Evaluation de Flux Spectraux pour la Sélection de Temps d'Ancrage Robustes aux Dégradations Sonores

Rémi Mignot

Evaluation de Flux Spectraux

pour la Sélection de Temps d'Ancrage Robustes aux Dégradations Sonores

Rémi Mignot

IRCAM -CNRS, rapport interne, projet BeeMusic

21 décembre 2015

Résumé

Ce document traite de l'évaluation de différents flux spectraux, dans le cadre précis de la sélection de *temps d'ancrage* pour le calcul d'*empreintes sonores* en indéxation audio. Ici est évalué la robustesse de ces temps d'ancrage par rapport à d'éventuelles dégradations sonores, telles que : ajout de bruit synthétique ou environnemental, égalisation, filtrage, changement d'échelles temporelles ou fréquentielles.

1 Introduction

L'un des buts de l'indexation audio est la recherche d'extraits sonores donnés, parmi une base de références contenant un très grand nombre d'items sonores, qui sont des morceaux de musique par exemple. Cette opération est en générale basée sur le calcul d'*empreintes sonores* obtenues séparément sur ces différents signaux. Ainsi, après avoir construit une base de données d'empreintes sonores des signaux de référence, items, la recherche de l'extrait se fait en quelques sortes, par une comparaison des valeurs.

Dans le travail effectué dans le cadre du projet *BeeMusic*, chaque empreinte sonore est calculée pour un intervalle de temps d'environ deux secondes, et les points de départ de chaque fenêtre d'analyse sont donnés par ce que nous appelons dans ce travail *points d'ancrage*, ou *temps d'analyse*. Pour éviter un trop grand volume de données à calculer, puis à stocker ou à transférer, au lieu de décaler ces fenêtres d'analyse avec un pas petit, nous utilisons un décalage plus grand d'environ $\delta_T = 0.25$ [s].

Cependant, si les temps d'analyse utilisés pour l'extrait sonore à rechercher ne coïncident pas aux temps d'analyse utilisés pour le son correspondant, indéxé dans la base de référence, le décalage temporel engendré peut produire des modifications dans les valeurs des empreintes sonores. Cela peut avoir pour conséquence de dégrader les performances de la reconnaissance. Par exemple, en choisissant des temps d'analyse uniformément répartis avec un pas constant de δ_T, il peut survenir un décalage temporel systématique entre les temps d'analyse du flux à reconnaître et ceux de l'item de la base. Ce décalage étant compris entre $-\delta_T/2$ et $\delta_T/2$, soit 125ms ici. Par cette méthode, la seule manière de garantir une correpondance des temps d'analyse serait de contraindre l'extrait sonore à commencer à un temps multiple entier de δ_T, ce qui n'est pas envisageable en pratique.

Pour éviter ce possible retard ou avancement, Mathieu Ramona propose dans [14] de déterminer ces points d'ancrage par une analyse du signal audio lui-même, semblable à une détection d'évènements sonores (*onsets* en anglais). En musique, ces onsets sont le plus souvent positionnés là où les notes commencent, c'est-à-dire au moment de l'attaque. Ainsi, tout décalage temporel du flux de δ_τ engendre un décalage des onsets de la même valeur, et de même pour les temps d'analyse, puisque ceux-ci sont synchronisés par le signal lui-même. Dans [14], cette détection est basée sur la recherche des maxima de l'énergie instantanée du son, calculée sur une fenêtre glissante.

Cependant, nous nous intéressons ici à la reconnaissance d'extraits sonores, avec de possibles dégradations. Ces dégradations peuvent être de différentes natures : bruits additifs (synthétiques ou environnementaux), égalisation, filtrage, encodage (formats MP3 ou GSM par exemple), saturation, compression des dynamiques (mono ou multi-bandes), réverbération, changement d'échelles temporelle ou fréquencielle, etc. Malheureusement, une éventuelle altération sonore peut modifier la détection de ces onsets, et ainsi produire soit un nouveau décalage de ces temps d'analyse, entre l'extrait et la référence, soit des absences, soit des détections supplémentaires. Par exemple dans le premier cas, cette imprécision peut engendrer une modification des empreintes sonores, et par conséquent diminuer les performances de l'indexation audio, ce que nous cherchions à éviter justement.

Dans la littérature, la détection d'onsets est très souvent basée par un calcul intermédiaire d'un *flux spectral*. Ceci est une fonction du temps qui informe sur la variabilité du spectrogramme. En général, un flux spectral de valeur faible correspond à un signal localement stationnaire, ne contenant donc pas d'évènement sonore, et des fortes valeurs correspondent à un spectre évoluant dans le temps, éventuellement du à des onsets.

Le travail ici présenté a pour but de déterminer parmi un certain nombre de flux spectraux, celui qui donne les points d'ancrage les plus stables, robustes, aux dégrations sonores. De plus, chacun d'entre eux est *généralisé* à l'aide de plusieurs paramètres, et les valeurs de ces paramètres sont également testées. La performance de chaque flux spectral, et de chaque ensemble de valeurs de paramètres, est évaluée en comparant les temps d'ancrage définies sur des extraits sonores originaux non dégradés, et sur ces mêmes extraits sonores avec dégradation. Remarquons qu'il ne s'agit pas d'une détection d'onsets au sens strict, pour laquelle la *vérité de terrain* serait un ensemble d'instants annotés au préallable. Ici, la vérité de terrain est l'ensemble des points d'ancrage données par un flux spectral pour un signal original

sans dégradation. A ceux là sont comparés le points d'ancrage donnés par le même flux spectral, avec les mêmes paramètres, pour le même signal mais ayant subi un certain nombre de dégradations.

Cette étape de calcul des points d'ancrage est cruciale. Il est absoluement nécessaire de déterminer le flux spectral le plus robuste aux dégradations, parce qu'une mauvaise robustesse peut engendrer un décalage des points d'analyse et des modifications désastreuses des valeurs des empreintes sonores. Voir la figure 1 pour une illustration d'un changement du temps d'analyse. C'est pour cette raison que nous avons étudié un grand nombre de flux spectraux différents, avec de nombreux paramètres pour leur *généralisation* : 3 ou 4 paramètres par flux, plus 6 paramètres communs pour les calculs de spectrogrammes et le post-filtrage. Cela peut paraître excessif, mais un bon réglage des paramètres pourra améliorer significativement la robustesse de l'analyse et par conséquent les performances de l'indéxation.

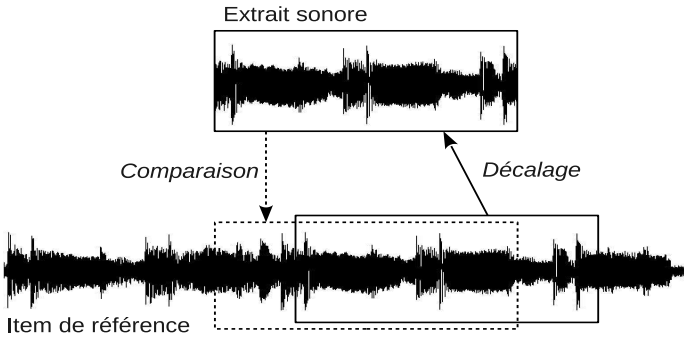

FIGURE 1 – Illustration de la modification du signal observé, par simple décalage temporel.

Ce document est organisé de la manière suivante : un résumé de la méthode de [14] est fait en section 2. Ensuite, la section 3 présente l'ensemble des paramètres communs relatifs au fenêtrage, au calcul du spectrogramme, et au post-filtrage servant de lissage des flux spectraux. Puis en section 4, l'inventaire des différents flux spectraux testés est fait. Notons que la plupart d'entre eux proviennent de la littérature sur le sujet de la détection d'onsets, et qu'ils sont ici tous *généralisés* grâce à l'utilisation de paramètres. La section 5 présente le protocole d'évaluation mis en oeuvre, ainsi que les résultats de performance de chacun des flux spectraux, et des valeurs de paramètres testées. Enfin la partie 6 concluera ce document.

2 Sélection des points d'ancrage de M. Ramona

Dans [13] Mathieu Ramona propose de synchroniser les points d'ancrage grâce à la détection d'onsets de [15]. Cependant, cet algorithme est relativement coûteux dans le cadre de l'indexation audio, ainsi dans [13], il propose une autre méthode similaire mais moins coûteuse basée sur la recherche de maxima d'énergie, et qui s'avére avoir des performances équivalentes. Cette section présente cette dernière méthode.

2.1 Energie spectrale

D'abord, le spectrogramme est calculée via une simple Transformée de Fourier à Court-Terme. Les fenêtre de Blackman sont utilisées pour la pondération avec une taille de 100ms, et un pas de 25ms. Puis, 6 sous-bandes fréquentielles sont définies sur lequels l'énergie est sommée. Ces bandes sont uniformément réparties dans l'échelle des barks, entre environ 500Hz et 1500Hz. Enfin, la fonction énergie *globale*, $E(n)$ où n est l'indice de la trame, est la moyenne des énergies des 6 bandes de fréquences.

2.2 Détection des maxima

Cette fonction énergie est dans un premier temps localement normalisée, en lui retirant sa médiane et en la divisant par son écart-type calculés avec une fenêtre glissante de taille 20 trames, soit 0.5s. Notons la $\overline{E}(n)$. Ensuite, un *filtre maximum*, semblable à un filtre médian, est appliqué. La sortie $P(n)$ du filtre maximum est la valeur maximale de $\overline{E}$ sur un intervalle de taille N trames autour de la trame n. La taille choisie du filtre est $N = 7$ trames, correspondant à 175ms. Enfin, les maxima locaux, donnant les points d'ancrage, correspondent aux trames ν pour lesquelles $\overline{E}(\nu) = P(\nu)$, voir la figure 2.

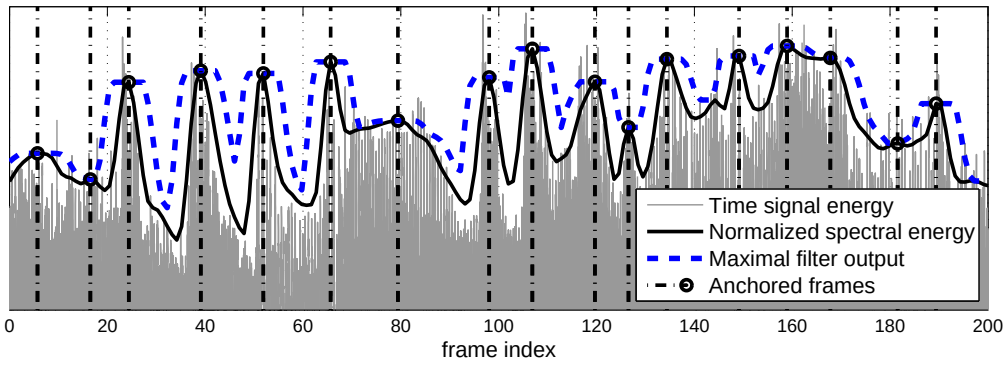

FIGURE 2 – Sélection des points d'ancrage, basée sur le sortie d'un filtre maximum de l'énegie spectrale.

Même si cette sélection des points d'ancrage semble satisfaisante, aucune étude poussée n'a été faite pour s'assurrer qu'il n'éxiste pas d'autres méthodes plus robustes aux dégradations sonores. Or, comparé au travaux de détection d'onsets, le problème est ici plus simple et donc facile à évaluer. En effet, pour la détection d'onsets, il est nécessaire d'avoir un base de signaux annotés de façon fiable. Or premièrement il est difficile et long d'obtenir une base de taille suffisante, et deuxièmement la définition même d'un onset n'est pas clairement définie. Ici, la vérité de terrain, avec laquelle nous comparont les points d'ancrage calculés sur un son dégradé, est simplement les positions des points d'ancrage obtenus sur le son original avec le même flux spectral et les mêmes paramètres. La constitution d'une grande base d'évaluation est alors très simple. Parmi une longue liste de flux spectraux définie en section 4, nous allons chercher en section 5 celui qui est le plus robuste aux dégradations.

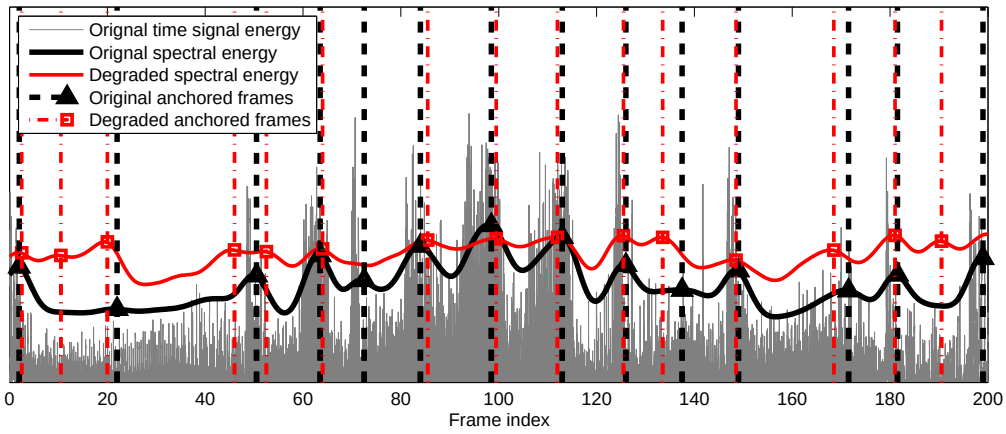

FIGURE 3 – Illustration de la sélection de points d'ancrage avec dégradation. Sur cet exemple, le point d'ancrage de référence à la trame 72 n'est pas détecté avec dégradation, un point d'ancrage est ajouté à la trame 10, et le point d'ancrage de la trame 139 est détecté avec un grand décalage.

3 Paramètres communs : spectrogramme et post-filtrage

Avant de présenter la liste des flux spectraux testés en section 4, nous présentons ici des paramètres communs qui sont également testés. Ces paramètres concernent premièrement le frenêtrage, le calcul du spectragramme, et le post-filtrage qui a pour but de lisser le flux spectral obtenu. De même, nous présentons ici le principe de *rectification en demi-ondes*, utilisé pour certains d'entre eux, qui a pour but à l'origine d'atténuer l'effet des extinctions de notes.

3.1 Fenêtrage et spectrogramme

Représentation temps/fréquences

Tous les flux spectraux ici présentés (à l'exception d'un), sont basés sur le calcul d'un spectrogramme, qui est une représentation temps-fréquences du son. Une fenêtre de pondération glissante permet l'observation de petits segments du signal. En déplaçant cette fenêtre d'un pas constant pour chaque trame, une transformée de Fourier discrète (dont l'acronyme anglais est DFT) est réalisée sur ce segment pondéré afin d'obtenir le contenu fréquentiel localement en temps. Voir par exemple [18, chap. 7]. Pour ce faire, nous avons besoin de définir : la taille w_s de la fenêtres d'analyse en secondes, le pas d'avancement (*hop size* en anglais) h_s en secondes, et la forme de la fenêtre.

Taille de fenêtre et pas d'avancement

Les deux paramètres w_s et h_s peuvent avoir un effet sur le flux spectral. En effet, w_s agit sur la résolution fréquentielle du spectre : une valeur élevée améliore globalement la résolution fréquentielle mais risque en contre-partie de *noyer* d'éventuels évènements sonores de très courte durée. Quant à h_s, une valeur faible améliore la résolution temporelle jusqu'à une certaine limite dépendant de w_s. Nous choisissons donc ces deux paramètres comme paramètres à tester.

Fenêtres asymétriques

Pour la forme de la fenêtre d'analyse, nous choisissons la fenêtre de Hann, souvent appelée à tort : fenêtre de *Hanning*. Nous pourrions aussi tester d'autres formes de fenêtres, tels que celles de [6], mais la fenêtre de Hann offre un bon compromis entre largeur du lobe principal et niveau des lobes secondaires, et elle est souvent choisie.

Cependant, nous avons observé pour certains flux spectraux, tel que la distance spectrale, cf. sec. 4.1, des maxima *doublées*. Effectivement, pour ce type de flux spectraux, lorque la fenêtre glissante rencontre un onset *saillant*, un premier maximum apparaît, ce qui est normal et souhaitable évidemment. Puis quand la fenêtre est centrée autour de cet onset, le flux spectral chutte, et remonte lorsque la fenêtre quitte l'onset. Il y a par conséquent un maximum de trop, cf. fig. 4a.

Une première solution consiste à appliquer une *rectification en demi-ondes* qui a pour but d'atténuer l'effet des extinctions de notes, cf. sec. 3.3, et donc supprimerait le second maximum. Cependant, ici nous ne cherchons pas les positions des évènements sonores en particulier, mais simplement des temps d'analyse quelconques, mais robustes aux altérations.

Nous avons donc implémenté une solution qui atténue le second maximum sans supprimer l'effet des extinctions de notes. Pour cela nous donnons la possibilité d'utiliser des fenêtres d'analyse asymétriques. Un paramètre $a \in [-1, 1]$ donne la position du maximum de la fenêtre par rapport à son centre. $a = 0$ correspond à la fenêtre de Hann symétrique, maximum en son centre, des valeurs négatives déplacent le maximum vers sa gauche, et des valeurs positives le déplacent vers sa droite. Voir la figure 5. Cette asymétrie est obtenue par interpolation linéaire de la fenêtre originale, symétrique.

Ainsi, en reprenant l'exemple de l'onset saillant, avec $a > 0$, le flux spectral présente un premier maximum plus marqué parce que l'*entrée* de l'onset dans la fenêtre y est plus brutal, et le second maximum est significativement réduit parce que la sortie de l'onset est plus douce. cf. fig. 4b.

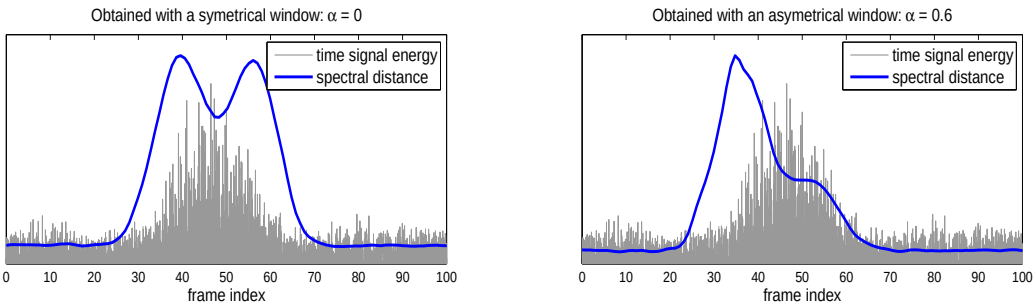

FIGURE 4 – Illustration du maximum double. Sur la figure de gauche, le maximum double apparaît, sans rectification et pour une fenêtre de pondération symétrique. Sur la figure de droite, un unique maximum apparaît, pour une fenêtre asymétrique avec un coefficient $a = 0.6$. Voir la figure 5 qui montre la forme de la fenêtre en fonction du coefficient.

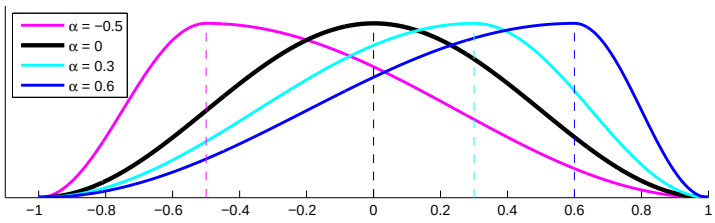

FIGURE 5 – Illustration de la fenêtre asymétrique de Hann. Notez qu'ici l'échelle des absisses est réglée de sorte à placer la fenêtre entre -1 et 1, si bien que la valeur du coefficient d'asymétrie donne directement la position du maximum.

Remarques et conclusion

Précisons que la fréquence d'échantillonnage F_s est fixée à 11025Hz qui est un bon compromis et qui est la valeur couramment utilisée en classification et indexation audio. Aussi, la taille de la DFT est automatiquement déterminée par la taille en échantillon W_s de la fenêtre d'analyse. Sa valeur M_s vaut 2 fois la première puissance de 2 supéreure ou égale à W_s. Par exemple, pour $w_s = 0.1$s, $W_s = 1102$ échantillons et $M_s = 4096$. On obtient une DFT de taille M_s par 0-padding du segment d'analyse.

Paramètres du fenêtrage : en conclusion les paramètres à tester sont
- ✓ w_s la taille de la fenêtre d'analyse en secondes,
- ✓ h_s le pas d'avancement de la fenêtre en secondes, et
- ✓ a le coefficient d'asymétrie de la fenêtre de pondération entre -1 et 1.

3.2 Post-filtrage

Pour rendre la sélection des temps d'acrange encore plus robuste aux dégradations sonores, nous avons testé la possibilité de lisser son évolution temporelle à l'aide d'un filtre linéaire numérique passe-bas. Avant de continuer, précisons que la fréquence d'échantillonnage des flux spectraux est $F_r = 1/h_s$ [Hz], nombre de trames par seconde, et non F_s.

Puisque ce post-traitement se réalise hors-ligne sur le flux complet, et donc sans contrainte de temps-réel ou de causalité, nous choisissons un filtre à moyenne ajustée symétrique et centré en $n = 0$. Par conséquent il est de retard de groupe nul, et les maxima sont lissés mais pas déplacés. Voici comment il est conçu : avec f_c la fréquence de coupure du filtre passe-bas, la réponse h_n du filtre idéal est

$$h_n = 2f_c \,\mathrm{sinc}(2f_c n/F_r),$$

et avec w_n une fenêtre de pondération symétrique, centrée en 0 et de taille $k+1$, la réponse du filtre réalisable de dimension fini est $b_n = h_n w_n$, ayant pour support $[-\frac{k}{2}, \frac{k}{2}]$. Nous prenons w_n une fenêtre de Hamming cf. [6]. Ce type de conception de filtre à moyenne ajustée est implémenté dans le logiciel Matlab par la fonction *fir1()*.

Le premier paramètre à tester est la fréquence de coupure f_c. Pour rendre le choix plus aisé, nous utilisons plutôt le temps de relaxation associé : $t_c = 1/f_c$, en seconde.

Ensuite, remarquons qu'un ordre k très élevé produirait des effets de Gibbs, évidemment pas souhaitable, et un ordre trop faible n'aurait aucun effet. Par conséquent, puisque le choix de k peut influer sur le résultat, et donc sur la robustesse, nous l'utilisons aussi comme paramètre à tester. Remarquons que k doit être pair.

Au préalable, le flux spectral ϕ_n est modifié avant filtrage par une caractéristique non-linéraire : $\widehat{\phi_n} = {\phi_n}^r$, avec $r > 0$ la puissance de la caractéristique. Une valeur de r inférieure à 1 a pour effet de *compresser* les valeurs du flux, et une valeur supérieure à 1 rend à l'inverse les maxima plus prédominants.

En résumé, le post-filtrage d'un flux spectral est réalisé de la manière suivante

$$\widetilde{\phi_n} = \Big(b \;*\; \phi^r\Big)(n) \tag{1}$$

Paramètres du post-filtrage : Les paramètres à tester sont
- ✓ t_c le temps de relaxation du filtre en secondes,
- ✓ k l'ordre du filtre (pair), et
- ✓ r la puissance de la caractéristique non-linéaire.

3.3 Rectification en demi-ondes

Comme nous verrons en section 4, beaucoup de flux spectraux utilisés en détection d'onsets sont basés sur une *distance* entre les spectres à court-terme de trames successives. Pour favoriser les attaques et atténuer les extinction de notes, un redressement de la différence des spectres est faite, cf. [4]. Les valeurs positives, pour des fréquences dont l'énergie croît, sont inchangées, et les valeurs négatives, pour une énergie décroissante, sont fixé à 0. On parlera ici de *rectification en demi-ondes*. Cette distance spectrale avec redressement $R\,()$, est donnée par

$$\phi_n = \sum_m R\left(X_{n,m} - X_{n-1,m}\right) \tag{2}$$

$$\text{avec} \quad R\left(X\right) = \tfrac{1}{2}\Big(X + |X|\Big) \tag{3}$$

Comme il est mentionné précédemment, nous ne cherchons pas à détecter les onsets en particulier, mais des points d'ancrage stables aux altérations sonores. Mais cette idée de redressement est tout de même utilisée et *généralisée* comme suit : plutôt que d'appliquer un redressement brutale qui annule les valeurs négatives, nous choisissons un redressement plus doux. Avec h le paramètre de valeur comprise entre -1 et 1, le redressement utilisé est

$$R_h\left(X\right) = \frac{1}{1+|h|}\Big(X + h|X|\Big) \tag{4}$$

Par exemple, pour $h = 1$, nous retrouvons le redressement de l'équation (3) qui annule les valeurs négatives, mais pour $h = 0.5$, les valeurs négatives sont cette fois-ci atténuées d'un facteur $1/3$ et les valeurs positives toujours inchangées. Pour $h = 0$, $R_h\,()$ est l'identité, et pour $h > 0$ les valeurs positives sont cette fois-ci atténuées et pas les valeurs négatives. Voir la figure 6 pour une illustration.

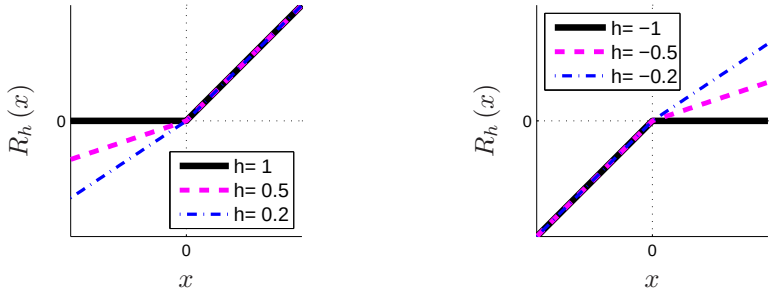

FIGURE 6 – Illustration du redressement de l'équation (4), en fonction du paramètre h de $R_h\,()$.

4 Liste des flux spectraux généralisés

Dans cette partie, seules les fréquences comprises entre 0 Hz et $F_s/2$ sont prises en compte. Ainsi les spectres discrets unilatéraux de taille $M = M_s/2 + 1$ sont obtenus par DFT et suppression des bins d'indice supérieur à M. Nous notons alors : X_n le vecteur de dimension $(M \times 1)$ du spectre à la trame d'indice n, et $X_{n,m}$ la valeur de la DFT à la fréquence indicée par $m \in [1, M]$.

Rappelons que les signaux temporels sont segmentés et leur spectrogramme X est calculé. Puis le flux spectral ϕ est calculé, et est enfin traité par le post-filtrage, ce qui donne le signal $\widetilde{\phi}$. Dans les sous-sections qui suivent, nous donnons la liste complète des flux spectraux testés, ϕ_n, avec leurs paramètres associés.

4.1 Distance Spectrale

La *distance spectrale* correspond à la norme ℓ_p de la différence des deux vecteurs de spectres de trames consécutives : $|X_n| - |X_{n-1}|$. Ce flux spectral est donc relatif à la distance

$$\| \, |X_n| - |X_{n-1}| \, \|_p = \begin{cases} \left(\sum_{m=1}^{M} \Big| |X_{n,m}| - |X_{n-1,m}| \Big|^p \right)^{1/p} & \text{si } p \neq +\infty, \\ \max_{m \in [1,M]} \Big| |X_{n,m}| - |X_{n-1,m}| \Big| & \text{si } p = +\infty. \end{cases} \tag{5}$$

Notons que la norme ℓ_1 est assez souvent utilisée en détection d'onsets, cf. [11]. Ici pour généraliser cette définition du flux spectral, l'ordre p n'est pas fixé, mais est choisi comme paramètre à tester.

Par ailleurs, cette distance est parfois *normalisée*. Cela consiste simplement à diviser la distance par la moyenne géométrique des normes des deux vecteurs, X_n et X_{n-1}. Dans ce travail, le choix de la normalisation se fera via un nouveau paramètre $d \in \{0, 1\}$ de la manière suivante :

$$\varphi = \frac{\| \, |X_n| - |X_{n-1}| \, \|_p}{(1-d) \; + \; d \, \sqrt{\|X_n\|_p \, \|X_{n-1}\|_p}}.$$

Cependant, avec $d = 1$ pour la version normalisée de la distance, cela peut engendrer de fortes perturbations pas toujours souhaitables. C'est par exemple le cas avec des signaux relativement faibles, en raison du produit des normes au dénominateur. Pour lisser le résultat, une idée est d'ajouter au dénominateur la somme des normes de X_n et X_{n-1}, pondérée par un nouveau paramètre $\beta \in [0, 1]$, cf. eq. (6).

Enfin, comme il est mentionné en section 3.3, nous appliquons l'opérateur $R_h\,()$ de rectification en demi-ondes sur la différence $|X_n| - |X_{n-1}|$, afin d'atténuer ou accentuer les attaques ou extinctions de notes, selon la valeur du paramètre $h \in [-1, 1]$.

Finalement, l'équation (6) donne la définition de la distance spectrale généralisée ϕ_n^{ds}.

Distance Spectrale :

$$\phi_n^{\text{ds}} = \frac{\| \, R_h \, (|X_n| - |X_{n-1}|) \, \|_p}{(1-d) \; + \; d \left(\sqrt{\|X_n\|_p \, \|X_{n-1}\|_p} + \beta \left(\; \|X_n\|_p + \|X_{n-1}\|_p \right) \; + \epsilon \right)}. \tag{6}$$

avec pour paramètres : l'ordre p de la norme, la paramètre $h \in [-1, 1]$ de la rectification en demi-ondes, ainsi que $d \in \{0, 1\}$ et $\beta \in [0, 1]$ pour le dénominateur.

Notons que ϵ est une constante fixe et petite évitant les division par zéro dans le cas de spectres nuls.

4.2 Corrélation Spectrale

La corrélation spectrale est basée sur le produit scalaire standard de deux spectres consécutifs :

$$\langle \, |X_n|, |X_{n-1}| \, \rangle = \sum_{m=1}^{M} |X_{n,m}||X_{n-1,m}|. \tag{7}$$

De la sorte, si les spectres de deux trames consécutives ont une forte dépendance linéaire, ils sont semblables et la corrélation est forte. Inversement s'ils sont quasi-orthogonaux, ils sont différents et la corrélation est porche de 0. Pour obtenir un flux spectral de fortes valeurs au moments des onsets, nous le définissons par, cf. [16, 5] :

Corrélation spectrale :

$$\phi_n^{\text{cs}} = 1 - \frac{\langle \, |X_n|, |X_{n-1}| \, \rangle}{\|X_n\|_2 \; \|X_{n-1}\|_2} \tag{8}$$

Maintenant, $\phi_n^{\text{cs}} = 1$ correspond à des vecteurs orthogonaux et $\phi_n^{\text{cs}} = 0$ à des vecteurs colinéaires, mais potentiellement de normes différentes.

Pour généraliser ce flux spectral, nous pourrions utiliser un produit scalaire généralisé défini pour tous vecteurs X et Y par : $\langle X, Y \rangle_A = XAY^T$, où $.^T$ est la transposition de matrices ou de vecteurs, et

A est une matrice de taille $(M \times M)$ symétrique définie positive. Cependant, nous ne l'avons pas fait. Par conséquent, la corrélation spectrale n'a pas de paramètre, autre que les paramètres communs définis en section 3.

4.3 Déviation de Phase

Avec $\psi_{n,m}$ la phase *déroulée* du spectrogramme, telle que $X_{n,m} = |X_{n,m}| \exp(j\psi_{n,m})$, la fréquence instantanée est donnée par sa dérivée $\psi'_{n,m}$ par rapport au temps. La méthode de la déviation de phase tire partie du fait qu'un onset a pour effet de faire fortement varier la fréquence instantanée, et qu'inversement, pour des signaux stationnaire à composantes sinusoïdales, les fréquences instantanées varient peu.

Ainsi, avec l'approximation de la dérivée seconde de la phase, $\psi''_n = \psi_{n-2} - 2\psi_{n-1} + \psi_n$, dans [1, 3] la déviation de phase est définie comme la norme ℓ_1 de ψ''_n :

$$\varphi_n = \frac{1}{N} \sum_{m=1}^{M} \left|\psi''_{n,m}\right| = \frac{1}{N} \left\|\psi''_n\right\|_1 \tag{9}$$

Cependant, cette définition ne tient pas compte du module, ce qui peut poser problème. Par exemple pour un signal harmonique stationnaire, même si la fréquence instantanée ne varie pas aux bins des lobes principaux, la présence d'un bruit faible, mais supérieur aux lobes secondaires, fait augmenter ψ'' entre les harmoniques, et donc la valeur totale de φ_n. Pour compenser ce problème, une pondération par le module de X a été proposé dans [2], et le flux spectral de la *déviation de phase pondérée* devient :

$$\varphi_n = \frac{\sum_{m=1}^{M} \left|X_{m,n}\ \psi''_{n,m}\right|}{\sum_{m=1}^{M} |X_{m,n}|} = \frac{\|\ X_m \otimes \psi''_n\ \|_1}{\|\ X_m\ \|_1}, \tag{10}$$

avec $\otimes$ représentant le produit élément par élément.

Pour généraliser la définition de ce flux spectral, premièrement, au lieu de forcer l'utilisation de la norme ℓ_1, nous donnons la possibilité de choisir l'ordre p de la norme, comme en sec. 4.1. Deuxièmement nous donnons la possibilité de choisir soit la déviation de phase simple, eq. (9), soit la déviation de phase pondérée, eq. (10). Ce dernier choix se fait par le paramètre a, cf. eq. (11) qui suit :

Déviation de phase :

$$\phi_n^{\mathrm{dp}} = \frac{1}{2\pi} \frac{\|\ (\xi_n + \mathrm{e}^{a}\, X_n) \otimes \psi''_n\ \|_p}{\|\ \xi_n + \mathrm{e}^{a}\, X_n\ \|_p}, \quad \text{où} \quad \xi_n = \frac{\|X_n\|_2}{\sqrt{M}}, \tag{11}$$

avec pour paramètres : l'ordre p de la norme, et a permettant de pencher soit pour la déviation de phase simple avec des valeurs négatives, soit pour la déviation de phase pondérée avec des valeurs positives.

Pour information, la dérivée seconde de la phase est approximée dans ce travail par : $\psi''_n = \psi_{n-1} - 2\psi_n + \psi_{n+1}$. Au lieu d'utiliser la phase des deux trames précédentes, nous utilisons ici les deux trames encadrant la trame courrante.

Remarquons que cette déviation de phase n'est valide que pour des signaux à composantes sinusoïdales, pour lesquels la phase des lobes principaux n'est pas perturbée par du bruit. Nous pouvons donc nous attendre à de mauvaises performances, cf. section 5.

4.4 Domaine Complexe

Le *Domaine Complexe* de [1, 2] consiste à faire une prédiction Y_n du spectre de la trame n à partir des spectres adjacents, et comparer cette estimation au spectre X_n réellement observé. Ce modèle de prédiction étant valide pour des signaux à composantes sinusoïdales stables, si l'erreur induite est forte la probabilité qu'un onset soit présent est forte.

La prédiction Y_n tient compte premièrement d'un module constant au cours du temps, ainsi $|Y_n| = |X_{n-1}|$. Deuxièmement, en considérant une fréquence instantanée pour chaque fréquence elle aussi constante au cours du temps en l'absence d'onset, la phase de Y_n est obtenue par déroulement de phase en fonction de la fréquence instantanée calculé aux trames adjacents. Avec $\psi'_{n,m} = \psi_{n-1,m} - \psi_{n-2,m}$ l'approximation de la fréquence instantanée du bin m à la trame n, la prédiction vaut :

$$Y_n = |X_{n-1}|\, \mathrm{e}^{\psi_{n-1} + \psi'_{n-1}}\,.$$

Dans [1, 2], le domaine complexe est alors donnée par la norme ℓ_1 normalisée de l'erreur de prédiction :

$$\varphi_n = \frac{\sum_{m=1}^{M} |X_{n,m} - Y_{n,m}|}{\sum_{m=1}^{M} |X_{n,m}|} = \frac{\|X_n - Y_n\|_1}{\|X_n\|_1}$$

Dans ce travail, nous généralisons ce flux spectral en introduisant les idées déjà mises en oeuvre pour la distance spectrale, mais ici pour définir une distance entre X_n et Y_n. Premièrement, nous testons plusieurs normes ℓ_p et non seulement la norme ℓ_1, deuxièmement nous considérons la recitification en demi-ondes R_h () de $X_n - Y_n$, via le paramètres $h \in [-1, 1]$, enfin nous donnons le choix de la normalisation par le paramètre d, et le paramètre β.

En conséquence, l'expression du domaine complexe généralisé est semblable à celle de la distance spectrale de eq. (6) avec Y_n au lieu de X_{n-1}, voir eq. (12).

Domaine Complexe :

$$\phi_n^{\text{dc}} = \frac{\| R_h (X_n - Y_n) \|_p}{(1-d) \; + \; d \left(\sqrt{\|X_n\|_p \, \|Y_n\|_p} + \beta \left(\|X_n\|_p + \|Y_n\|_p \right) \; + \epsilon \right)}. \tag{12}$$

avec pour paramètres : l'ordre p de la norme, le paramètre $h \in [-1, 1]$ de la rectification en demi-ondes, ainsi que $d \in \{0, 1\}$ et $\beta \in [0, 1]$ pour le dénominateur.

Comme pour la déviation de phase, la fréquence instantanée est approximée dans ce travail par la moitié de la différernce de phase des trames encadrant la trame courant, cad : $\psi'_n = \frac{1}{2} (\psi_{n+1} - \psi_{n-1})$.

Remarquons que cette méthode n'est aussi valide que pour des signaux à composantes sinusoïdales.

4.5 Difference des moments spectraux

Une idée pour étudier la variabilité d'un spectrogramme, et donc détecter d'éventuels onsets, est d'observer la différence trame à trame du barycentre spectral ou de l'*écart-type* spectral. En effet, en présence d'onsets, le spectre varie suffisamment pour modifier le contenu spectrale, et donc ces deux quantités. Inversement, pour un signal stationnaire, ni le barycentre ni l'écart-type ne change.

Afin de généraliser le nouveau flux spectral défini en eq. (15), nous considérons alors les moments spectraux de la puissance q du spectrogramme. Par exemple pour $q = 2$, $|X_{n,m}|^q$ n'est autre que l'énergie du spectrogramme. Le barycentre spectral $\mu_n^{(q)}$ et l'écart-type $\sigma_n^{(q)}$ sont alors définis par :

$$\mu_n^{(q)} \;=\; \frac{1}{\|X_n\|_q^q} \sum_{m=1}^{M} \omega_m \; |X_{n,m}|^q, \tag{13}$$

$$\sigma_n^{(q)} \;=\; \left(\frac{1}{\|X_n\|_q^q} \sum_{m=1}^{M} \left(\omega_m - \mu_n^{(q)} \right)^2 |X_{n,m}|^q \right)^{1/2}. \tag{14}$$

Et leurs évolutions dans le temps sont données par leur différence trame à trame :

$$\delta_\mu(n) = |\mu_n^{(q)} - \mu_{n-1}^{(q)}| \quad \text{et} \quad \delta_\sigma(n) = |\sigma_n^{(q)} - \sigma_{n-1}^{(q)}|.$$

Premièrement, au lieu d'utiliser $\delta_\mu(n)$ ou $\delta_\sigma(n)$ seuls, le flux spectral ϕ_n^{ms} est défini par une combinaison linéaire de ces deux fonctions. Nous introduisons alors le facteur e^α qui est le coefficient de pondération de $\delta_\sigma(n)$ dans la somme. Ainsi, une valeur négative de α favorise δ_μ, et une valeur positive favorise δ_σ. Deuxièmement, au lieu de calculer une simple somme, c'est la norme ℓ_p du vecteur $[\, \delta_\mu, \, \text{e}^\alpha \, \delta_\sigma \,]^T$ qui est considérée.

Nous obtenons par conséquent un nouveau flux spectral généralisé, relatif aux différences trame à trame des moments spectraux d'ordre 1 et 2. Cette fonction est définie en eq. (15).

Difference des moments spectraux :

$$\phi_n^{\text{ms}} = \frac{1}{1 + \text{e}^\alpha} \left(\; \left| \mu_n^{(q)} - \mu_{n-1}^{(q)} \right|^p \; + \; \text{e}^\alpha \; \left| \sigma_n^{(q)} - \sigma_{n-1}^{(q)} \right|^p \; \right)^{1/p} \tag{15}$$

avec pour paramètres : $q > 0$ la puissance du spectre, α le coefficient de pondération permettant de prendre en compte la variance spectrale, ou pas, et p l'ordre de la norme.

4.6 Norme spectrale

Dans [14], les maxima de l'énergie sont utilisés pour la détection d'onsets. Nous proposons ici de tester la même idée mais en considérant cette fois-ci la norme ℓ_p du spectre. Donc, avec la définition de la norme en eq. (5), le flux spectral est simplement donné par eq. (16).

Norme Spectrale :

$$\phi_n^{\text{ns}} = \|X_n\|_p. \tag{16}$$

avec pour paramètre : p l'ordre de la norme.

4.7 Difference de la norme spectrale

Cette fois-ci, il s'agit simplement de calculer la différence trame à trame de la norme spectrale de eq. (5) : $\delta_{\|.\|_p}(n) = |\ \|X_n\|_p - \|X_{n-1}\|_p\ |$. Contrairement à la norme spectrale, cf. sec. 4.6, ici les onsets sont détectés là où la norme du spectre varie le plus.

Toujours dans le but de généraliser ce nouveau flux spectral, nous reprenons les idées mises en oeuvres pour la distance spectrale, cf. sec. 4.1. Nous introduisons donc :

- La rectification en demi-ondes $R_h\,()$ de la différence $\delta_{\|.\|_p}$ des normes, de paramètre h. Notons qu'ici le redressement se fait sur $\delta_{\|.\|_p}(n)$ qui est scalaire et non un vecteur.
- Le choix de la normalisation ou pas, via le paramètre d.
- L'ajout de la somme des normes au dénominateur, avec coefficient β.

Par conséquent, le flux spectral associé à la différence de la norme spectrale est donné par eq. (17).

Difference de la norme spectrale :

$$\phi_n^{\text{dns}} = \frac{|\ R_h\ (\ \|X_n\|_p - \|X_{n-1}\|_p\ )\ |}{(1-d)\ +\ d\left(\sqrt{\|X_n\|_p\,\|X_{n-1}\|_p} + \beta\left(\ \|X_n\|_p + \|X_{n-1}\|_p\right) + \epsilon\right)}. \tag{17}$$

avec pour paramètres : l'ordre p de la norme, la paramètre $h \in [-1, 1]$ de la rectification en demi-ondes, ainsi que $d \in \{0, 1\}$ et $\beta \in [0, 1]$ pour le dénominateur.

4.8 Difference de la norme temporelle

En section 4.7 nous avons défini un flux spectral basé sur la différence de la norme. Cette norme y est calculée dans le domaine fréquentiel. L'idée simple exploitée ici est de réaliser la même différence de normes mais calculées dans le domaine temporel. L'un des avantages est de gagner du temps puisque la transformée de Fourier à court-termes n'est pas nécessaire.

Pour la généralisation du flux, nous introduisons les mêmes concepts : rectification en demi-ondes, et dénominateur paramétrè par d et β. Avec $\overrightarrow{x_n}$ le vecteur temporel de la trame indicé par n, pondéré par la fenêtre de pondération de taille W_s, cf. sec. 3.1, le flux spectral associé à la différence de la norme *temporelle* est donné par une expression équivalente à eq. (17), pour laquelle X_n et X_{n-1} sont remplacés par $\overrightarrow{x_n}$ et $\overrightarrow{x_{n-1}}$.

Difference de la norme temporelle :

$$\phi_n^{\text{dnt}} = \frac{|\ R_h\ (\ \|\overrightarrow{x_n}\|_p - \|\overrightarrow{x_{n-1}}\|_p\ )\ |}{(1-d)\ +\ d\left(\sqrt{\|\overrightarrow{x_n}\|_p\,\|\overrightarrow{x_{n-1}}\|_p} + \beta\left(\ \|\overrightarrow{x_n}\|_p + \|\overrightarrow{x_{n-1}}\|_p\right) + \epsilon\right)}. \tag{18}$$

avec pour paramètres : l'ordre p de la norme, la paramètre $h \in [-1, 1]$ de la rectification en demi-ondes, ainsi que $d \in \{0, 1\}$ et $\beta \in [0, 1]$ pour le dénominateur.

Evidemment, le résultat est le même dans le cas de la norme euclidienne ℓ_2 en raison du théorème de Parseval. Cependant pour $p \neq 2$, le résultat est différent.

4.9 IS-divergence Spectrale

En section 4.1, nous avons défini le flux spectral par une distance entre les vecteurs de spectres consécutifs. Nous proposons ici de remplacer la norme ℓ_p de la différence des spectres, par la divergence d'Itakura-Saito $\mathrm{D_{IS}}(X_n \| X_{n-1})$, donnée dans [7] par :

$$\mathrm{D_{IS}}(X_n \| X_{n-1}) = \frac{1}{M} \sum_{m=1}^{M} \left(\left| \frac{X_{n,m}}{X_{n-1,m}} \right| - \log \left| \frac{X_{n,m}}{X_{n-1,m}} \right| - 1 \right). \tag{19}$$

Cette divergence est une mesure de la dissimilarité de deux spectres, par conséquent elle peut être également utilisée pour définir un flux spectral.

Néanmoins, cette divergence n'est pas symétrique, dans le sens où pour A et B deux vecteur quelconques, $\mathrm{D_{IS}}(A\|B) \neq \mathrm{D_{IS}}(B\|A)$. Nous n'avons pas fait d'étude détaillée, mais nous pouvons supposer que cela a pour effet soit de favoriser les onsets, soit les fins de notes, comme c'est le cas pour la rectification en demi-ondes. Toujours dans le but de généraliser le flux spectral défini par la divergence d'Itakura-Saito, nous faisons ici une combinaison linéaire de $\mathrm{D_{IS}}(X_n\|X_{n-1})$ et de son symétrique : $\mathrm{D_{IS}}(X_{n-1}\|X_n)$. Avec γ comme paramètre, nous avons :

$$\varphi_n = \frac{1+\gamma}{2}\, \mathrm{D_{IS}}(X_n\|X_{n-1}) \;+\; \frac{1-\gamma}{2}\, \mathrm{D_{IS}}(X_{n-1}\|X_n). \tag{20}$$

Par exemple pour $\gamma > 0$, c'est la divergence directe $\mathrm{D_{IS}}(X_n\|X_{n-1})$ qui est favorisée. Evidemment, γ est l'un des paramètres à tester. De plus, nous proposons d'étudier la divergence non seulement du spectre $|X_{n,m}|$, mais plus généralement de sa puissance : $|X_{n,m}|^q$.

En résumé, le flux spectral généralisé associé à la divergence d'Itakura-Saito, est donné par eq. (21).

IS-divergence Spectrale :

$$\phi_n^{\mathrm{isd}} = \frac{1+\gamma}{2}\, \mathrm{D}_{\mathrm{IS}}^{(q)}(X_n\|X_{n-1}) \;+\; \frac{1-\gamma}{2}\, \mathrm{D}_{\mathrm{IS}}^{(q)}(X_{n-1}\|X_n), \tag{21}$$

$$\text{avec}\quad \mathrm{D}_{\mathrm{IS}}^{(q)}(A\|B) = \frac{1}{M} \sum_{m=1}^{M} \left(\left| \frac{A_m}{B_m} \right|^q - \log \left| \frac{A_m}{B_m} \right|^q - 1 \right). \tag{22}$$

avec pour paramètres : γ le coefficient de la somme et q la puisssance du spectre.

Remarquons que pour accélérer le calcul, une forme simplifiée de l'équation (21) a été développée :

$$\phi_n^{\mathrm{isd}} = \frac{1}{2M} \sum_{m=1}^{M} \left((1+\gamma) \left| \frac{X_{n,m}}{X_{n-1,m}} \right|^q + (1-\gamma) \left| \frac{X_{n-1,m}}{X_{n,m}} \right|^q + 2\gamma q \log \left| \frac{X_{n,m}}{X_{n-1,m}} \right| - 2 \right).$$

4.10 KL-divergence Spectrale

Une autre divergence bien connue est la divergence de *Kullback-Leibler*, cf. e.g. [8]. Elle est par exemple utilisée en *Factorisation de Matrices Non-négatives*, cf. [9]. Pour deux vecteurs A et B quelconques de dimension M, elle est donnée par :

$$\mathrm{D_{KL}}(A\|B) = \frac{1}{M} \sum_{m=1}^{M} |A_m| \log \left| \frac{A_m}{B_m} \right|$$

Nous proposons donc ici d'utiliser cette divergence pour définir un flux spectral comme cela a été fait en section précédente pour la divergence d'Itakura-Saito. Pour généraliser le flux, nous faisons également la somme pondérée de $\mathrm{D_{KL}}(X_n\|X_{n-1})$ et de son symétrique $\mathrm{D_{KL}}(X_{n-1}\|X_n)$, et nous considérons la puissance q du spectre.

En résumé, le flux spectral associé à la divergence de Kullback-Leibler, est donné par eq. (23).

KL-divergence Spectrale :

$$\phi_n^{\text{kld}} = \frac{1+\gamma}{2}\, \mathrm{D}_{\mathrm{KL}}^{(q)}(X_n \| X_{n-1}) \;+\; \frac{1-\gamma}{2}\, \mathrm{D}_{\mathrm{KL}}^{(q)}(X_{n-1} \| X_n), \tag{23}$$

$$\text{avec}\;\; \mathrm{D}_{\mathrm{KL}}^{(q)}(A\|B) = \frac{1}{M}\sum_{m=1}^{M} |A_m|^q \log \left|\frac{A_m}{B_m}\right|^q. \tag{24}$$

avec pour paramètres : γ le coefficient de la somme et q la puisssance du spectre.

Notons que pour accélérer le calcul, nous utilisons la forme

$$\phi_n^{\text{kld}} = \frac{q}{2M}\sum_{m=1}^{M} \left((1+\gamma)\,|X_{n,m}|^q + (1-\gamma)\,|X_{n-1,m}|^q\right) \log \left|\frac{X_{n,m}}{X_{n-1,m}}\right|.$$

4.11 KLn-divergence Spectrale

Toujours dans l'idée de définir un flux spectral pour une divergence de deux spectres consécutifs, nous proposons ici d'utiliser une forme normalisée de la divergence de Kullback-Leibler. Dans [17], elle est donnée pour deux vecteurs A et B par :

$$\mathrm{D}_{\mathrm{KLn}}(A\|B) = \frac{1}{M}\left[\sum_{m=1}^{M} |A_m| \log \left|\frac{A_m}{B_m}\right| + \log\left(\sum_{m=1}^{M} |B_m|\right)\right].$$

Ainsi, le flux spectral associé à la divergence de Kullback-Leibler normalisé est alors défini comme nous l'avons fait pour la divergence d'Itakura-Saito et la divergence de Kullback-Leibler, cf. sections 4.9 et 4.10.

KLn-divergence Spectrale :

$$\phi_n^{\text{klnd}} = \frac{1+\gamma}{2}\, \mathrm{D}_{\mathrm{KLn}}^{(q)}(X_n \| X_{n-1}) \;+\; \frac{1-\gamma}{2}\, \mathrm{D}_{\mathrm{KLn}}^{(q)}(X_{n-1} \| X_n), \tag{25}$$

$$\text{avec}\;\; \mathrm{D}_{\mathrm{KLn}}^{(q)}(A\|B) = \frac{1}{M}\left[\sum_{m=1}^{M} |A_m|^q \log \left|\frac{A_m}{B_m}\right|^q + \log\left(\sum_{m=1}^{M} |B_m|^q\right)\right]. \tag{26}$$

avec pour paramètres : γ le coefficient de la somme et q la puisssance du spectre.

4.12 I-divergence Spectrale

Cette fois-ci, nous proposons d'utiliser la forme dite généralisée de la divergence de Kullback-Leibler, aussi appelée : *I-divergence*, [17]. Elle est donnée pour deux vecteurs A et B par :

$$\mathrm{D}_{\mathrm{I}}(A\|B) = \frac{1}{M}\sum_{m=1}^{M} \left(|A_m| \log \left|\frac{A_m}{B_m}\right| - |A_m| + |B_m|\right)$$

De la même manière, le flux spectral associé à la I-divergence, est alors défini par eq. (27).

I-divergence Spectrale :

$$\phi_n^{\text{id}} = \frac{1+\gamma}{2}\, \mathrm{D}_{\mathrm{I}}^{(q)}(X_n \| X_{n-1}) \;+\; \frac{1-\gamma}{2}\, \mathrm{D}_{\mathrm{I}}^{(q)}(X_{n-1} \| X_n), \tag{27}$$

$$\text{avec}\;\; \mathrm{D}_{\mathrm{I}}^{(q)}(A\|B) = \frac{1}{M}\sum_{m=1}^{M} \left(|A_m|^q \log \left|\frac{A_m}{B_m}\right|^q - |A_m|^q + |B_m|^q\right). \tag{28}$$

avec pour paramètres : γ le coefficient de la somme et q la puisssance du spectre.

4.13 LP-divergence Spectrale

Dans [10], l'erreur de *prédiction linéaire* est mise sous forme fréquentielle. Avec P_m et $\widetilde{P_m}$ les densités spectrales de puissance du signal et du filtre prédictif, l'erreur s'écrit sous la forme :

$$E_{LP} = \frac{1}{M} \sum_{m=1}^{M} \frac{P_m}{\widetilde{P_m}}. \tag{29}$$

Nous proposons ici d'en définir une divergence, $\mathrm{D_{LP}}$, et de l'utiliser pour construire un nouveau flux spectral. Il est facile de vérifier dans eq. (29), que les fréquence où $P_m < \widetilde{P_m}$ contribuent d'avantage à augmenter l'erreur, que les fréquences où $P_m > \widetilde{P_m}$. Cela est notamment la raison pour laquelle la prédiction linéaire fonctionne relativement bien pour des spectres discrets. Dans ce travail, cela peut être utile pour favoriser les onsets, ou les fins de notes.

Le flux spectral associé à la divergence LP, est alors défini par eq. (30).

LP-divergence Spectrale :

$$\phi_n^{\mathrm{lp}} = \frac{1+\gamma}{2}\, \mathrm{D}_{\mathrm{LP}}^{(q)}(X_n \| X_{n-1}) \;+\; \frac{1-\gamma}{2}\, \mathrm{D}_{\mathrm{LP}}^{(q)}(X_{n-1} \| X_n), \tag{30}$$

$$\text{avec} \quad \mathrm{D}_{\mathrm{LP}}^{(q)}(A \| B) = \frac{1}{M} \sum_{m=1}^{M} \left| \frac{A_m}{B_m} \right|^q. \tag{31}$$

avec pour paramètres : γ le coefficient de la somme et q la puisssance du spectre.

5 Protocol d'évaluation et résultats

Pour rappel, dans ce travail nous utilisons l'approche de Mathieu Ramona, cf. [14] résumée en sec. 2 : premièrement, le spectrogramme est calculé sur le signal et l'un des flux spectraux donnés en sec. 4 est calculé. Puis, après le post-filtrage de la section 3.2, le *filtre maximum* expliqué en sec. 2.2 donne les points d'ancrage.

Dans ce document, nous cherchons le meilleur flux spectral, ainsi que les valeurs de ses paramètres, pour la robustesse de la sélection des points d'ancrage. Il ne s'agit pas ici de faire une optimisation de paramètres, qui serait trop coûteux. Nous avons en effet 3 paramètres communs pour le fenêtrage, environ 3 ou 4 paramètres pour chacun des 13 flux spectraux, et à nouveau 3 paramètres communs pour le post-filtrage. Alors, nous testons un certain nombre d'ensembles de valeurs, puis en fonction du résultat, d'autres tests sont relancés avec des paramètres ajustés manuellement, et en éliminant peu à peu les flux spectraux les moins robustes. Il s'agit donc plus d'une approche *essai/erreur*.

5.1 Procédure d'évaluation

5.1.1 Résumé

Pour les tests effectués, nous utilisons environ une centaine de morceaux choisis aléatoirement dans une base de données. Pour chacun d'entre eux, 10 secondes de signal sont extraites au milieu, et l'analyse retourne une valeur environ toutes les 250ms, dépendant de la taille du filtre maximum. En conclusion, environ 4000 points d'ancrage sont extraits des signaux pour l'évaluation.

Pour chaque flux spectral et chaque ensemble de valeurs de paramètres, les points d'ancrage sont calculés une première fois sur les signaux originaux sans altération. Cela constitue la *vérité de terrain*. Notons qu'ici cette vérité de terrain n'est pas *absolue* mais *relative* à un flux et à des valeurs de paramètres. Puis ces signaux originaux sont dégradés par différents types d'altérations et plusieurs degrés de dégradation, allant d'une altération légère à une altération plus forte. Les points d'ancrage sont alors recalculés sur ces nouveaux signaux. Enfin, les temps ainsi obtenus sont comparés à la vérité de terrain, donnée sur le son original avec le même flux et les mêmes valeurs de paramètres.

Pour l'évaluation des résultats, nous définissons plusieurs critères. Par exemple l'un d'entre eux est basé sur la F-mesure informant sur le nombre de bonnes détections et de mauvaises détections. Pour faciliter la visualisation des performances, une interface graphique a été developpée. Elle permet d'afficher les valeurs des critères obtenus, avec éventuellement un tri sélectif, pour montrer les meilleurs résultats, parmi des milliers calculés.

Les sections suivantes détaillent plusieurs des points précédemment décrits : altérations, critères, et interface graphique. En outre en section 5.1.2 nous donnons quelques éléments non exploités pour l'indexation mises en oeuvre par la suite, mais qui auraient pu être utiles avec une autre approche, finalement abandonnée.

5.1.2 Information supplémentaire

L'indexation audio mise en oeuvre dans le projet BeeMusic est basée sur le calcul d'empreintes sonores représentant des segments d'analyse long-terme de 2 secondes environ espacés toutes les 250ms environ, selon la taille du filtre maximum. Dans le cadre d'une autre approche, nous pensions baser les empreintes sonores sur une comparaison de spectres deux à deux, espacés d'environ 1 à 2 secondes. Sans entrer dans les détails, cette comparaison aurait permis de représenter par exemple les changements de notes ou d'accords, ce qui constitue une information particulièrement pertinente en musique.

Cependant, au lieu de chercher des points d'ancrage associés à des onsets, il aurait aussi été intéressant d'associer ces temps d'analyse à des parties stationnaires du signal, c'est-à-dire à des moments où les harmoniques des notes sont suffisamment visibles et stables, et non masqués par les transitoires d'attaques. Pour ce faire, il suffit de chercher les minima du flux spectral et non les maxima, en replaçant le filtre maximum de la section 3.2 par un *filtre minimum* reprenant le même principe.

Remarquons que c'est l'une des raisons pour laquelle nous avons introduit la généralisation de la rectification en demi-ondes avec le paramètre h, cf. sec. 3.3.

Evidemment, avec cette approche, la robustesse des points d'ancrage est tout autant nécessaire. Ainsi les évaluations mises en oeuvre dans ce travail ne traitent pas seulement les point d'ancrage associés à des onsets, maxima des flux spectraux, mais aussi des points d'ancrage associés aux parties stationnaires du signal, minima des flux spectraux. Ces deux types de point d'ancrage sont respectivement associés à ce que nous appelons : *transients* et *sustains*, termes anglophones signifiant *transitoires* et *entretiens*. En section 5.1.5, nous verrons que ces deux types de points d'ancrage sont premièrement calculés séparément, mais qu'une opération simple permet aussi de les lier.

Finalement, cette idée d'empreintes sonores basées sur des comparaisons de spectres deux-à-deux a été abandonnée en raison de la forte sensibilité au bruit. En conséquence, seuls les points d'ancrage associés aux onsets sont utilisé dans le projet BeeMusic.

5.1.3 Dégradations

Pour le test de robustesse aux altérations, nous avons réalisé plusieurs dégradations différentes des sons originaux. Ces dégradations, sont réalisées par la boîte à outils : *BeeAlter Toolbox*, cf. [12]. Pour obtenir un certain réalisme, nous avons mis à la chaîne plusieurs types de dégradations, avec à chaque fois une dégradation dominante. Ici sept types de dégradations ont été testés.

En outre, un coefficient α permet aussi de régler le niveau d'altération. Chaque dégradation a été réglée de sorte que $\alpha = 1$ produit une altération marquée mais raisonable, et $\alpha = 2$ produit une altération très forte, voir éxagérée. Cette section présente la liste des dégradations testées et l'effet de α.

Bruits ambiants
Il s'agit d'un bruit additif enregistré dans un restaurant à un moment d'affluence. Ici le coefficient α agit directement sur le rapport signal-à-bruit, dont l'acronyme anglais est SNR. Par exemple il vaut environ 3dB pour $\alpha = 1$.

Bruit rose synthétique
Ce bruit synthètique a un spectre de pente $-10dB$ par décade. Il est bien connu puisqu'il donne une impression perceptive de bruit blanc. De même, ici α agit sur le SNR.

Distorsion
Ici l'altération est donnée par un écrêtage des valeurs extrèmes, par rapport à une valeur de seuil donné par α. Par exemple pour $\alpha = 1$, le seuil est réglé de sorte à ce que 30 % des échantillons soient écrêtés.

Egalisation
Un égaliseur graphique est utilisé, pour lequel la courbe de gain est modifiée par α de la manière suivant : la réponse varie entre $\alpha[-15, +15]$ décibels.

Encodage MP3
Pour simuler la dégradation produite par le codage MP3, le son original est encodé puis décodé. Cette-fois le coefficient α agit sur le taux d'encodage.

Transposition fréquentielle
La transposition consiste à déplacer le contenu fréquentiel sans modifier l'échelle du temps comme le ferait un changement de vitesse de lecture. Ici les fréquences sont transposées vers les basses fréquences, avec une valeur de -2α demi-tons.

Effet *wow*
Cette altération est simulée par un filtre à retard variable pouvant représenter par exemple l'effet d'un désaccordage, *detuning*, ou un effet Doppler *cyclique*. Ici α agit directement sur la fréquence de modulation et l'amplitude du retard.

Comme il a été dit, une chaîne de dégradations a été mise en oeuvre, avec une dégradation dominante. Les dégradations listées précédemment sont tour à tour dominantes. Dans la suite, lorsque nous parlons du test d'une altération, il s'agit en réalité d'une altérations dominante, testée simultanément avec d'autres altérations faibles. Par exemple, pour l'ajout du bruit ambient, le son est au préalable altéré par une légère égalisation, un petit effet *wow*, et une faible transposition, puis un fort ajout de bruit est fait, qui est dans ce cas dominant. Cela permet de prendre en compte plusieurs dégradations simultanément, pour plus de réalisme, mais avec une dégradation plus marquée.

5.1.4 Critères

Cette section présente l'ensemble des critères permettant de comparer les points d'ancrage obtenus sur le signal original et ceux du signal dégradé. Deux types de critères sont mis en oeuvre ici : le premier type est basé sur un calcul de la F-mesure, le second est basé sur un calcul de distance spectrale.

F-mesure et mesure modifiée

La F-mesure est basée sur le calcul intermédiaire de la *précision* et du *rappel* qui correspondent au rapport du nombre de bonnes détections, *Vrais Positifs*, sur le nombre d'éléments à détecter d'une part, et sur le nombre total de détections d'autre part.

$$\text{Rappel} : \rho = \frac{N_{\text{VP}}}{N_{\text{VP}} + N_{\text{FN}}}, \quad \text{et} \quad \text{Précision} : \pi = \frac{N_{\text{VP}}}{N_{\text{VP}} + N_{\text{FP}}}. \tag{32}$$

où N_{VP} est le nombre de Vrais Positifs, donc le nombre de points d'ancrage inchangé après dégradations, $N_{\text{VP}} + N_{\text{FN}}$ est le nombre total de points d'ancrage du signal original, FN signifiant *Faux Négatifs*, et $N_{\text{VP}} + N_{\text{FP}}$ est le nombre total de points d'ancrage du signal dégradé, FP signifiant *Faux Positifs*.

D'une part, une valeur du rappel proche de 1 signifie que la plupart des éléments ont été détectés, mais potentiellement avec un certain nombre de Faux Positifs. D'autre part, une valeur de précision proche de 1 signifie que la plupart des détections sont bonnes, sans dire si tous les éléments ont été détectés. Bien entendu, le meilleur des cas correspond à une précision et un rappel tous deux proches de 1. Pour résumer cette information par un seul critère, il est souvent fait la moyenne *harmonique* de ces deux quantités. Cette *F-mesure* est définie par :

$$\text{F-mesure} : F = 2\frac{\rho\pi}{\rho + \pi}. \tag{33}$$

Pour déterminer le nombre de Vrais Positifs, donc le nombre de point d'ancrage inchangés après altération du son, nous considérons qu'un point d'ancrage du signal dégradé est correctement détecté s'il est placé à moins de $D/2$ d'un point d'ancrage du signal original. Pour une valeur faible de D, la F-mesure nous indique alors la robustesse de l'algorithme avec une forte sélectivité, alors que pour une valeur plus grande de D, la mesure accepte une sensibilité moins forte. Par exemple, avec ν^* l'indice de trame d'un point d'ancrage obtenu sur le signal original, et $\widetilde{\nu}$ l'indice de trame d'un point d'ancrage du signal dégradé le plus proche en temps de ν^*, on considére une bonne détection, vrai positif, si $|\widetilde{\nu} - \nu^*|h_s < D/2$. Sinon, le point d'ancrage en $\widetilde{\nu}$ est un faux positif. Dans les évaluations de la section 5.2, nous calculons la F-mesure pour 3 valeurs de D : 21ms, 42ms, et 84ms. Remarquons que même la valeur la plus grande de $D/2$ ne représente pas plus de 16% de l'écart moyen des points d'ancrage.

Sans donner de détails, nous proposons aussi une mesure similaire pour laquelle chaque détection est pondérée par sa distance normalisée au point d'ancrage original le plus proche. Cette mesure, que nous appelons bêtement la *G-mesure*, dépend aussi d'un paramétre D pour lequel nous prenons les mêmes valeurs. Remarquons que nous n'obervons pas de réelles différences avec la F-mesure.

Mesure basée sur le spectre

Dans le cadre de la sélection de points d'ancrage associés aux parties stationnaires du signal, sustains, au lieu de nous intéresser au concept de Vrais Positifs, définis ici par une erreur maximale de $D/2$, nous avons développé une mesure de similarité de spectres. En effet, dans ce cas là, la sensibilité temporelle de la détection a moins d'importance que la similarité des spectres. L'important est qu'au point d'ancrage donné après dégradation, le spectre soit le plus proche possible du spectre au point d'ancrage associé sans dégradation.

Avec ν^* l'indice de trame d'un point d'ancrage obtenu sur le signal original, et $\widetilde{\nu}$ l'indice de trame d'un point d'ancrage du signal dégradé le plus proche en temps de ν^*, cette nouvelle mesure correspond à la similarité entre $X_{\widetilde{\nu}}$ et X_{ν^*}. Dans ce travail, nous utilisons deux similarités : la première est basée sur une distance euclidienne de ces deux spectres, équivalente à la norme ℓ_2 de la section 4.1

$$\zeta_e = 1 - \frac{\| \, |X_{\widetilde{\nu}}| - |X_{\nu^*}| \, \|_2}{\sqrt{\|X_{\widetilde{\nu}}\|_2 \, \|X_{\nu^*}\|_2} + \|X_{\widetilde{\nu}}\|_2 + \|X_{\nu^*}\|_2} \tag{34}$$

et la seconde est basée sur la corrélation des spectres, équivalente à celle de la section 4.2

$$\zeta_c = \frac{\langle \, |X_{\widetilde{\nu}}|, |X_{\nu^*}| \, \rangle}{\sqrt{\|X_{\widetilde{\nu}}\|_2 \, \|X_{\nu^*}\|_2}} \tag{35}$$

Remarquons que pour ne pas introduire l'effet de la dégradation dans le calcul de ces similarités, le spectre dégradé n'est pas utilisé pour cette mesure. X_{ν^*} et $X_{\widetilde{\nu}}$ sont deux spectres du signal original, le premier à l'indice ν^* du point d'ancrage obtenu sur l'original, le second à l'indice $\widetilde{\nu}$ du point d'ancrage obtenu sur le signal dégradé.

5.1.5 Interface graphique et outils informatiques

Pour la visualisation des résultats, nous avons développé une interface graphique sous Matlab, via l'utilitaire GUIDE, qui affiche les différentes mesures pour plusieurs flux spectraux et valeurs de paramètres. Voir la figure 7 pour une vue d'ensemble de l'interface.

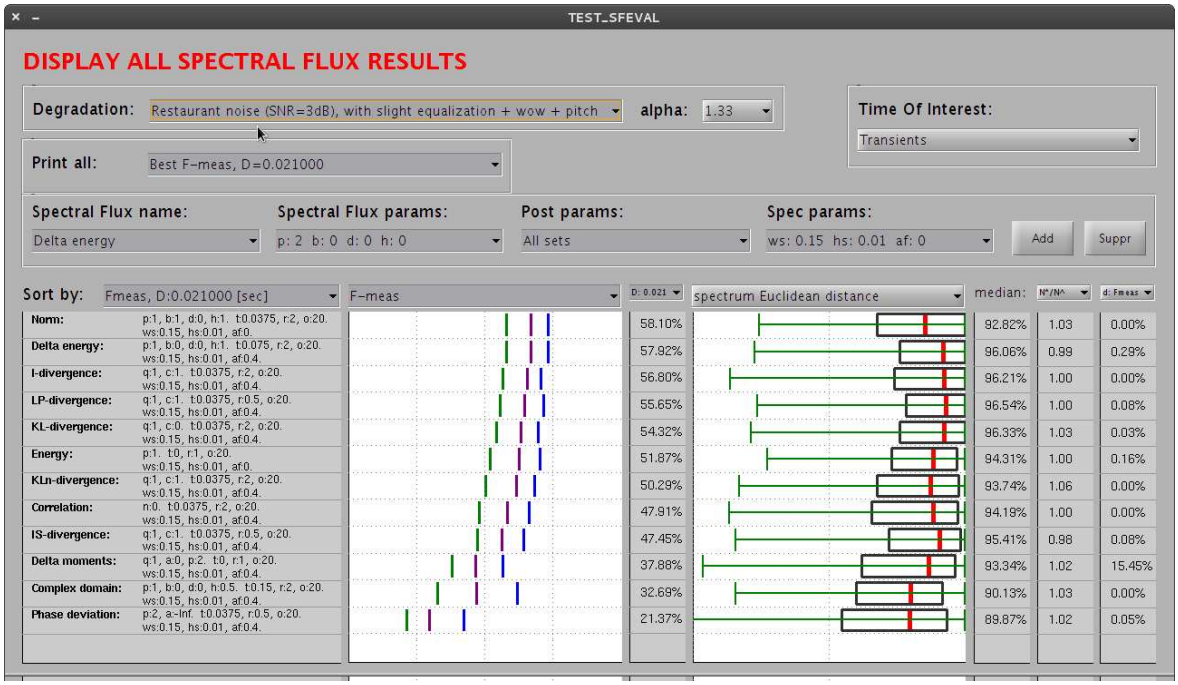

FIGURE 7 – Interface graphique pour l'affichage des résultats.

Détails de la fenêtre

La partie supérieure de la fenêtre permet de choisir le type de dégradation dominante, et le coefficient α pour la force d'altération. Voir la figure 8.

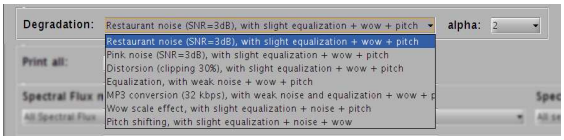

FIGURE 8 – Choix de la dégradation dominante et de α.

Le menu déroulant *Print All*, détaillé en figure 9, permet de sélectionner les résultats à afficher. Il est par exemple possible d'afficher pour un flux les résultats de tous les ensembles de valeurs de paramètres testés, ou bien de sélectionner pour chaque flux le meilleur ensemble de valeurs de paramètres selon une mesure à choisir : F-mesure, G-mesure, avec les différentes valeurs de D, ou bien la mesure de similarité donnée par la distance euclidienne ou par la corrélation spectrale, cf. sec. 5.1.4.

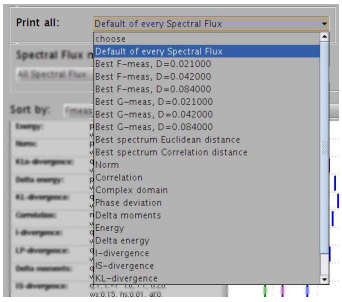

FIGURE 9 – Choix des flux spectraux à afficher.

La partie intermédiaire composée de quatres menus et de deux boutons, illustrée en figure 10, permet d'ajouter ou de retirer un ensemble spécifique de tests, à ceux déjà affichés. En partant de gauche, le premier menu permet de sélectionner le nom du flux spectral ; le suivant, l'ensemble des paramètres relatifs au flux choisi ; le troisième, l'ensemble de paramètre communs pour le post-filtrage ; et le dernier, l'ensemble des paramètres relatifs au calcul du spectrogramme. Les deux boutons permettent soit d'ajouter, soit de retirer les tests sélectionnés à la liste de la partie inférieure, cf. figs. 11 et 12. Remarquons que les noms des paramètres ont un peu changé par rapport aux noms donnés en section 4.

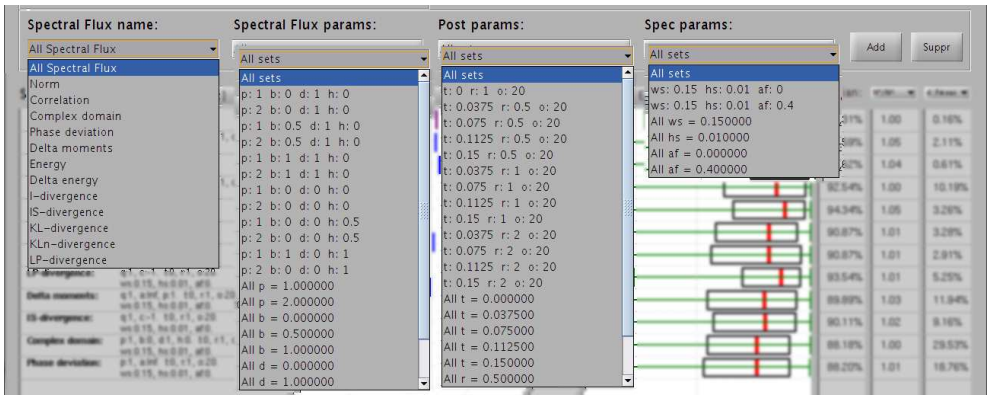

FIGURE 10 – Ajout spécifique de flux spectraux et paramètres. Remarquons qu'il n'est pas possible de dérouler simultanément tous les menus. Il s'agit ici d'un montage. Aussi, le contenu du menu nommé *Spectral Flux Params* dépend du flux spectral sélectionné par le menu *Spectral Flux Name*.

La partie inférieure gauche de la fenêtre affiche l'ensemble des tests sélectionnés, cad nom du flux spectral et valeurs des paramètres. Ils sont triés selon un ordre spécifié par le menu du dessus *Sort by*, cf. fig. 11. Le graphique à sa droite représente alternativement les valeurs de la F-mesure ou de la G-mesure,

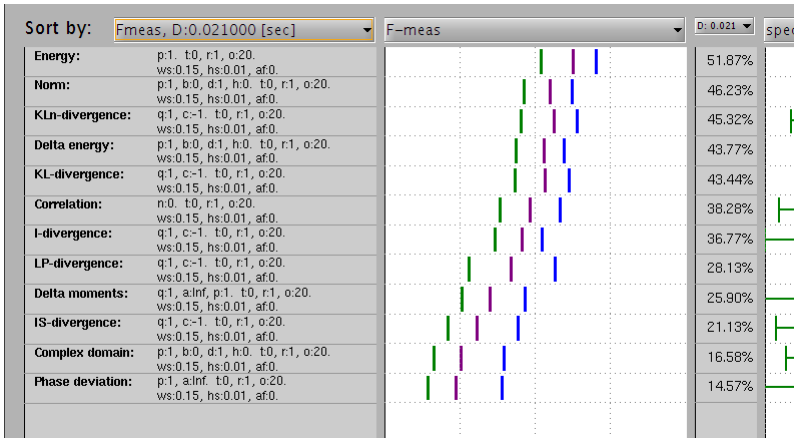

FIGURE 11 – Affichage des tests sélectionnés, et représentation de la F-mesure.

pour les trois valeur de D. Ce choix de mesure dépend du menu du dessus. La colonne suivante donne la valeur numérique de la mesure pour un D renseigné par le menu du dessus.

Le graphique de la partie suivante, cf. fig. 12, représente la distribution des mesures associées à la similarité des spectres. Les *boîtes à moustaches* représentent les centiles à : 5%, 25%, 50%, 75% et 95%. Par exemple le quantile de 50% donne la valeur médiane. Le menu du dessus permet de choisir entre la similarité basée sur la distance euclidienne ou bien la corrélation. La colonne à sa droite affiche la valeur numérique de la médiane.

Vient ensuite une colonne représentant la valeur du ratio : $N^*/\widetilde{N}$, où N^* est le nombre moyen de points d'ancrage attendu par seconde, égal à l'inverse de la taille du filtre maximum, et $\widetilde{N}$ est le nombre moyen de points d'ancrage réellement observés. Si ce ratio est très supérieur à 1, alors le test a sélectionner peu de points d'ancrage. De manière générale 1 est une valeur souhaitée. Aussi, cette même colonne peut afficher le ratio du nombre moyen de sélections entre le signal original et le signal dégradé. Cela permet de savoir si la dégradation produit plus ou moins de point d'ancrage. Encore une fois un ratio proche de 1 est préférable.

La dernière colonne affiche une mesure de similarité entre les points d'ancrage obtenus pour les transients et pour les sustains. En effet, il est nécessaire de s'assurer que ces deux distributions soient très différentes. Pour ce faire nous utilisons simplement la F-mesure ou la G-mesure déjà expliquées. Cette fois-ci la mesure doit être la plus proche de 0.

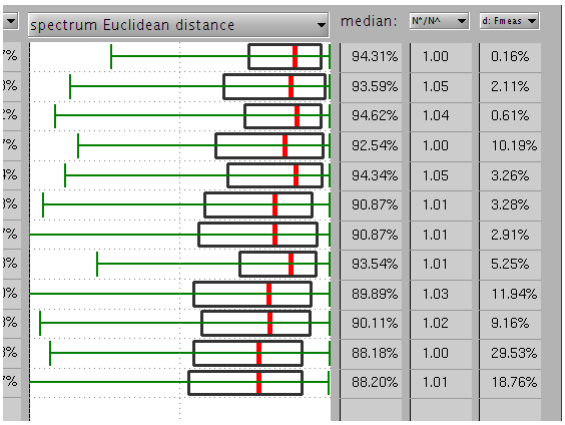

FIGURE 12 – Affichage de la similarité des spectres, *boîtes à moustaches*, valeurs de la médiane, ratio du nombre de sélections, et similarité des distributions de transients et de sustains. Remarque : ici l'échelle des absisses va de 0.5 à 1.

Enfin, la partie du haut à droite de la fenêtre, cf. fig. 13, permet de sélectionner l'affichage des évaluations de robustesse en considérant soit les transients soit les sustains. Cette partie se nomme *Time Of Insterest*, qui est l'ancien nom donné aux *points d'ancrage* Les deux premiers choix consistent aux résultats du calcul direct, basé respectivement sur les maxima ou minima des flux spectraux.

Les trois suivant correspondent à des méthodes de calculs légèrement différents. Le problème est que

pour l'approche expliquée en section 5.1.2, il est préférable d'avoir des points d'ancrage pour transients et sustains alternés. Ainsi pour le troisième choix, *Sustains (linked, min between transients)*, premièrement les point d'ancrage des transients sont calculés, et les sustains correspondent au minima entre chaque transients, ce qui impose alors l'alternance. Le quatrième choix correspond à la recherche de maxima entre sustains préalablement calculés. Quant au dernier choix, il s'agit de points d'ancrage placés exactement au centre de deux transients consécutifs.

Dans la pratique, on observe parfois des sustains placés à des instants où le signal est faible, et ce en général en fin de note peu avant une attaque. Malheureusement à ces moment, le signal est faible et le rapport signal-à-bruit est faible par conséquent, ce qui n'est pas souhaitable. C'est la raison pour laquelle nous avons également testé ce placement.

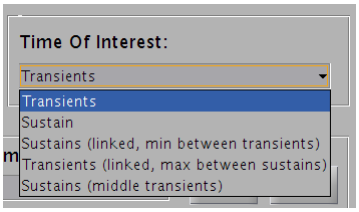

FIGURE 13 – Choix de type de points d'ancrage à afficher.

Code source implémenté

Le code source développé sous Matlab contient plusieurs scripts et fonctions. Nous en faisons un bref aperçu dans cette partie.

Dans un premier temps, les calculs de dégradation et de points d'ancrage sur plusieurs morceaux sont successivement lancés par le script `MAKE_SFEVAL`. Notons, qu'il a été lancé simultanément sur plusieurs ordinateurs de l'IRCAM, et que les résultats ont été stockés provisoirement sur un même disque.

Puis le script `MAKE_POSTEVAL` permet de rassembler les données enregistrées, grâce à la fonction `F_Merge_Results()`, et l'ensemble des statistiques, tels que la F-mesure et les similarités de spectren, est calculé par la fonction `F_Make_Stat()`. Une fois les statistiques obtenues, elles sont enregistrées dans un fichier.

Enfin, le script `TEST_SFEVAL` charge les résultats précédents et les affiche dans l'interface graphique précédemment détaillée.

Chacun de ces scripts utilise le script `COMMON_NAMES` qui rassemble les noms des différents dossiers et fichiers. De même les valeurs des paramètres utilisées pour les flux sont renseignées dans `PARAMS_SFEVAL`. Il fait appel à la fonction `F_Get_SFList()` qui contient l'ensemble des valeurs de paramètres de chaque flux. Le script `PARAMS_POSTEVAL` initialise les paramètres de l'analyse faite par `MAKE_POSTEVAL`. Enfin, la fonction `F_Get_Alterations()` retourne un tableau pour le réglage des altérations de la librairie *BeeAlter*.

Le répertoire *lib* contient plusieurs fichiers utiles :

- ✓ `spectanal()` Cette fonction se contente de calculer un spectrogramme.
- ✓ `F_specflux()` Cette fonction calcule le flux spectral sur un spectrogramme donné en entrée. Tous les flux spectraux donnés en sec. 4 sont codés ici. Le nom du flux spectral utilisé est donné en entrée, ainsi que les valeurs des paramètres.
- ✓ `F_sffilter()` Cette fonction réalise le post-filtrage du flux spectral. Elle prend en entrée le flux spectral ϕ et les paramètres du filtrage, t_c, k et r, et retourne le flux ϕ', voir sec. 3.2.
- ✓ `F_Compute_SFTOI()` Cette foncion prend en argument un spectrogramme, d'un son original ou altéré, ainsi que les valeurs des paramètres testées, et retourne une struture contenant toutes les positions des point d'ancrage. Pour rappel, les points d'ancrage sont nommés dans le code : *Time Of Interest* ayant pour acronyme TOI. Aussi l'acronyme SF signifie *Spectral Flux*.
- ✓ `F_fmeasure()` Calcul de la F-mesure, à partie des distributions des points d'ancrage obtenus sur le signal original et les points d'ancrage obtenus sur le signal dégradé.
- ✓ `F_gmeasure()` Il s'agit cette fois-ci de la G-mesure.
- ✓ `F_Compute_SpectSimil()` Cette fonction calcule les deux similarités de spectres donnés basées sur la distance euclidienne et la corrélation spectrale. Cf. eqs. (34) et (35).
- ✓ `F_toitimeproxim()` Cette fonction calcule la G-mesure pour évaluer la proximité des deux distributions correspondant aux tansients et aux sustains, voir la remarque sur la toute dernière

colonne de l'interface.

- ✓ `F_Merge_Results()` Cette fonction se charge de recharger les structures calculées par la fonction précédente, pour chaque morceau un à un, et concatène les point d'ancrage correspondant à une même dégradation, un même flux spectral et un même ensemble de valeurs de paramètres. Cette fonction est lancée par `MAKE_POSTEVAL`.
- ✓ `F_Make_Stat()` Cette fonction réalise l'ensemble des mesures sur les points d'ancrage obtenus. Cette fonction est aussi lancée par `MAKE_POSTEVAL`.
- ✓ `plotsf` Il s'agit de la classe correspondant à l'affichage des résultats pour un test. Un objet de cette classe contient et modifie les données correspondant à une ligne de la partie inférieure de l'interface.
- ✓ `handleData` Classe permettant l'encapulation données dans un handle. Cela simule un pointeur de structure, comme en langage C, sans avoir besoin de déclarer la structure avant, comme c'est possible en Matlab.
- ✓ `TEST_SpecFlux` Ceci est un script pour tester le calcul des flux spectraux. Quelques exemples y sont implémentés, tels que celui de la figure 14.
- ✓ `test_getsflist` Fonction semblable à `F_Get_SFList()`, mais utilisée pour le test précédent.
- ✓ `F_toispecsimil()` Fonction probablement obsolète, utilisée pour le calcul des deux similarités de spectres.
- ✓ `sfeval_details()` Sous-interface graphique utilisée par `TEST_SFEVAL`.

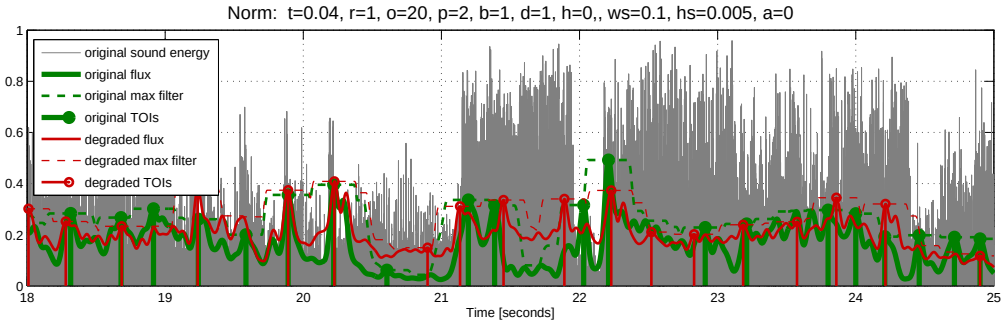

FIGURE 14 – Illustration de la sélection de pont d'ancrage. Ici la norme est utilisée, les valeurs des paramètres sont donnés en titre.

5.2 Résultats

Le choix du *meilleur* flux spectral et des *meilleures* valeurs de paramètres, se fait ici empiriquement par visualisation graphique des résultats. Comme il a été dit précédemment, il s'agit d'une approche *essai/erreur*. Une première étude a permis de conserver les flux visiblement les plus robustes, et d'autres ont ensuite permis de restreindre encore le nombre de flux candidats, mais avec un nombre d'ensembles de paramètres plus grand.

Remarquons que le choix s'est fait de façon *subjective* en considérant simultanément : de bonnes valeurs des 4 mesures (F-mesure, G-mesure, et les 2 similarités de spectres), une relative stabilité des résultats quelque soit les valeurs de paramètres, et un bon ratio $N^*/\widetilde{N}$ informant sur le nombre moyen de points d'ancrage sélectionnés, cf. sec. 5.1.5. Dans le cas des points d'ancrage associés aux sustains, pour l'approche abandonnée et expliquée en sec. 5.1.2, il est aussi vérifié une bonne dissimilarité des 2 distributions correspondant aux transients et aux sustains, cf. la dernière colonne de la fenêtre.

5.2.1 Premier lot de tests

Un premier lot de tests a été lancé sur tous les flux spectraux et avec un certain nombre d'ensembles de valeurs de paramètres. Dans ce premier lot de tests, nous avons choisi 2 ensembles de valeurs de paramètres pour le spectrogramme, une dizaine en moyenne pour chacun des flux spectraux, et 12 pour le post-filtrage, ce qui donne environ 3200 tests. A cela s'ajoute les différentes dégradations testées, seulement 3 sur les 7 (bruits ambiants, égalisation et effet wow), pour 2 valeurs de α : 1.33 et 2, ce qui représente de très fortes altérations. Notons qu'ici, tous les flux de la section 4 ont été testés, excepté la *différence de norme temporelle* qui a été testée en dernier.

Par exemple ici les valeurs des paramètres communs sont : $w_s = 150$ms, $h_s = 10$ms, $a \in \{0, 0.4\}$, $k = 20$, $t_c \in \{37.5, 75, 112.5, 150\}$ms, $r \in \{0.5, 1, 2\}$.

Résultats : Selon l'étude empirique des résultats, basée sur les critères donnés plus haut, ce premier lot de tests a d'ores et déjà permis d'éliminer un bon nombre de flux spectraux : la déviation de phase, le domaine complexe, la corrélation spectrale, la différence des moments spectraux, la norme spectrale, ainsi que la divergence d'Itakura-Saito, la divergence de Kulback-Leibler normalisée, et la LP-divergence. Restent alors en lice : la **distance spectrale** ϕ^{ds} de la section 4.1, la **différence de la norme spectrale** ϕ^{dsn} de la section 4.7, la **KL-divergence** (Kulback-Leibler non normalisée) ϕ^{kl} de la section 4.10, et la **I-divergence** ϕ^{id} de la section 4.12.

5.2.2 Deuxième lot de tests

De nouveaux tests ont été lancé sur les quatres flux retenus : distance spectrale, différence de la norme spectrale, KL-divergence, la I-divergence. Cette fois-ci les 7 dégradations ont été appliquées, avec $\alpha = 2$ uniquement, 8 ensembles de valeurs pour le spectrogramme, 3 pour le post-filtrage, et 12 ou 15 pour chaque flux spectral. Ici, $w_s \in \{130, 175\}$ms, $h_s \in \{8, 12.5\}$ms, $a \in \{0, 0.4\}$, $k = 20$, $t_c \in \{37.5, 75\}$ms, $r = 1$. Ici nous avons testé davantage de valeurs de paramètres, excepté pour t_c parce que on peut observé d'assez mauvais résultats pour $t_c > 100$ms.

Résultats : Cette fois-ci le choix est plus délicat, mais nous avons une petite préférence pour : la **différence de la norme spectrale** et la **KL-divergence**.

5.2.3 Troisième lot de tests

Enfin, nous avons lancé un dernier lot de tests pour départager les deux derniers flux retenus : la différence de la norme spectrale et la KL-divergence. Ici 4 dégradations ont été testées (bruits ambiants, distortion, égalisation, et effet wow), pour $\alpha \in \{1.33, 2\}$, 18 ensembles de valeurs pour le spectrogramme, 9 pour le post-filtrage, 16 pour le différence de la norme spectrale et 15 la KL-divergence. Ici, $w_s \in \{100, 150, 200\}$ms, $h_s \in \{5, 10, 20\}$ms, $a \in \{0, 0.5\}$, $k = 20$, $t_c \in \{0, 25, 50, 100, 150\}$ms, $r \in \{1, 2\}$. Nous donnons les valeurs testées pour différence de la norme spectrale : $p \in \{1, 2\}$, $h \in \{-1, -0.5, 0, 0.5, 1\}$, $d \in \{0, 1\}$, et $\beta \in \{0, 0.5, 1\}$; et pour la KL-divergence : $q \in \{1, 2, 3\}$ et $\gamma \in \{-1, -0.5, 0, 0.5, 1\}$.

Résultats : Malheureusement, nous n'avons pas pu *départager* ces deux derniers flux. Ils ont des résultats équivalents pour la plupart des critères fixés. Cependant, on peut remarquer que la différence de la norme spectrale est un peu moins coûteuse que la KL-divergence notamment en raison du logarithme. Donc pour y gagner un peu en ressources CPU, la différence de la norme spectrale semble un bon choix. Nous donnons maintenant les valeurs des paramètres des meilleures résultas obtenus :

Différence de la norme spectrale :
Les meilleures valeurs de paramètres obtenues sont : $p = 1$, $d = 0$, $\beta = 0$, $h = 1$, $t_c = 50$ms, $k = 20$, $h_s = 10$ms, $a = 0.5$, $w_s = 150$ ou 200 ms et $r = 1$ ou 2. L'expression du flux spectral se simplifie alors en

$$\phi_n^{\text{dns}} = \mid R_1 \left(\|X_n\|_1 - \|X_{n-1}\|_1 \right) \mid . \tag{36}$$

KL-divergence :
Les meilleurs valeurs de paramètres obtenues sont : $q = 1$, γ quelconque, $t_c = 50$ms, $k = 20$, $h_s = 10$ms, $a = 0.5$, $w_s = 200$ms et $r = 1$. Ici nous observons que la valeur de γ n'agit pas sur la robustesse, si bien qu'avec $\gamma = 1$, l'expression du flux se simplifie en

$$\phi_n^{\text{kld}} = \frac{1}{M} \sum_{m=1}^{M} |X_{n,m}| \log \left| \frac{X_{n,m}}{X_{n-1,m}} \right| . \tag{37}$$

5.2.4 Remarques

Différence de norme : temporelle vs. fréquentielle. A l'origine nous n'avions pas testé la différence de norme *temporelle*. En voyant les bons résultats de la différence de norme *spectrale*, nous avons imaginé

une version calculée dans le domaine temporel afin de réduire le temps de calcul. Cependant un dernier lot de tests pour comparer ces deux flux, a montré que la différence de norme *temporelle* a des performances significativement inférieure quelque soit les valeurs de ses paramètres, et ainsi elle a été rejetée. Remarquons que puisque la meilleure valeur de l'ordre p de la norme est 1, le théorème de Parseval ne s'applique pas, ce qui explique des performances différentes entre les deux flux.

Différence de norme fréquentielle pour les sustains. Dans le cas des points d'ancrage associés aux sustains, pour l'approche abandonnée expliquée sec. 5.1.2, nous remarquons que la meilleure valeur de h, associé à la rectification en demi-ondes, pour la différence de norme *spectrale* est 1. Cette valeur de h a pour effet d'annuler le flux en dehors des onsets, transients. Par conséquent la différence de norme *spectrale* est à éviter pour les sustains associé aux minima du flux, on préfèrera alors la KL-divergence.

Liaison des sustains et des transients. Toujours, dans le cas des points d'ancrage associés aux transients et au sustains, si l'alternance est requis, nous constatons que le fait de placer les points d'ancrage pour les sustains exactement au centre de 2 maxima du flux, transients, donne d'assez bons résultats. En plus d'améliorer le rapport signal-à-bruit, ce principe permet d'utiliser le différence de norme fréquentielle, parce que ce sont les maxima qui sont cherchés, et non les minima, cf. paragraphe précédent.

Domaine complexe. Au moment de la rédaction de ce document, je me suis rendu compte d'une erreur de programmation pour le domaine complexe. Le problème vient du fait que la rectification ne fonctionne pas des nombres complexes. Le résultat est que quelque soit la valeur entrée de h, c'est $h = -1$ qui est effectivement utilisé. Cependant, vu les résultats obtenus, nous pouvons supposer que la correction de ce problème ne changera pas la conclusion.

6 Conclusion

Dans ce travail nous avons généralisé le concept de flux spectral, et en avons défini un grand nombre, paramétrables par un très grand nombre de paramètres. Le but est d'obtenir une sélection de points d'ancrage pour l'analyse d'empreintes sonores, la plus robuste possible à toutes forme de dégradations sonores. L'idée de la sélection de ces points d'ancrage reprend le principe de la méthode de Mathieu Ramona de [14], mais pour laquelle l'énergie spectrale est remplacée tour à tour par chacun des flux spectraux testés.

Le nombre de flux spectraux et leur paramétrisation pour les généraliser peuvent paraître excessifs, mais la robustesse de cette étape est absoluement cruciale pour l'indexation audio réalisée lors du projet BeeMusic. Ce travail nous a permis de sélectionner 2 flux spectraux et les valeurs de paramètres conduisant à de bons résultats.

La batterie de tests présentée en section 5.2 nous a permis de sélectionner empiriquement mais de façon fiable, les 2 flux spectraux les mieux adaptés. Il s'agit de la *différence de norme spectrale* et la *KL-divergence* basée la divergence de Kulback-Leibler, cf. [8]. De même les bonnes valeurs des paramètres ont été choisie, cf. leurs valeurs en section 5.2.

Dans le cadre du projet BeeMusic, puisque nous nous intéressons au point d'ancrage associés aux onsets, autrement appelés transiets, nous choisissons la *différence de norme spectrale* avec les bonnes valeurs de paramètres.

Pour accélérer les calculs d'empreintes sonores, nous choisissons de calculer une seule fois le spectrogramme. Ainsi, le spectrogramme est à la fois utilisé pour la sélection des points d'ancrage et pour les empreintes sonores. Cela signifie que les valeurs des paramètres du spectrogramme doivent aussi être adaptées aux empreintes. Nous choisissons donc un pas d'avancement de 10ms, et une taille de fenêtre de 150ms, mais puisque nous préférons des fenêtres de Hann symétriques pour les empreintes, finalement, le flux spectral est calculé avec $a = 0$, et non $a = 0.5$ comme il a été trouvé en section 5.2. Nos tests montrent que cela produit une dégradation des performances de robustesse négligeable.

Remarquons qu'une autre contribution de ce travail est la définition de nouveaux flux spectraux basés sur des divergences telles que la divergence d'Iatkura-Saito ou de Kulback-Leibler. En effet, même si ces mesures de dissimilarité spectrale ne sont pas nouvelles, à notre connaissance, elles n'ont jamais été utilisées dans le cadre de la détection d'onsets. Notamment, on observe dans notre cas précis de très bons résultats avec la divergence de Kullback-Leibler. Nous pourrions alors imaginé de l'appliquer à la *vraie* détection d'onsets, et voir si cela apporte des performances intéressantes.

Références

[1] J.P. Bello, C. Duxbury, M. Davies, and M. Sandler. On the use of phase and energy for musical onset detection in the complex domain. *Signal Processing Letters, IEEE*, 11(6) :553–556, 2004.

[2] S. Dixon. Onset detection revisited. In *Proc. Int. Conf. on Digital Audio Effects (DAFx'06)*, pages 133–137, 2006.

[3] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler. A combined phase and amplitude based approach to onset detection for audio segmentation. In *Proc. 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03)*, pages 275–280. World Scientific, 2003.

[4] C. Duxbury, M. Sandler, and M. Davies. A hybrid approach to musical note onset detection. In *Proc. Int. Conf. on Digital Audio Effects (DAFx'02)*, pages 33–38, 2002.

[5] S. Essid. *Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique*. PhD thesis, Université Pierre et Marie Curie, 2005.

[6] F.J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1) :51–83, Jan 1978.

[7] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proc. 6th of the International Congress on Acoustics*, pages C17–C20, 1968.

[8] S. Kullback and R.A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.

[9] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791, 1999.

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

[11] P. Masri. *Computer modelling of sound for transformation and synthesis of musical signals*. PhD thesis, University of Bristol, 1996.

[12] R. Mignot. BeeAlter Toolbox : Boîte á outils de dégradations sonores, pour tests. Technical report, IRCAM – CNRS, 2015. BeeMusic Project, document interne.

[13] M. Ramona and G. Peeters. Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'11)*, pages 477–480, 2011.

[14] M. Ramona and G. Peeters. Audioprint : An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'13)*, pages 818–822, May 2013.

[15] A. Röbel. Onset detection in polyphonic signals by means of transient peak classification. *MIREX Online Proceedings (ISMIR 2005)*, 2005.

[16] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331–1334, Apr. 1997.

[17] Z. Yang, H. Zhang, Z. Yuan, and E. Oja. Kullback-leibler divergence for nonnegative matrix factorization. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 250–257. Springer, 2011.

[18] U. Zölzer. *DAFX : Digital Audio Effects*. Wiley, 2nd edition, 2011. 624 pages.