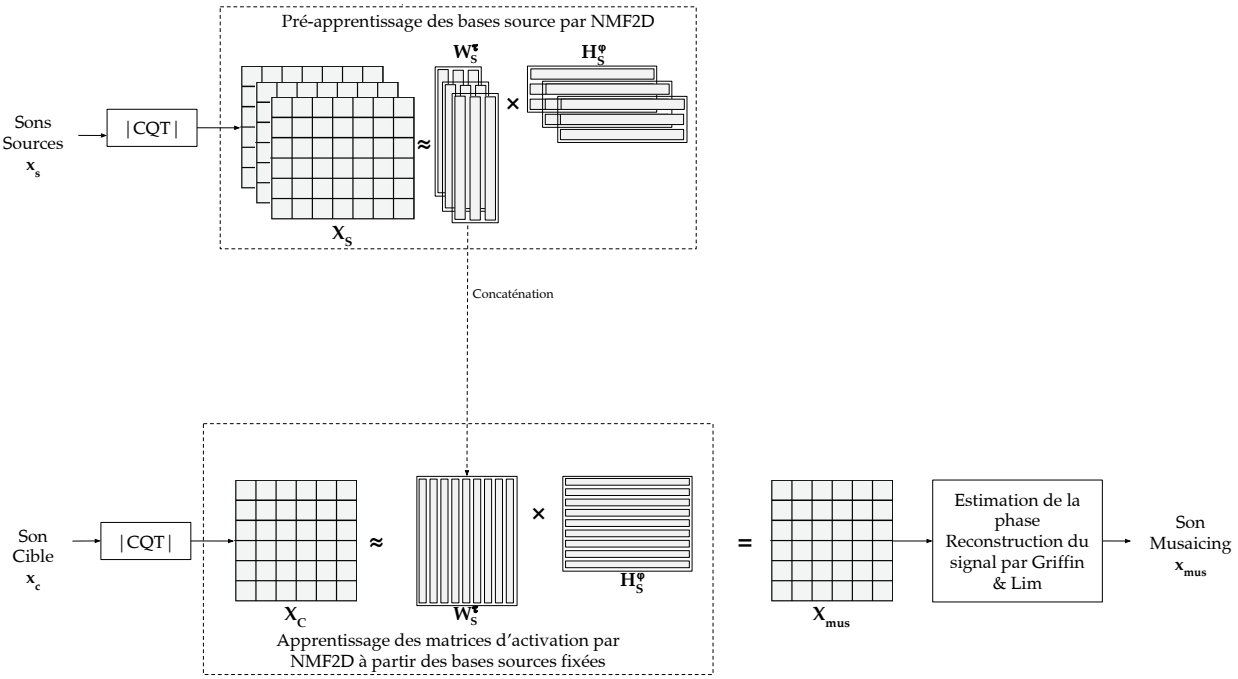

FIGURE 14 – Schéma explicatif du musaïcing par NMF2D.

De la même manière que pour la NMFD, l'extension de la méthode au multi-source est faite par pré-apprentissage des bases des différents enregistrements puis par concaténation de celles ci.

L'utilisation de la NMF2D pour le musaïcing est pertinente du fait que les bases issues des sources sont automatiquement transposées. En comparaison avec les méthodes précédentes, on obtient avec la NMF2D un signal de musaïcing x_{mus} dont la structure est plus fidèle à celle du morceau cible.

4.2.1 *Illustration*

A titre d'exemple, la figure 15 illustre notre algorithme de musaïcing-par-NMF2D avec la même cible qu'utilisée précédemment (le morceau "Let it be") et des voyelles chantées ("a", "e", "i", "ou", "u") en tant que sources. $K_S = 3$ bases sont apprises par source ce qui fait un total de $K = 15$ bases.

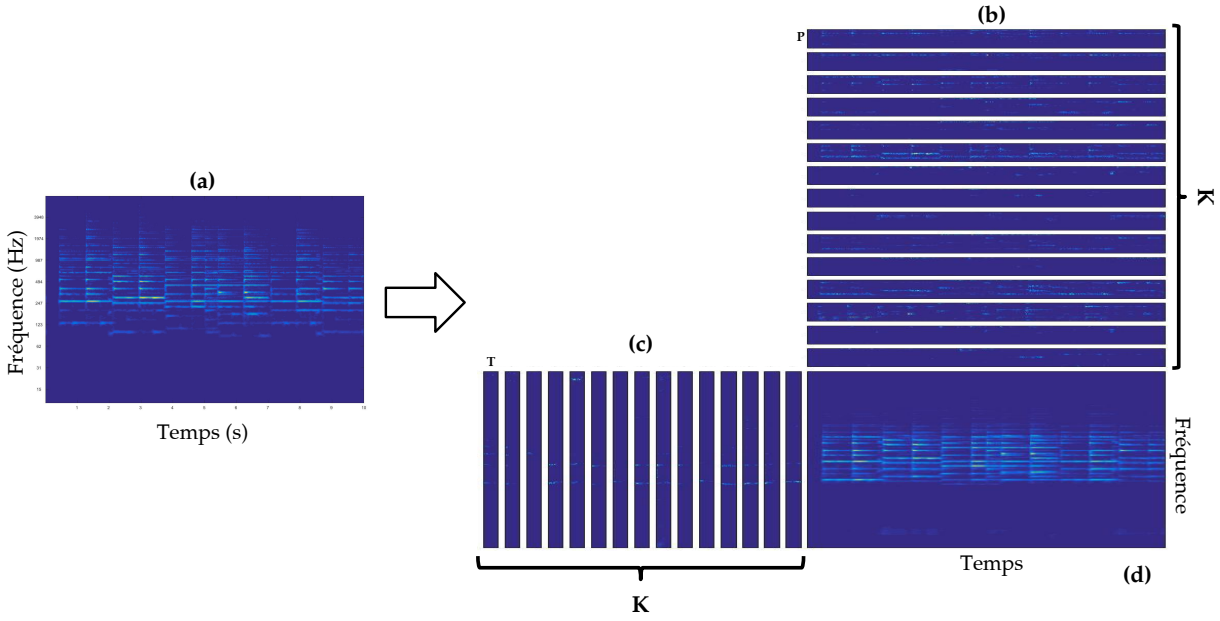

FIGURE 15 – Exemple de musaïcing par NMF2D.
(a) : matrice cible X_C, "Let it be" - The Beatles ; **(b)** : tenseur de bases (pré-apprise par NMF2D sur les matrice source) W^τ avec $K = 15$; **(c)** :matrice de musaïcing reconstruite X_{mus} ; **(d)** : tenseur d'activation H^ϕ. Les matrices sont calculé à l'aide de la CQT.

Sur cette figure, on observe une stationnarité évidente des sons sur les matrices de bases pré-apprises W^τ. Chacune est de taille $T = 200$ ce qui équivaut à 500ms environ. Visuellement les modules de CQT de la matrice cible X_C et de la matrice de musaïcing X_{mus} présentent une différence. En effet la matrice de musaïcing s'étend sur une plage fréquentielle plus limitée. Cela s'explique du fait que la plage fréquentielle couverte par chaque base n'inclue pas forcément les mêmes harmoniques que le son cible à repoduire.

4.3 MUSAÏCING PAR NMF2D CONTRAINTE

Lors de l'apprentissage de la matrice d'activation par NMF2D, on observe la formation d'activations successives ou superposées. Ce phénomène entraine une superposition des bases pré-apprises. Le résultat sonore reproduit alors le son cible avec une grande fidélité. Ce phénomène est observable sur la matrice de musaïcing figure 15. Cependant, rappelons que l'objectif du musaïcing est de permettre la conservation des caractéristiques acoustiques des sources utilisées de telles sorte à reproduire la structure du morceau cible. Lorsque le reconstruction est trop fidèle au signal cible, cela empêche d'identifier correctement les caractéristiques acoustiques des sources. Cela entraîne de plus un phénomène d'écho multiple.

4.3.1 *Contrainte sur la répétition successive des bases W^{τ}*

Comme précédemment, nous introduisons une contrainte favorisant la parcimonie temporelle des tenseurs d'activation. Cette contrainte est assez similaire à celle apposée en section 1.2.2.3 (voir l'équation 4) et est modifiée afin d'être appliquée à chaque matrice du tenseur d'activation :

$$R^{\phi}_{km} = \begin{cases} H^{\phi}_{km} & \text{si } H^{\phi}_{km} = \mu^{r,\phi}_{km} \\ H^{\phi}_{km}(1-\lambda_{it}) & \text{sinon} \end{cases} \tag{23}$$

avec $\mu^{r,\phi}_{km}$ la valeur maximale de H^{ϕ} dans un voisinage horizontal $[m-r; m+r]$.

Chaque valeur présente dans une matrice d'activation composant le tenseur H^{ϕ} déclenche l'activation de la base relative (transposée selon ϕ dans la CQT). Afin d'éviter la superposition de plusieurs bases de longueur T, on fixe $r = T/2$.

4.3.2 *Contrainte sur la superposition des bases relatives à chaque enregistrement source*

Une seconde contrainte, elle aussi basée sur celle de Driedger, permet de limiter le nombre d'activations simultanées. On parle ici encore de limitation polyphonique. Un degré de polyphonie p est fixé et pour chaque matrice composant le tenseur d'activation on applique la contrainte suivante lors de la mise à jour de l'algorithme de NMF2D :

$$P^{\phi}_{km} = \begin{cases} R^{\phi}_{km} & \text{si } k = \Omega^{p,\phi}_{m} \\ R^{\phi}_{km}(1-\lambda_{it}) & \text{sinon} \end{cases} \tag{24}$$

avec $\omega^{p,\phi}_{m}$ contenant les indices des p-plus grands éléments de la $m^{ième}$ colonne de R^{ϕ}.
Finalement cette contrainte nous permet de limiter la superposition des activations sur chaque trame spectrale.

En plus de cette contrainte nous avons fait des test avec la contrainte sur l'énergie relative à chaque source décrite dans 3.3.2

4.3.3 *Illustration*

La figure 16 illustre l'application de ces contraintes sur le même exemple de la figure 15. On peut observer que la matrice de musaïcing est très proche de la matrice cible (mis à part l'effet de filtrage). Cela induit donc une reconstruction fidèle de la structure du morceau cible tout en utilisant les caractéristiques acoustiques des bases sources.

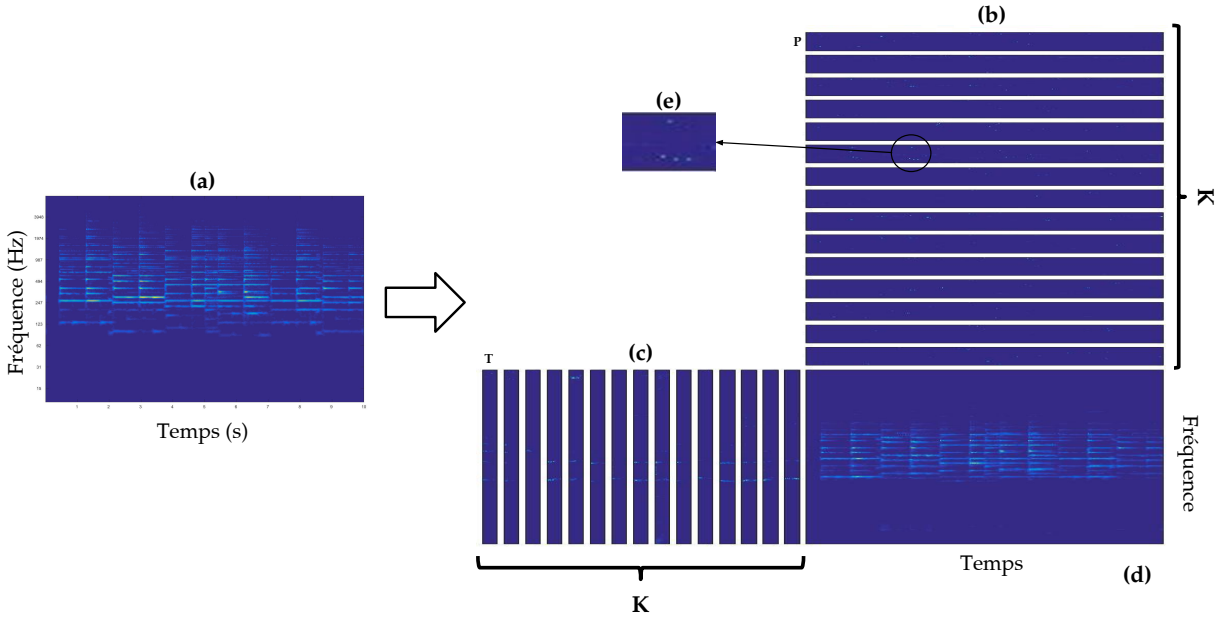

FIGURE 16 – Exemple de musaïcing par NMF2D contrainte.
(a) : matrice cible X_C, "Let it be" - The Beatles ; **(b)** : tenseur de bases (pré-apprise par NMF2D sur les matrice source) W^τ avec $K = 15$; **(c)** :matrice de musaïcing reconstruite X_{mus} ; **(d)** : tenseur d'activation H^ϕ ; **e** : zoom sur une matrice d'activation parcimonieuse. Les matrices sont calculé à l'aide de la CQT. Les contraintes appliquées correspondent à $r = T/2 = 100$ et $p = 12$

5

MUSAÏCING POST-SÉPARATION HARMONIQUE/PERCUSSIVE

Les travaux récents de Clément Laroche [12] montrent que l'utilisation d'algorithmes NMF dédiés aux caractéristiques du signal permettent l'obtention d'une meilleure décomposition. Pour cela, l'auteur propose l'utilisation d'une NMF orthogonale pour la partie harmonique et d'une NMFD pour la partie percussive.

C'est cette idée que nous avons choisi de développer dans notre algorithme de musaïcing par NMF dans l'optique de cibler de manière plus précise les caractéristiques acoustiques des éléments sources qui vont reconstituer le signal de musaïcing. Pour cela, le signal est d'abord séparé dans ses deux composantes harmoniques et percussives. On parle de séparation de source harmonique/percussive lorsqu'à partir d'un morceau de musique on extrait d'un coté la partie harmonique (les sons instrumentaux stationnaire) et de l'autre la partie percussive (les sons de percussion plus impulsifs). Cette séparation est effectuée à partir d'une représentation temps/fréquence obtenue par TFCT. Cette opération que nous appellerons HPSS par la suite peut être réalisée via diverses méthodes. Nous nous sommes intéressé à celle de Filzgerald [9] en particulier pour sa simplicité et son efficacité.

Les deux signaux résultants sont ensuite décomposés à l'aide d'algorithme NMF spécifiques à leurs caractéristiques : un algorithme NMF2D pour la partie harmonique, un algorithme NMFD pour la partie percussive.

5.1 DÉCOMPOSITION HPSS DE FILTZGERALD

Dans [9], Fitzgerald propose une méthode de séparation de source harmonique/percussive (HPSS) basée sur la morphologie du spectrogramme correspondant à chacune de ces deux parties. Il constate que les sons musicaux harmoniques stationnaires présentent des motifs **horizontaux**, alors que les sons percussifs (tels que les sons de batteries) sont plus impulsifs et présentent des motifs **verticaux**.
Tenant compte de ces observations, il considère le spectrogramme d'un morceau $X_{m,n}$ comme l'addition d'un spectrogramme harmonique $H_{m,n}$ et d'un spectrogramme percussif $P_{m,n}$:

$$X_{m,n} = H_{m,n} + P_{m,n} \tag{25}$$

La première étape vise à la création des spectrogrammes relatifs à chaque partie. Afin d'isoler la partie harmonique, pour chaque fréquence, un filtrage médian est appliqué à travers le temps (celui-ci en retire les composantes percussives considérées comme des discontinuités à travers le temps). Afin d'isoler la partie percussive, pour chaque temps, un filtrage médian est appliqué à travers les fréquences (celui-ci en retire les composantes harmoniques considérées comme des discontinuités à travers les fréquences). Pour rappel un filtrage médian permet de remplacer chaque point d'entrée par la valeur médiane de son voisinage.

A partir de H et P, deux masques sont générés et appliquer au spectrogramme du morceau. Fitzgerald décrit deux types de masques :

— Un masque binaire (ou masque "dur"), considérant que chaque bin fréquentiel est une composante soit de H, soit de P :

$$M_{H_{m,n}} = \begin{cases} 1, & si \quad H_{m,n} > P_{m,n} \\ 0, & sinon \end{cases} \tag{26}$$

$$M_{P_{m,n}} = \begin{cases} 1, & si \quad P_{m,n} > H_{m,n} \\ 0, & sinon \end{cases} \tag{27}$$

— Un masque basé sur le filtrage de Wiener [24] (masque "doux") :

$$M_{H_{m,n}} = \frac{H_{m,n}^{p}}{(H_{m,n}^{p} + P_{m,n}^{p})} \tag{28}$$

$$M_{P_{m,n}} = \frac{P_{m,n}^{p}}{(H_{m,n}^{p} + P_{m,n}^{p})} \tag{29}$$

avec p la puissance de chaque élément individuel au sein du spectrogramme. Usuellement, p vaut 1 ou 2.

Enfin on applique les masques au spectrogramme du morceau à séparer afin de retrouver les spectrogrammes complexes relatifs à chaque partie :

$$\hat{H}_{m,n} = \hat{X}_{m,n}.M_{H_{m,n}} \tag{30}$$

$$\hat{P}_{m,n} = \hat{X}_{m,n}.M_{P_{m,n}} \tag{31}$$

Finalement une simple TFCT inverse nous permet d'obtenir les signaux harmonique $h(t)$ et percussif $p(t)$ constituant le morceau cible $x(t)$.

5.1.1 *Illustration*

La figure 17 résume la méthode de HPSS utilisée. Elle y est appliquée sur un morceau de jazz contenant piano et batterie.

Les paramètres choisis pour la TFCT et la longueur du filtre médian sont ceux utilisés par Fitzgerald dans son article [9] (après plusieurs tests ce sont les paramètres optimaux pour une majorité de signaux audio différents). Les deux types de masque ont été testé et les résultats avec un masque "doux" et $p = 2$ sont plus probants.

5.2 ALGORITHME DE MUSAÏCING POST HPSS

Dans notre méthode musaïcing, la méthode HPSS est appliquée en amont. Elle permet de pouvoir isoler les parties harmonique et percussive de notre signal cible et de les remplacer par des sources plus adaptées à chaque partie.

Nous décrivons ci-dessous notre algorithme modifié afin de prendre en compte les avantages du HPSS.

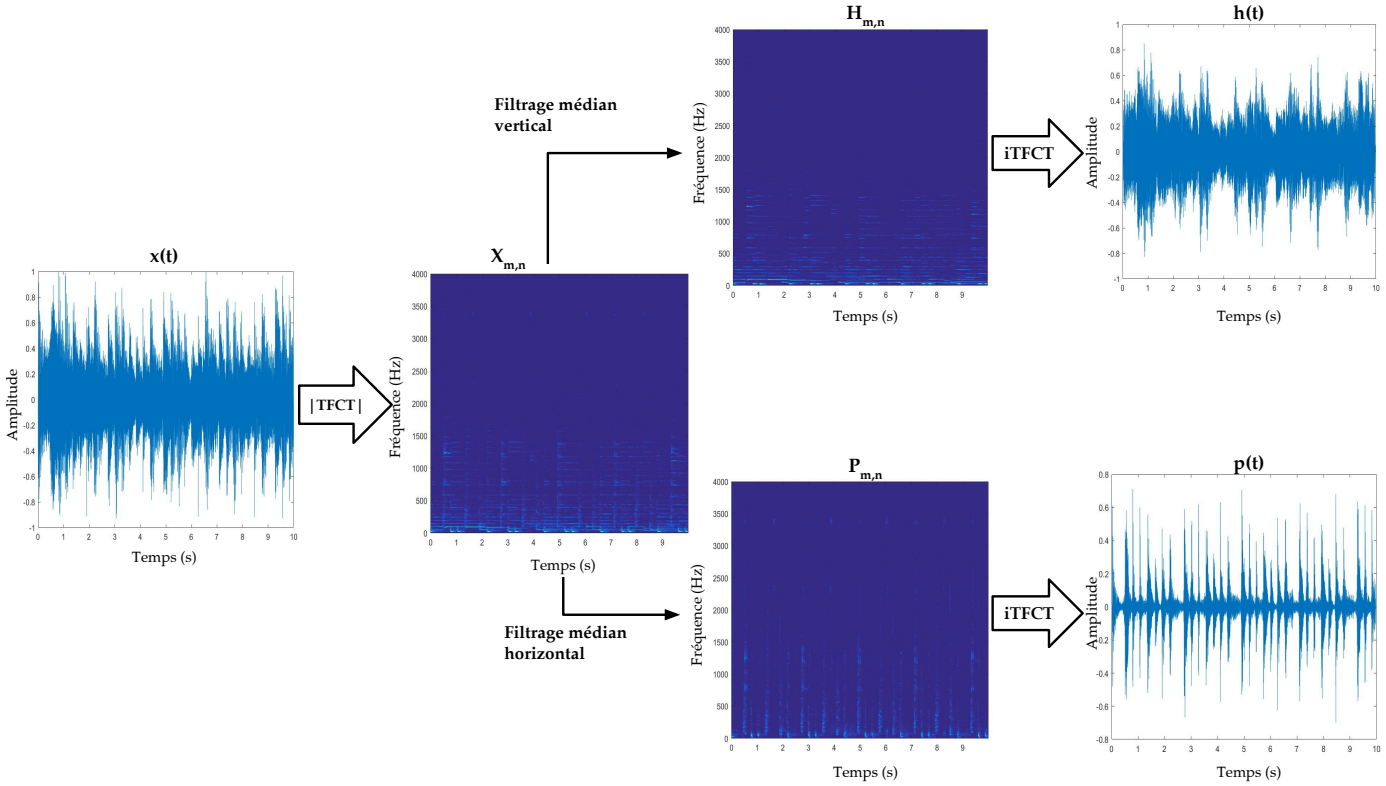

FIGURE 17 – HPSS de Filtgerald.

1. On applique la méthode HPSS au signal cible x_C. Nous obtenons un signal cible harmonique x_{C_h} et un signal cible percussif x_{C_p}
2. On pré-apprends des bases relatives à chaque partie sur des enregistrements sources.
   — Des bases harmoniques sont pré-apprise par NMF2D sur des enregistrements présentant des sonorités stationnaires (tel que des voyelles chantées) W_{S_h}.
   — Des bases percussives sont pré-apprise par NMFD sur des sons plus impulsifs (tel que des congas) W_{S_p}. On utilise la NMFD en s'inspirant du travail de Clément Laroche [12]. Ce dernier utilise cette méthode à des fins d'extraction des sons percussifs d'un morceau. Les sons percussifs ne nécessitent pas de transposition automatique, d'où l'utilisation de la NMFD.
3. On apprends les activations de la partie harmonique H_{C_h} par NMF2D et celles de la partie percussive H_{C_p} par NMFD.
4. On re-synthétise les signaux indépendemment par l'algorithme de Griffin & Lim correspondant (sachant que la NMFD utilise le module de la TFCT comme entrée alors que la NMF2D utilise celui de la CQT).
5. On additionne les deux signaux obtenus pour reconstituer notre signal de musaicing final : $x_{mus} = x_{mus_h} + x_{mus_p}$.

La figure 18 résume notre algorithme modifié utilisant le HPSS en amont d'une décomposition par NMF2D et par NMFD.

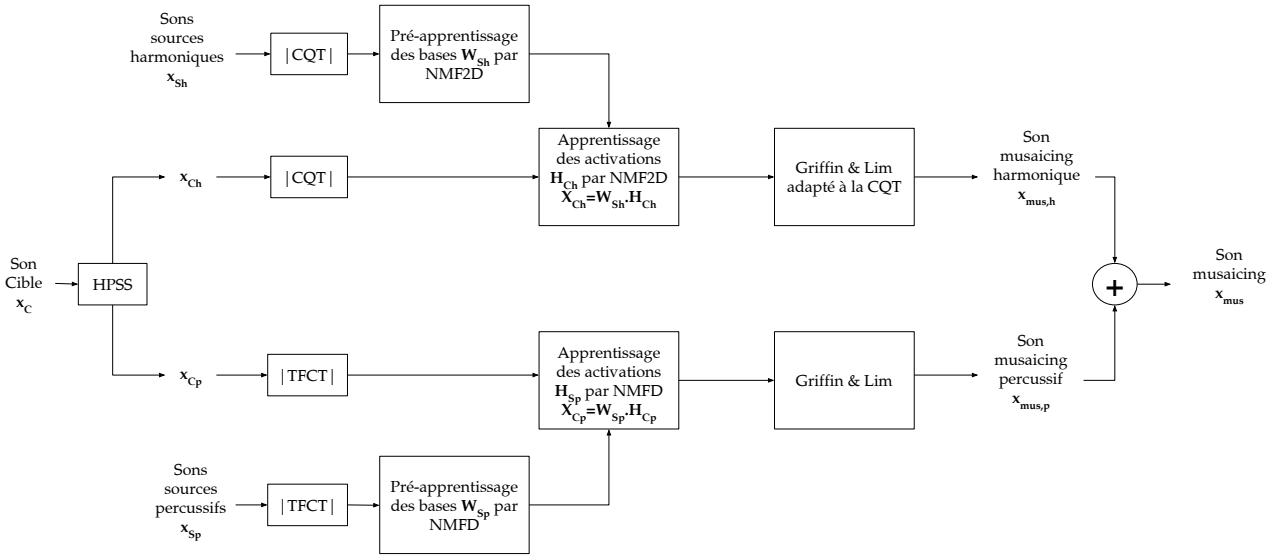

FIGURE 18 – Musaicing mixte post-HPSS.
Les différentes méthodes utilisées sont détaillées dans les chapitres correspondant.

6 RÉSULTATS

Dans cette partie, nous présentons - l'implémentation des différentes méthodes utilisées, - ainsi que le test d'écoute mis en place pour évaluer perceptivement la qualité des différents algorithmes. Nous commenterons finalement les résultats du test obtenus.

6.1 IMPLÉMENTATION DES MÉTHODES

L'ensemble des méthodes étudiées et proposées au cours de ce stage ont été implémentées sur Matlab. Nous avons occasionnellement utilisé le logiciel Audacity pour traiter certains extraits audio.

6.1.1 *Musaïcing de Driedger*

Dès le début du stage, nous avons ré-implémenté le musaïcing de Driedger afin de s'immerger au mieux dans la méthode de l'état de l'art. Nous l'avons ré-implémenté en suivant les étapes de son article à savoir :

— La NMF basique,
— Le musaïcing par NMF basique,
— Les différentes contraintes,
— Le musaïcing par NMF contrainte.

Nous avons également ré-implémenté l'algorithme de Griffin & Lim tel que décrit dans l'article [10]. Une fois cette étape achevé nous étions capable d'estimer la phase d'un signal à partir du module de sa TFCT afin d'en reconstruire le signal. Chaque élément constitue une fonction Matlab.

6.1.2 *Musaïcing par NMFD*

La fonction de NMFD a été ré-implémentée conformément à l'article de [21]. Les contraintes appliquées lors de l'apprentissage de la matrice d'activations ont chacune été implémentées dans une fonction spécifique.
L'architecture du programme est organisée à l'image du musaïcing de Driedger :

— Une fonction NMFD utilisant les fonctions contraintes ou non selon les paramètres utilisés en entrée.
— Un script appelant successivement les différentes étapes de la méthode à savoir - le pré-apprentissage des bases, - leur concaténation en cas de multi-sources, - l'apprentissage des activations (contraintes ou non) et enfin - la reconstruction du signal par Griffin & Lim.

Afin de tester l'efficacité du musaïcing par NMFD, plusieurs tests ont été effectué - ceci afin d'estimer les meilleurs paramètres de la méthode - mais également afin de tester l'adéquation des différents choix de sources et de cibles possibles.

1. Afin de comparer l'algorithme NMFD à celui de Driedger, nous avons dans un premier temps utilisé les mêmes cibles que celui-ci (les 30 premières secondes de "Let it be" et "Funk jazz") et les mêmes sources (Abeilles et Baleine).
2. Ensuite, afin de mettre en évidence le musaicing multi-source, nous utilisons un panel plus grand d'enregistrement pour le pré-apprentissage des bases (une base de morceau de reggae, des enregistrements de voyelles).

Nous avons également comparé l'utilisation de la TFCT et de la CQT comme représentation en entrée.

Nous avons pu constater que le temps d'exécution de l'algorithme utilisant la CQT est bien plus long que la TFCT pour des résultats finaux similaires.

6.1.3 *Musaïcing par NMF2D*

Pour ce qui est de la NMF2D, nous avons utilisé la Toolbox Matlab de Schmidt & Mørup [15], les auteurs de l'article de référence [16].

Dans un premier temps nous avons adapté le programme au musaïcing (originellement appliqué à la séparation de source). Nous avons ensuite suivit la même logique que pour l'implémentation des méthodes précédentes. Notons que du faite de l'invariance fréquentielle recherchée, seule la CQT est utilisée en entrée. L'utilisation de la CQT et les multiplications matricielles induites par la convolutions en deux dimensions entrainent en revanche un temps de calcul pour le pré-apprentissage des bases et l'apprentissage de la matrice d'activation relativement long. Afin de diminuer ce temps nous avons remplacer les multiplications matricielles pour la reconstruction de matrice par NMF2D par des fichiers MEX codés en C [1]. Cela a permis de diviser le temps d'exécution par 10 pour l'ensemble de l'algorithme de musaïcing par NMF2D.

6.1.4 *Musaïcing post HPSS*

Dans le cas du musaïcing post HPSS, le signal cible est d'abord séparé en ses deux composantes (harmonique et percussive) par l'algorithme HPSS.

Les paramètres de cet algorithme sont la taille du filtre médian et le type de filtre utilisé pour réaliser la séparation (binaire ou de Wiener).

Pour cela, les étapes décrites au Chapitre 5 sont codées dans un script matlab. Celui-ci appel ensuite les fonctions décrites ci-dessus pour la la NMFD et la NMF2D.

6.2 TESTS D'ÉCOUTE

L'évaluation des résultats obtenus avec les différentes méthodes est une tâche compliquée. Notre objectif étant l'utilisation des techniques NMF, NMFD, NMF2D et HPSS pour le musaïcing et non pour la séparation ou la transcription, nous ne pouvons basé notre évaluation sur des mesures de type erreur de reconstruction, rappel, précision ou les mesures de de séparation fournies par la tooblox bss_eval.

Nous avons donc décidé de mettre en place un test perceptif afin de quantifier quatre mesures qualitatives indiquées ci-dessous. Le test d'écoute permet à des

1. Merci à Frédéric Cornu pour l'implémentation des fichiers MEX.

auditeurs de juger sous forme de questionnaire à points les exemples générés par nos différentes méthodes.

6.2.1 *Matériaux sonores*

Nous avons sélectionné parmi l'ensemble des signaux sonores testés pendant ce stage, un sous-ensemble recouvrant celui utilisé par Driedger pour sa méthode.

Pour les enregistrements **cibles** nous avons donc sélectionné

— "Let it be" des Beatles et

— "Funk jazz" de Music Delta.

Tous deux font partie des exemples utilisés par Driedger pour évaluer sa méthode [2]. Pour les deux sons cibles nous limitons les test au 30 premières secondes. Let it be contient alors des notes de piano ainsi qu'une partie chantée. Funk jazz est plus instrumentale, on y retrouve des percussions et de la trompette. Ce choix s'impose par soucis de comparaison entre nos méthodes et l'état de l'art.

Pour les enregistrements **sources** nous avons utilisé les deux enregistrements sources de Driedger

— un enregistrement de bourdonnement d'abeille,

— un enregistrement de chants de baleine à bosses.

Afin de tester le musaïcing multi-sources (pré-apprentissage des bases sur une collection audio), nous avons également utilisé les deux collections audio suivantes :

— un ensemble de dix morceaux de reggae selectionnés aléatoirement dans la base GTZAN [22],

— un enregistrement de cinq voyelles chantées.

Les signaux audio sélectionnés sont préalablement sous-échantillonnés à 16kHz afin de réduire le temps de calcul des méthodes.

6.2.2 *Questionnaire*

Le test d'écoute s'effectue sous forme d'un questionnaire accessible en ligne. Il est demandé aux auditeurs de juger sous forme de questionnaire à points les exemples générés par nos différentes méthodes.

Ce questionnaire est programmé en php et html5 [3] est accessible à l'adresse suivante : `http://recherche.ircam.fr/anasyn/expe/expeNMFv02.php`.

Le questionnaire a été envoyé non seulement à des chercheurs de l'IRCAM (familiers des techniques de traitement du signal audio) mais aussi à des personnes extérieures à la recherche musicale.

Après une brève présentation du projet et du questionnaire, 28 exemples sonores (présentés dans un ordre aléatoire et anonymisés) sont donnés à écouter. Chacun de ces sons correspond au résultat du musaicing obtenu par une méthode spécifique (NMF, NMFD, NMF2D avec ou sans contrainte) appliqué à une cible spécifique et utilisant des sources spécifiques. Les sons originaux de la cible et des sources sont mis en correspondance afin d'être écouter.

Pour chacun des 28 exemples sonores, il lui est donc demandé de répondre à 4 questions en notant chacune sur une échelle de 1 (Bad) à 5 (Excellent) :

2. Les exemples des résultats de Driedger sont accessibles à l'adresse `https://www.audiolabs-erlangen.de/resources/MIR/2015-ISMIR-LetItBee/`

3. Un grand merci à Luc Ardaillon pour nous avoir fournis le modèle de questionnaire ainsi qu'à Geoffroy Peeters pour l'implémentation de celui ci.

1. *Comment jugez-vous la qualité du son ?*
2. *La structure temporelle et harmonique du son cible est-elle reconnaissable ?*
3. *Le(s) son(s) source(s) sont ils identifiables(s) ?*
4. *Comment jugez vous l'intérêt créatif de l'exemple sonore ?*

La première question relative à la qualité sonore permet de rendre compte de l'audibilité de l'exemple. La seconde et la troisième sont directement jointe à l'objectif global du projet de stage. La dernière permet d'obtenir un avis artistique pour chaque exemple, les applications du projet ayant pour attrait principal la créativité artistique. Une fois le questionnaire rempli par l'auditeur, nous recevons les réponses en fichier texte par mail. Un script Matlab permet enfin d'analyser des réponses obtenues.

6.2.3 *Musaicïng post-HPSS*

Nous n'avons pas inclut l'évaluation du musaïcing post-HPSS dans ce test perceptif puisque cette comparaison ne serait pas honnête au vue de la plus grande sophistication de cet algorithme. En effet, il utilise non pas une collection pour le pré-apprentissage des bases mais deux collections correspondant aux sons harmoniques et percussifs.

6.3 RÉSULTATS

6.3.1 *Participation*

Nous avons obtenu 30 réponses de notre panel d'auditeurs. En plus des questions relatives aux exemples sonore, il était demandé quelques informations sur les participants :

— *Sexe ?*
25 hommes et 5 femmes ont participé au test.

— *Age ?*
La moyenne d'age des participants est de 26.07 ans. Le plus jeune a 18 ans et le plus âgé 44.

— *Êtes vous professionnel du milieu du traitement du signal audio ?*
La majorité des participants sont des professionnels dans le domaine du traitement du signal audio. On en compte 21 pour 9 participants n'étant pas dans ce domaine.

— *Êtes vous familier des test perceptif ?*
16 participants ont déjà effectué un test perceptif. 14 n'on jamais participé à cette expérience.

— *A l'aide de quel type de matériel avez vous réalisez ce test ?*
27 participants ont utilisé un casque audio pour réaliser le test, 2 personnes ont utilisé des enceintes et une seule personne s'est servi d'écouteurs.

6.3.2 *Analyse des résultats*

Les réponses ont été analysé via un script Matlab appliqué aux notes attribuées à chaque exemple. Afin de les interpréter au mieux, nous affichons quatre graphes. Chacun de ces graphes représente les réponses relatives à une question. Les figures

19, 20, 21 et 22 nous montrent pour chaque cible, source et méthode les résultats statistiques liés aux réponses du panel d'auditeurs. Pour plus de clarté, les figures sont fractionnées en sources (en gris) et en cible (en rouge) utilisées pour les algorithmes de musaïcing.

Nous observons les résultats sous forme de moyenne (traits horizontaux) et d'intervalle de confiance (traits verticaux variables). La notion d'intervalle de confiance permet d'obtenir des informations synthétiques sur une "population" que l'on ne connait pas entièrement. Il permet donc d'évaluer la précision de l'estimation d'un paramètre statistique sur un échantillon. On peut le calculer en connaissant la moyenne, l'écart-type, le nombre de participants et en fixant un niveau de confiance (95%) selon la formule suivante :

$$IC = moyenne \pm t\frac{\sigma}{\sqrt{n}} \tag{32}$$

avec t un coefficient fonction du niveau de confiance, σ l'écart-type et n la taille de l'échantillon.

Dans notre cas, cette notion est utile pour souligner les différences statistiques plus ou moins importantes de nos résultats. Lorsque les intervalles de confiance se recouvrent, les résultats sont plus difficiles à comparer.

Afin d'analyser les résultats, nous observons les statistiques de chacune des méthodes question par question.

QUESTION 1 : COMMENT JUGEZ-VOUS LA QUALITÉ DU SON ? (FIGURE 19) La qualité des exemples audio est jugée assez similaire d'une méthode à une autre. Les moyennes des réponses sont en grande majorité situées entre 2 et 3. Comme nous l'avons évoqué précédemment, les signaux audio utilisés en entrée des algorithmes ont été sous-échantillonné par gain de temps. Cette opération explique les similitudes quand à la qualité sonore des exemples.

QUESTION 2 : LA STRUCTURE TEMPORELLE ET HARMONIQUE DU SON CIBLE EST-ELLE RECONNAISSABLE ? (FIGURE 20) Le panel d'utilisateur a répondu de manière plus variée selon la méthode employée pour le musaïcing. En effet, les exemples générés par NMF2D sont jugés relativement efficaces en terme de conservation de la structure acoustique du morceau cible particulièrement lors de l'utilisation de Reggae en tant que source.

Le musaïcing de Driedger possède également un bon score.

Nous constatons cependant que les exemples générés par ces deux méthodes contraintes sont jugés moins efficaces que les non-contraintes correspondantes.

En revanche, les exemples générés par NMFD possèdent un moins bon score pour la conservation de la structure de la cible.

L'utilisation de cibles différentes n'influe pas sur les scores de manière générale. Les sources quant à elles, de par leur nature différentes (voyelles plus stationnaire que baleines par exemple), entrainent des résultats perceptifs différents. Ainsi les scores de la méthode par NMF2D utilisant des sources telles que les voyelles sont statistiquement plus élevés que par NMFD. Que ce soit pour les abeilles ou pour les baleines (mono-source), la méthode de Driedger non-contrainte possède les meilleurs score.

QUESTION 3 : LE(S) SON(S) SOURCE(S) SONT ILS IDENTIFIABLES(S) ? (FIGURE 21) Pour ce qui est de l'utilisation de bourdonnement d'abeilles en tant que

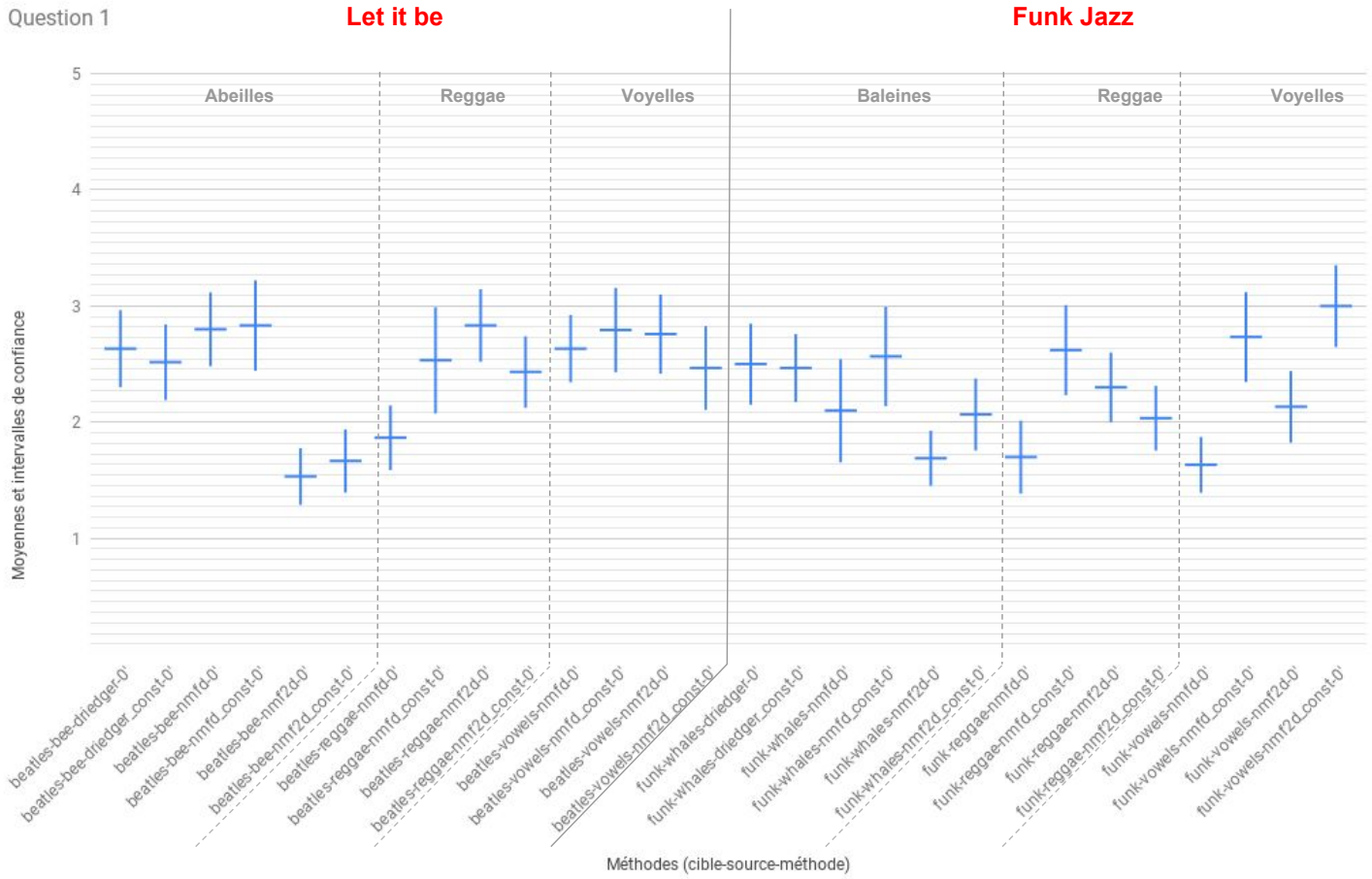

FIGURE 19 – Graphique des réponses à la question 1 : Comment jugez-vous la qualité du son ?

sons sources, la méthode de Driedger s'avère plus efficace que nos méthodes. En revanche l'utilisation de chant de baleine est jugée aussi efficace par toutes les méthodes. Le panel considère en majorité que la NMFD contrainte sur Funk Jazz est efficace dans la reconnaissance des sources Reggae. La reconnaissance des sons de voyelles chantée sort du lot pour l'utilisation du musaïcing par NMFD et NMF2D contraintes pour les deux sons cibles.

QUESTION 4 : COMMENT JUGEZ VOUS L'INTÉRÊT CRÉATIF DE L'EXEMPLE SONORE ? (FIGURE 22) Cette question est plus abstraite et ainsi plus relative à chaque auditeur. Cela se traduit sur l'observation des scores. Les intervalles de confiance se recouvrent en grande majorité. L'utilisation de voyelles en tant que source est tout de même plus appréciée pour son aspect créatif comme nous le montre les scores, plus particulièrement pour la NMFD contrainte et pour la NMF2D contrainte.

En recoupant ces différentes observations nous pouvons conclure que :

— la méthode de Driedger est à même de remplir ses objectifs à savoir la reconnaissance des sources et des cibles mais elle reste limitée à une source en entrée de l'algorithme.

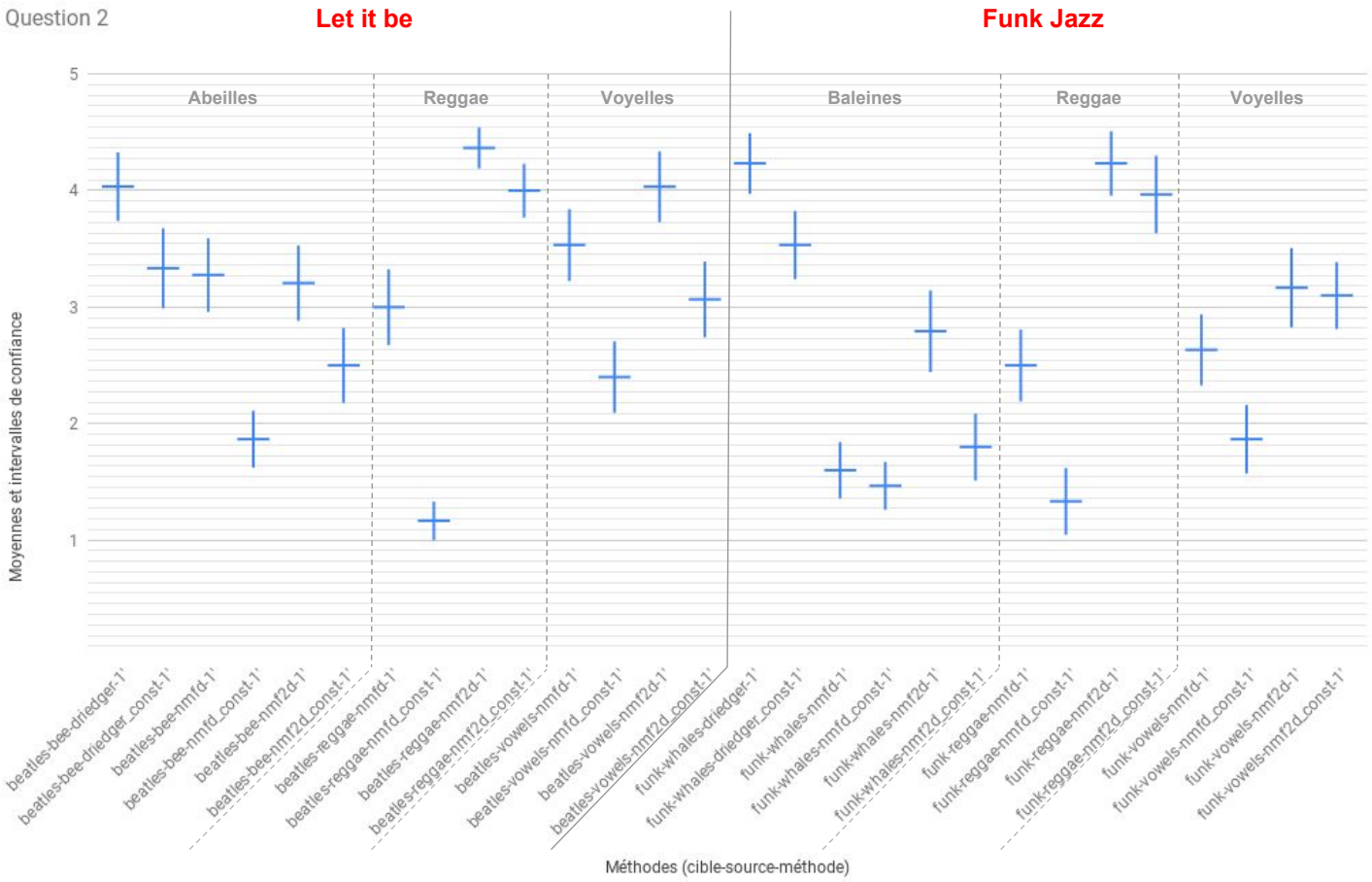

FIGURE 20 – Graphique des réponses à la question 2 : La structure temporelle et harmonique du son cible est-elle reconnaissable ?

— la NMFD possède de bon résultats sur la reconnaissance des sources également. En revanche qu'elle soit contrainte ou non, l'identification de la structure de la cible est plus délicate.
— la NMF2D possède de bons résultats moyens sur la reconnaissance de structure de la cible et sur la reconnaissance des sources également surtout dans le cas du multi-sources.

Il est évident que le sous-échantillonnage des signaux nuit à la qualité audio des exemples.

Pour ce qui est de la créativité des méthodes, les résultats de nos contributions (NMFD et NMF2D) sont plutôt satisfaisants. Particulièrement dans le cas du multi-sources avec les voyelles chantées que ce soit pour la NMFD et la NMF2D contraintes.

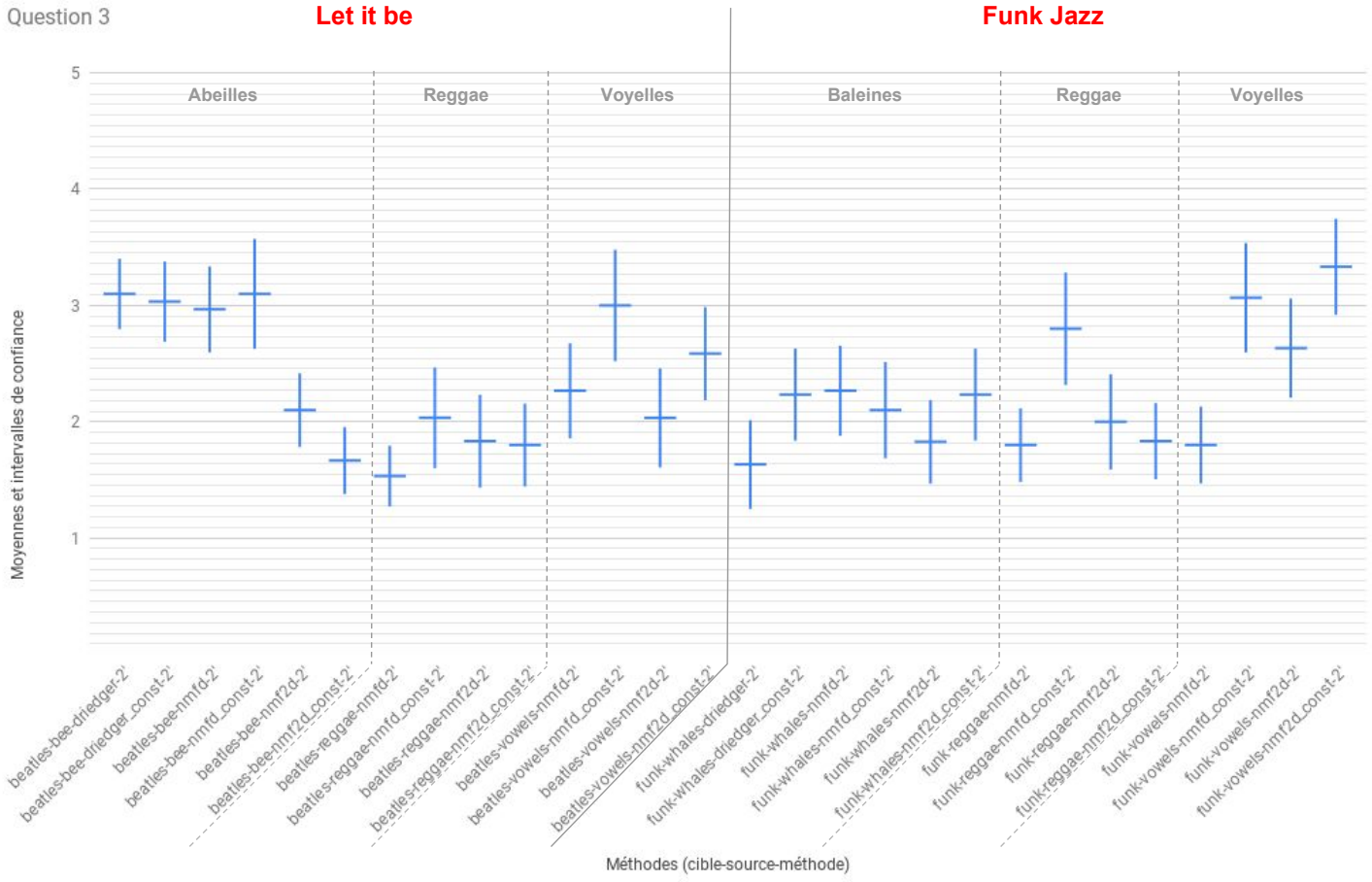

FIGURE 21 – Graphique des réponses à la question 3 : Le(s) son(s) source(s) sont ils identifiables(s) ?

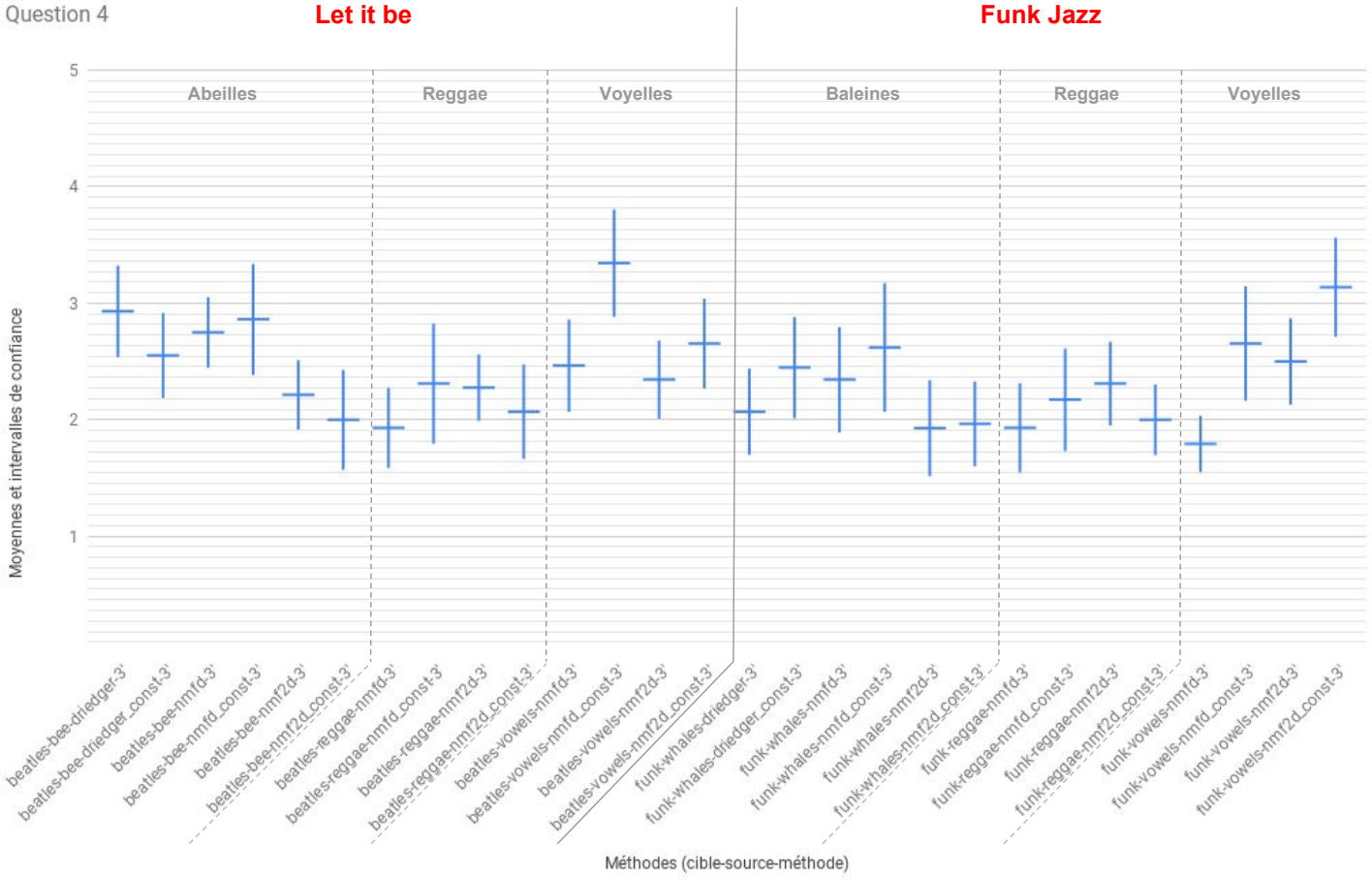

FIGURE 22 – Graphique des réponses à la question 4 : Comment jugez vous l'intérêt créatif de l'exemple sonore ?

7 CONCLUSION

7.1 DISCUSSION

Durant ce stage au sein de l'équipe analyse/synthèse, nous avons pu étudier en détails la documentation relative aux techniques de synthèse liées au musaïcing. Le musaïcing, rappelons le, est une technique de synthèse concaténative permettant de reproduire la structure temporelle d'un morceau cible avec des extraits sources possédant les caractéristiques acoustiques de la cible à chaque instant.

Cette étude approfondie nous a permis d'étendre la méthode de Driedger [8] ou encore de Burred [3] (voir chapitre 1).

A partir des méthodes décrites dans l'état de l'art nous avons apporté plusieurs contributions :

- Un musaïcing par NMFD permettant la conservation de la temporalité des bases sans contrainte de parcimonie diagonale et applicable au multi-sources.
- Un musaïcing par NMF2D permettant la transposition automatique des bases sources sur l'échelle chromatique lors de la reconstruction de la matrice de musaïcing.
- Un musaïcing prenant en entrée un signal préalablement divisé en partie harmonique et percussive grâce à l'algorithme d'HPSS de Fitzgerald [9].
- Nous pouvons ajouter l'utilisation de la CQT en tant que représentation temps/fréquence des matrices d'entrée de nos algorithmes.
- L'adaptation de l'algorithme de Griffin & Lim adapté à la CQT a lui aussi été une contribution nécessaire afin de pouvoir estimer le spectre de phase des signaux audio obtenus et ainsi créer notre panel d'exemples.

La figure 23 montre de manière simplifiée d'un coté les méthodes de l'état de l'art et de l'autre nos méthodes de musaïcing (par NMFD et NMF2D). Le schéma nous montre bien l'utilisation des méthodes de factorisation en matrices non-négatives et les opérations effectuées sur celle-ci pour mener à bien une synthèse "hybride"/concaténative.

Le score de chaque question relative à un exemple sonore généré par nos différentes méthodes est analysé grâce à un questionnaire perceptif. Les tests révèlent donc des résultats comparatifs intéressants entre la méthode de Driedger et la notre.

Pour résumer le panel d'auditeur ayant répondu au questionnaire a jugé globalement que : la méthode de Driedger (limitée au mono-source) mais permet une bonne identification de la cible et des sources (lorsqu'elles sont plus stationnaires), la NMFD contrainte ou non-contrainte permet une bonne identification des caractéristiques acoustiques des sources mais une mauvaise reconstruction de la structure de la cible, la NMF2D permet une très bonne reconstruction de la structure temporelle et harmonique de la cible ainsi qu'une bonne conservation des sons sources et particulièrement stationnaires tels que les voyelles chantées.

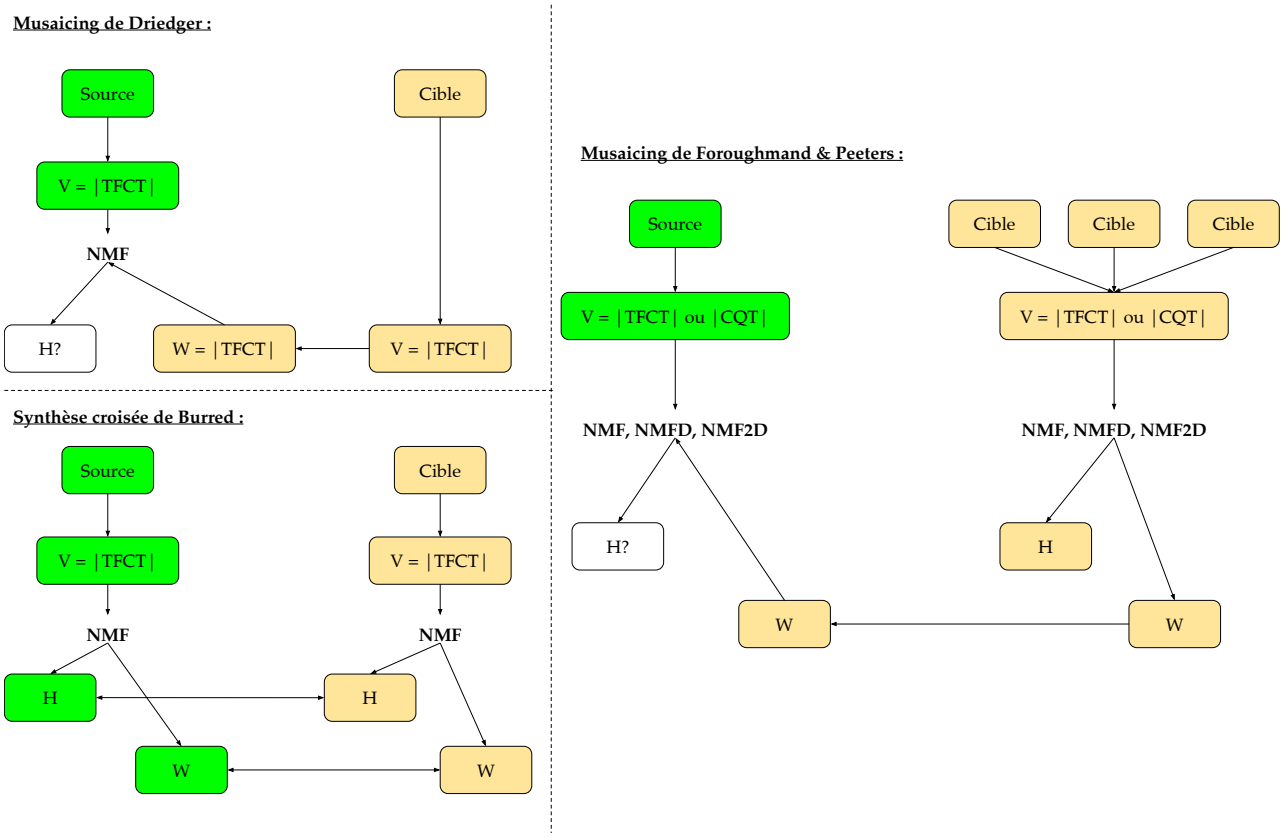

FIGURE 23 – Schéma comparatif entre les méthodes de l'état de l'art et les nôtres. Les différentes méthodes utilisées sont détaillées dans leur partie respective.

7.2 PERSPECTIVES

Nous pourrions étendre ce travail et principalement la méthode de musaïcing par NMF2D.

Le temps de calcul induit par la méthode a parfois limité l'étendue des tests possibles. Il serait tout d'abord intéressant de pouvoir effectuer des pré-apprentissage de bases sources sur des bases de données plus complètes et plus vastes. Nous pourrions également tester la méthode sans avoir recours à un pré-apprentissage et donc utiliser des extraits sources de longueur fixe (d'une manière similaire à la méthode de Laroche dans son article [12] qui utilise des signaux réel de batterie pour détecter les événements percussifs présents dans un morceau).

Des applications du même type que CataRT [19] (section 1.2.1) permettrait d'intégrer notre méthode de musaïcing au sein d'un logiciel interactif de création musicale.

BIBLIOGRAPHIE

[1] Jonathan Allen. « Short term spectral analysis, synthesis, and modification by discrete Fourier transform ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.3 (1977), p. 235–238.

[2] Judith C Brown et Miller S Puckette. « An efficient algorithm for the calculation of a constant Q transform ». In : *The Journal of the Acoustical Society of America* 92.5 (1992), p. 2698–2701.

[3] Juan José Burred. « A framework for music analysis/resynthesis based on matrix factorization. » In : *ICMC*. 2014.

[4] Graham Coleman, Esteban Maestre et Jordi Bonada. « Augmenting sound mosaicing with Descriptor-Driven transformation ». In : *Proceedings of DAFx*. 2010.

[5] Edward Costello, Victor Lazzarini et Joseph Timoney. « A streaming audio mosaicing vocoder implementation ». In : (2013).

[6] Roy De Maesschalck, Delphine Jouan-Rimbaud et Désiré L Massart. « The mahalanobis distance ». In : *Chemometrics and intelligent laboratory systems* 50.1 (2000), p. 1–18.

[7] Wim D'haes, Dirk Van Dyck et Xavier Rodet. « PCA-based branch and bound search algorithms for computing K nearest neighbors ». In : *Pattern Recognition Letters* 24.9 (2003), p. 1437–1451.

[8] Jonathan Driedger, Thomas Prätzlich et Meinard Müller. « Let it Bee-Towards NMF-Inspired Audio Mosaicing. » In : *ISMIR*. 2015, p. 350–356.

[9] Derry Fitzgerald. « Harmonic/percussive separation using median filtering ». In : (2010).

[10] Daniel Griffin et Jae Lim. « Signal estimation from modified short-time Fourier transform ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), p. 236–243.

[11] Jordi Janer et Maarten De Boer. « Extending voice-driven synthesis to audio mosaicing ». In : *5th Sound and Music Computing Conference, Berlin*. T. 4. 2008.

[12] Clément Laroche, Hélene Papadopoulos, Matthieu Kowalski et Gaël Richard. « Drum extraction in single channel audio signals using multi-layer non negative matrix factor deconvolution ». In : *ICASSP*. 2017.

[13] Daniel D Lee et H Sebastian Seung. « Learning the parts of objects by non-negative matrix factorization ». In : *Nature* 401.6755 (1999), p. 788–791.

[14] Beth Logan et al. « Mel Frequency Cepstral Coefficients for Music Modeling. » In : *ISMIR*. 2000.

[15] M. Mørup et M. N. Schmidt. *Non-negative Matrix Factor 2-D Deconvolution*. 2006. url : http://www2.imm.dtu.dk/pubdb/p.php?4521.

[16] Mikkel N Schmidt et Morten Mørup. « Nonnegative matrix factor 2-D deconvolution for blind single channel source separation ». In : *International Conference on Independent Component Analysis and Signal Separation*. Springer. 2006, p. 700–707.

[17] Christian Schörkhuber, Anssi Klapuri, Nicki Holighaus et Monika Dörfler. « A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution ». In : *Audio Engineering Society Conference : 53rd International Conference : Semantic Audio*. Audio Engineering Society. 2014.

[18] Diemo Schwarz. « A system for data-driven concatenative sound synthesis ». In : *Digital Audio Effects (DAFx)*. 2000, p. 97–102.

[19] Diemo Schwarz, Grégory Beller, Bruno Verbrugghe et Sam Britton. « Real-time corpus-based concatenative synthesis with catart ». In : *9th International Conference on Digital Audio Effects (DAFx)*. 2006, p. 279–282.

[20] Robert Silver. *Photomosaics*. http://www.photomosaic.com/.

[21] Paris Smaragdis. « Non-negative matrix factor deconvolution ; extraction of multiple sound sources from monophonic inputs ». In : *International Conference on Independent Component Analysis and Signal Separation*. Springer. 2004, p. 494–499.

[22] George Tzanetakis et Perry Cook. « Musical genre classification of audio signals ». In : *IEEE Transactions on speech and audio processing* 10.5 (2002), p. 293–302.

[23] Peter Welch. « The use of fast Fourier transform for the estimation of power spectra : a method based on time averaging over short, modified periodograms ». In : *IEEE Transactions on audio and electroacoustics* 15.2 (1967), p. 70–73.

[24] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*. T. 7. MIT press Cambridge, MA, 1949.

[25] Aymeric Zils et François Pachet. « Musical mosaicing ». In : *Digital Audio Effects (DAFx)*. T. 2. 2001, p. 135.