Chapitre 1

Méthodes d'analyse du signal musical

1.1. Introduction

L'un des objectifs de l'analyse du signal musical est d'en fournir une description aussi complète que possible. Les méthodes d'analyse sont très nombreuses et très diverses et permettent d'accéder, de façon plus ou moins robuste, aux divers paramètres de cette description.

Dans quel but effectuer une analyse du signal musical ? Nous en distinguerons trois. Tout d'abord, reproduire le signal musical par une machine après d'éventuelles transformations. Cela concerne l'analyse-synthèse ou le codage, l'objectif de l'analyse étant d'obtenir une représentation du signal lui-même qui autorise de telles transformations. Cela concerne aussi la synthèse et son contrôle, l'objectif de l'analyse étant d'estimer des paramètres utiles à la synthèse ou rendant son contrôle automatique.

Ensuite, permettre la réinterprétation du signal musical par un humain (ou par une machine MIDI). Il s'agit de transcrire le signal sous forme de partition (ou en signal MIDI), et donc d'obtenir une représentation de la notation musicale.

Finalement, accroître nos connaissances sur la production, la perception ou la modélisation des signaux musicaux. Il s'agit par exemple de concevoir et de valider des modèles de signaux ou encore d'établir des règles d'interprétation.

Chapitre rédigé par Boris DOVAL.

Examinons la notation musicale. C'est une forme de description de la musique, mais les musiciens savent qu'elle n'est que schématique et ne permet pas, seule, d'assurer une production de qualité. Le musicien est appelé à « interpréter » cette notation. Fournir une description du signal musical nécessite donc d'associer à la notation musicale des informations sur la façon dont le son est produit.

Explicitons donc en quoi la notation musicale décrit le signal. Elle est constituée principalement de notes définissant la fréquence et la durée relative du son à produire, d'indications d'articulation (liaisons, piqués, etc.) associées à la forme de l'enveloppe temporelle de chaque note, de nuances définissant, sur une échelle sommaire, la force du son, et du tempo permettant de relier les durées relatives et les durées absolues.

Parallèlement, la seule indication sur la production du son est le type d'instrument qui détermine un grand nombre de paramètres souvent rassemblés sous la notion de « timbre » : son percussif ou entretenu, enveloppe spectrale des partiels, évolution de cette enveloppe, présence d'inharmonicités, de bruits transitoires ou entretenus, de non-linéarités, etc.

Enfin, le jeu de l'instrumentiste n'est en général pas précisé non plus : par exemple le vibrato, les attaques, la tenue des sons (sons filés, soufflets), bref, tout ce qui fait la « vie » d'une note.

De cet examen rapide de la notation musicale se dégagent les analyses qui permettent de décrire le signal musical. Pour représenter le signal sonore, il faut savoir estimer l'enveloppe temporelle, les partiels, la fréquence fondamentale, l'enveloppe spectrale des partiels, la partie non périodique, et pouvoir suivre leurs évolutions dans le temps. Pour analyser les paramètres de jeu du musicien, il faut savoir en déduire les caractéristiques de l'enveloppe temporelle de chaque note, du vibrato (fréquence et amplitude) et d'autres variations du fondamental. Enfin, la notation musicale nécessite de savoir segmenter la phrase musicale en notes, estimer la durée, la fréquence et la force de chaque note, extraire le tempo (battue/rythme).

Ce chapitre ne prétend pas fournir des réponses à tous les problèmes évoqués ci-dessus, mais présente des méthodes d'analyse qui s'articulent autour de systèmes d'analyse-synthèse ou de « MIDIfication » du signal acoustique. Il est conçu pour que le lecteur puisse programmer les techniques élémentaires d'analyse permettant par exemple de « MIDIfier » un signal musical monophonique, tout en donnant des références de méthodes plus performantes. Les prérequis nécessaires sont les notions élémentaires de traitement du signal et d'algorithmique de niveau deuxième cycle universitaire. Les ouvrages suivants peuvent être consultés en complément : [BLA 98, PIC 98, RAB 75].

1.2. Outils d'analyse du signal sonore

1.2.1. *Le signal musical et sa numérisation*

Pour analyser le signal musical, il est nécessaire de connaître les ordres de grandeurs des mesures effectuées sur ce signal. En voici donc quelques grandeurs caractéristiques.

L'intensité I exprimée en dB SPL (*sound pressure level*) correspond à la mesure en décibels du carré de la variation de pression acoustique p_a prise en un point rapportée à la pression de référence $P_{\text{ref}} = 2 \times 10^{-5}$ Pa, c'est-à-dire $I = 10 \log_{10}((p_a/P_{\text{ref}})^2) = 20 \log_{10}(p_a/P_{\text{ref}})$. Les sons produits par les instruments ou par la voix ont une intensité acoustique pouvant varier sur une étendue d'environ 120 dB SPL, bien que cette étendue soit rarement utilisée ou même atteinte par un seul instrument.

La hauteur des notes est mesurée par leur fréquence exprimée en hertz. L'intervalle musical IM séparant deux notes de fréquences f_1 et f_2 est exprimé en demi-tons tempérés par $IM = 12 \log_2(f_1/f_2)$ où $\log_2(x) = \log(x)/\log(2)$. L'intervalle d'octave correspond à un rapport 2 entre les fréquences. Les sons produits par des instruments à hauteur de note définie couvrent une étendue de fréquence pouvant être très grande : de 27,5 Hz (période = 36 ms) à 4 186 Hz (0,239 ms) pour le piano, de 65,4 Hz (15 ms) à 349 Hz (2,86 ms) pour une tessiture de basse Do1-Fa3, de 261,6 Hz (3,82 ms) à 1 046 Hz (0,95 ms) pour une tessiture de soprano Do3-Do5.

Cependant, lors de la production d'une note à une fréquence F_0, sont aussi produites des fréquences plus élevées. Lorsque le son est périodique, ces fréquences sont multiples de F_0 et sont appelées harmoniques. Le nombre d'harmoniques d'amplitude significative varie beaucoup selon l'instrument et la hauteur de la note : une ou deux pour un sifflement, moins d'une dizaine pour la flûte, plusieurs dizaines pour le piano dans le grave.

Les analyses présentées ci-après font toutes l'hypothèse que le signal analysé est la variation de tension issue d'un capteur de type microphone mesurant les variations de pression acoustique en un point. De plus, le signal étant traité par algorithme, il est converti en un signal numérique noté $s(k)$, $k = 0, \ldots, K-1$ [1], à la fréquence d'échantillonnage $F_e = 1/T_e$. Ce signal est donc à bande limitée $[-F_e/2, F_e/2]$. Comme il est supposé réel, la bande de fréquence « utile » à analyser se restreint à $[0, F_e/2]$. De plus, chaque échantillon est codé sur un nombre limité de bits, ce qui limite la dynamique et introduit un bruit de quantification. La dynamique augmentant

1. Dans toute la suite, nous utiliserons la convention suivante (utilisée notamment en langage C) pour les indices : les indices d'un tableau de taille N vont de 0 à $N-1$.

de 6 dB par bit, les meilleurs systèmes d'enregistrement utilisent 20 bits pour couvrir la dynamique de 120 dB, mais les plus courants 16 pour 96 dB de dynamique (CD audio, DAT).

Les fréquences audibles couvrent la bande de 20 Hz à 20 000 Hz. La fréquence d'échantillonnage du CD audio, $F_e = 44\,100$ Hz, permet donc de représenter l'ensemble des fréquences audibles. Cependant, dans de nombreuses analyses, la bande des fréquences utiles est nettement inférieure et des fréquences d'échantillonnage de 16 000 Hz, 8 000 Hz ou même inférieures seront utilisées pour réduire le nombre d'échantillons à traiter.

1.2.2. *Analyse à court terme dans le domaine temporel*

La musique s'inscrit dans le temps. Elle est une succession d'événements généralement distincts à des échelles de temps très diverses, allant d'une milliseconde à une vingtaine de millisecondes pour les périodes fondamentales, d'un dixième de seconde à plusieurs secondes pour la durée des notes. La plupart des analyses étudient donc le signal localement avec une échelle temporelle adaptée de façon à caractériser ces événements séparément.

L'analyse à court terme permet de se focaliser sur une partie du signal concentrée dans le temps. Pratiquement, il suffit de multiplier le signal par une fonction nulle en dehors d'un support temporel borné. Une telle fonction, notée $w(k)$, $k = 0, \ldots, M-1$, est appelée « fenêtre de pondération ». Les paramètres importants sont la taille (M échantillons de signal correspondant à une durée de MT_e secondes) et le type de la fenêtre (rectangulaire, Hanning, Hamming, Blackmann, Blackmann-Harris, Kaiser, etc.). La fenêtre de Hanning est définie par $w(k) = 0{,}5 - 0{,}5\cos(2\pi k/M)$, celle de Hamming par $w(k) = 0{,}54 - 0{,}46\cos(2\pi k/M)$, celle de Blackmann par $w(k) = 0{,}42 - 0{,}50\cos(2\pi k/M) + 0{,}08\cos(4\pi k/M)$. Plus la taille de la fenêtre sera grande, plus l'analyse sera « délocalisée ». Le type de fenêtre a une influence sur l'étalement fréquenciel provoqué par le fenêtrage (voir paragraphe 1.2.3).

Si, en théorie, l'analyse peut être effectuée à chaque instant, en pratique elle est effectuée à des instants particuliers appelés « instants d'analyse » et notés t_l, $l = 0, \ldots, L-1$. La durée entre deux instants d'analyses successifs est donc un paramètre supplémentaire qui règle la précision temporelle de l'analyse. Dans le cas où les instants d'analyses sont régulièrement répartis, cette durée est constante et s'appelle période d'analyse T_a. Son inverse est la fréquence d'analyse F_a. Le nombre de points entre deux fenêtres successives est $D = T_a/T_e$. Dans le cas contraire, les instants d'analyse sont en général déterminés par les résultats d'une autre analyse. C'est le cas de l'analyse synchrone à la période fondamentale, utilisée notamment pour certaines transformations de durée et de hauteur de la voix, où les instants d'analyse correspondent à un point particulier d'une période fondamentale du signal.

Il est souvent nécessaire d'attribuer un instant précis à l'analyse effectuée sur une fenêtre. Dans la mesure où les fenêtres sont presque toujours symétriques, et pour respecter la régularité des instants d'analyse, l'instant précis d'analyse est très souvent choisi au centre de la fenêtre : $t_l = (lD + (M-1)/2)T_e$. Notons que dans certaines applications, le premier instant doit être à zéro, auquel cas le signal est complété par des zéros pour les temps négatifs.

Finalement, l'analyse à court terme consiste à construire la suite des signaux s_l à support temporel borné, appelés « trames » et définis par :

$$s_l(k) = w(k)s(lD + k) \qquad \text{pour } k = 0, \ldots, M-1$$

où $l = 0, \ldots, L-1$ désigne le numéro de la trame. Remarquons que la durée entre deux instants d'analyses successifs doit être inférieure à la taille de la trame pour éviter d'« oublier » certains échantillons de signal, ce qui est assuré dans le cas d'une analyse régulière par $D \leqslant M$.

1.2.3. *Analyse à court terme dans le domaine fréquenciel*

La structure de la musique se fonde aussi et surtout sur la notion essentielle de fréquence. Que ce soit pour déterminer la hauteur des notes, leur contenu harmonique ou la répartition de l'énergie par bande de fréquence, l'outil d'analyse fréquencielle est très utile.

Effectuer une analyse à court terme dans le domaine fréquenciel consiste simplement à appliquer une transformée de Fourier (TF) sur chaque trame issue de l'analyse précédente. En fait, pour qu'elle soit calculable, la transformée utilisée est la transformée de Fourier discrète (TFD), calculée par un algorithme de transformée de Fourier rapide (TFR, en anglais FFT). Le résultat est une représentation temps-fréquence qui est appelée transformée de Fourier à court terme (TFCT) [ALL 77, POR 80] et qui décrit le contenu fréquenciel du signal à chaque instant d'analyse :

$$\widetilde{s_l}(n) = \mathrm{TFD}(s_l)(n) = \sum_{k=0}^{N-1} e^{-j2\pi nk/N} s_l(k)$$

pour $n = 0, \ldots, N-1$.

Aux N points de signal, la TFD associe N points de spectre complexe que l'on représente habituellement sous forme polaire (module et phase). Ces N points, numérotés de 0 à $N-1$, correspondent à la bande de fréquence $[0, F_e[$. Le signal étant réel, le spectre est à symétrie hermitienne et seuls les points $[0, N/2]$ sont utiles. Le n-ième point de TFD correspond donc à la fréquence $f = nF_e/N$.

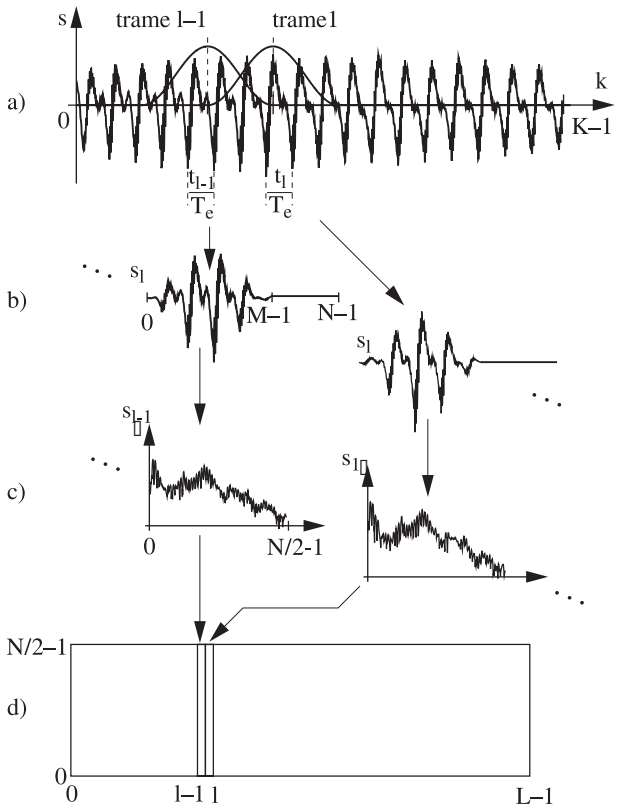

Figure 1.1. *Représentations du signal : a) représentation temporelle, b) représentation temporelle à court terme, c) représentation fréquencielle à court terme, d) spectrogramme*

Quel que soit N, les points de TFD représentent toujours la même bande de fréquences. Le rapport N/M agit donc comme un facteur de « suréchantillonnage » de la TF. Remarquons que dans l'équation précédente, la TF est prise sur N points de signal, bien que la fenêtre soit de taille M. En pratique, cela revient à compléter le signal fenêtré par des zéros (*zero padding* en anglais). Finalement, pour le calcul de la TFCT, il faut effectuer les choix de la taille et du type de fenêtre, de la fréquence d'analyse et de la précision spectrale d'affichage, ce qui revient à choisir M, D et N.

Le choix de M est un compromis : suffisamment grand pour englober le phénomène étudié et suffisamment petit pour ne pas intégrer ses variations dans une même trame. La quantité $1/MT_e$ est la résolution intrinsèque de l'analyse spectrale.

Le choix de N est moins critique : il doit être supérieur ou égal à M, suffisamment grand pour obtenir une bonne précision spectrale, mais limité par le temps de calcul (en $N \log(N)$) et par les algorithmes de TFR les plus courants qui imposent que N soit une puissance de 2. La quantité $1/NT_e$ est la précision d'affichage de l'analyse spectrale.

De son côté, le choix de $T_a = DT_e$ doit permettre d'« échantillonner » le phénomène étudié suffisamment finement pour suivre ses variations. Ainsi, d'après le théorème de Shannon, au-dessus d'une certaine valeur de T_a, l'analyse ne traduit pas les variations rapides du phénomène, par contre en dessous de cette valeur, diminuer T_a fournit plus de points d'analyse mais pas plus de précision dans les variations. Il y a donc une valeur optimale qui dépend de la taille et du type de fenêtre. Une heuristique [ALL 77] consiste à faire en sorte que les points soient pondérés également par la superposition, d'où le choix de $D = M/P$, où P est la largeur du lobe principal de la fenêtre de M points, exprimée en points de FFT (voir tableau 1.1).

Finalement, le choix de la fenêtre de pondération est dicté par des considérations spectrales. Les fenêtres sont utilisées pour réduire l'étalement spectral et les rebonds dus à la troncature temporelle. La TF de la fenêtre de pondération étant constituée d'un lobe principal et de lobes secondaires, il faudra faire en sorte que la largeur du lobe principal soit la plus faible possible, que l'amplitude des plus hauts lobes secondaires soit la plus faible possible, et enfin, selon les cas, que la pente des lobes secondaires soit la plus raide possible. La largeur du lobe principal exprimée en hertz est inversement proportionnelle à MT_e, le coefficient de proportionnalité (en points, voir tableau 1.1) dépendant du type de fenêtre. Par contre, elle est indépendante de N, ce qui montre bien qu'augmenter N améliore la précision de l'affichage (il y a plus de points) mais ne change en rien la résolution spectrale. L'amplitude et la pente des lobes secondaires, elles, ne dépendent que du type de fenêtre. Ceci implique notamment qu'agrandir la fenêtre ne réduit pas l'influence des lobes secondaires. Le choix de la fenêtre est alors un compromis entre ces critères.

Le tableau 1.1 résume les propriétés de quelques fenêtres parmi les plus utilisées. Une étude très détaillée peut être trouvée dans un article dû à F.J. Harris [HAR 78].

Il faut remarquer que la représentation graphique du module en dB de la TFCT, $20 \log_{10} |\widetilde{s_l}(n)|$ sous la forme d'une image de niveaux de gris, appelée spectrogramme, est très utilisée pour l'analyse visuelle des sons. Les temps discrets, $l = 0, \ldots, L-1$, sont en abscisse, les fréquences discrètes, $n = 0, \ldots, N-1$, en ordonnée et les amplitudes en dB sont représentées par les niveaux de gris de l'image. Une normalisation et un seuillage des amplitudes sont nécessaires pour une meilleure lisibilité.

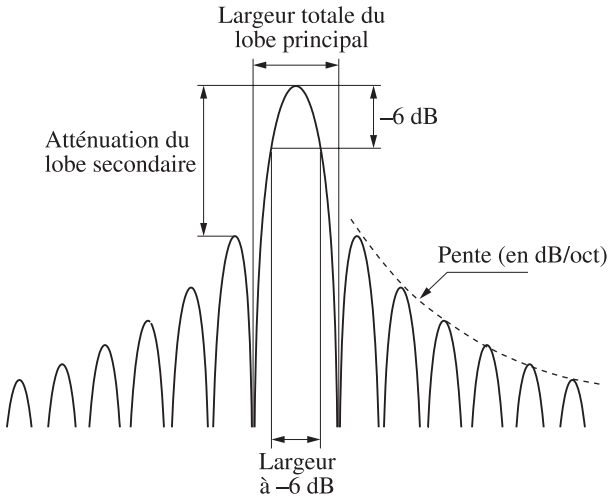

Figure 1.2. *Caractéristiques des fenêtres de pondération*

Fenêtre	**Largeur -6 dB lobe principal (pts)**	**Largeur totale lobe principal (pts)**	**Atténuation lobes second. (dB)**	**Pente (dB/Oct)**
Rectangulaire	1,21	2	-13	-6
Triangulaire	1,78	4	-26	-12
Hanning	2,00	4	-31	-18
Hamming	1,81	4	-42	-6
Blackmann	2,35	6	-58	-18
Kaiser $\alpha = 2{,}0$	1,99	4,5	-45	-6
Kaiser $\alpha = 3{,}0$	2,39	6,4	-69	-6

Tableau 1.1. *Caractéristiques de quelques fenêtres de pondération. La largeur à -6 dB (d'après [HAR 78]) et la largeur totale approximative du lobe principal (entre les deux minimums les plus proches) sont exprimées en nombre de points dans le cas où $N = M$; si $N \neq M$, les valeurs indiquées sont à multiplier par N/M.*

1.2.4. *Robustesse*

Le signal musical est extrêmement complexe et les analyses de ce signal sont souvent délicates. Il est donc important de s'assurer de la robustesse des analyses effectuées.

La notion de robustesse peut être vue comme le fait que les résultats de l'analyse seront peu sensibles à de petites variations du signal d'entrée. Pour illustrer cela, supposons que nous désirions connaître l'amplitude et la fréquence du partiel le plus fort dans le spectre du signal sonore. L'algorithme est évident : il consiste à déterminer le maximum du spectre et de prendre l'indice de ce maximum. Nous pourrons dire que l'estimation de cette amplitude est robuste, mais que celle de la fréquence correspondante ne l'est pas, car une modification infime de l'amplitude du partiel le plus fort peut modifier radicalement la fréquence du maximum, comme par exemple dans le cas de deux partiels ayant presque la même amplitude mais des fréquences très différentes.

Cependant, cette notion de robustesse est plutôt attachée à la grandeur mesurée qu'à l'analyse effectuée : ainsi, l'estimation de l'énergie est robuste, mais celle de la fréquence fondamentale ne l'est généralement pas.

1.3. Analyse de l'enveloppe temporelle

L'amplitude du signal acoustique peut être analysée à différentes échelles. La plus large qui nous concerne ici est l'échelle de la note et l'enveloppe temporelle d'amplitude traduit les variations d'intensité des différentes notes successives. Une échelle plus fine permet de mettre en évidence les variations d'intensité au cours de chaque note. Enfin, une échelle très fine révèle les variations de chaque période dans le cas d'un signal pseudo-périodique.

1.3.1. *Enveloppe temporelle*

La structure de l'enveloppe temporelle d'amplitude d'une note (on parle simplement d'enveloppe temporelle) est souvent décrite, en informatique musicale, par quatre phases successives : l'attaque, la décroissance, la tenue et le relâchement, plus connues en anglais sous les termes *attack*, *decay*, *sustain* et *release* (ADSR). Cette décomposition est notamment utilisée pour la synthèse par échantillonnage où les phases A, D et R sont reproduites telles quelles et où la phase S, la plus variable, est bouclée pour satisfaire à la durée de la note à synthétiser et comporte éventuellement des variations d'intensité. Le modèle sous-jacent est celui d'un signal périodique [2] $x_{T_0}(t)$ modulé en amplitude par cette enveloppe temporelle $A(t)$:

$$s(t) = A(t)x_{T_0}(t)$$

2. En toute généralité, il s'agit d'un signal composé d'une somme finie de sinusoïdes non nécessairement harmoniques, ce qui permet de traiter aussi les sons percussifs et/ou inharmoniques de type cloche. Il faut alors remplacer F_0 par la fréquence de la composante la plus grave (c'est-à-dire de plus petite fréquence).

Bien entendu, l'enveloppe temporelle est supposée varier lentement par rapport aux variations de x_{T_0}. L'enveloppe temporelle $A(t)$ est donc un signal à bande limitée $[-B, +B]$ telle que la largeur de bande soit inférieure à la plus petite fréquence de x_{T_0} : $2B < F_0$.

1.3.2. *Estimation de l'enveloppe temporelle*

Un des systèmes les plus simples pour estimer l'enveloppe temporelle est analogique : il s'agit du redressement simple ou double alternance suivi d'un filtrage passe-bas, réalisable par un simple pont de diode et un circuit RC. Ce principe est utilisé en radio pour effectuer une démodulation AM. Mais il est aussi utilisé en musique dans les systèmes analogiques d'analyse du signal musical.

Le seul paramètre à régler ici est la fréquence de coupure f_c du filtre passe-bas. Elle doit être choisie pour séparer au mieux le contenu fréquenciel de l'enveloppe $(A(t))$ de celui du signal périodique sous-jacent $(x_{T_0}(t))$ et devra donc appartenir à l'intervalle $[B, F_0/2]$. De plus, pour éviter des déphasages non linéaires importants, le filtre passe-bas n'aura pas une bande de transition très étroite, ce qui conduit à choisir f_c plutôt proche de B. L'ordre de grandeur de B est de 10 à 30 Hz.

Le principal défaut de cette technique est la difficulté d'estimer correctement le gain à appliquer à la sortie du filtre (supposé de gain 1 en dessous de f_c) pour que l'estimation « enveloppe » vraiment le signal. Dans le cas simple où $x_{T_0}(t)$ peut être considéré comme constitué d'une seule harmonique, il suffit de diviser le résultat par la valeur moyenne du cosinus redressé, à savoir par $1/\pi$ ou $2/\pi$ selon qu'il s'agit de simple ou de double alternance. Mais si $x_{T_0}(t)$ est riche en harmoniques, ce calcul ne convient plus et l'estimation de l'enveloppe ne sera correcte qu'à un coefficient multiplicatif près.

Une variante qui ne présente pas cet inconvénient consiste à suivre le signal dans sa phase montante et à redescendre en exponentielle décroissante de constante de temps τ [KUH 90]. La version numérique de cette technique s'écrirait alors : $\hat{A}(k) = \max(\hat{A}(k-1)e^{-T_e/\tau}, |s(k)|)$, où $s(k)$ représente l'échantillonnage de $s(t)$ à la période T_e, et $\hat{A}(k)$ l'estimation de l'enveloppe temporelle. La constante de temps τ doit être suffisamment grande pour éliminer les harmoniques de $x_{T_0}(t)$, mais suffisamment petite pour suivre la décroissance de l'enveloppe. Elle correspond, à un coefficient près, à l'inverse de la fréquence de coupure précédemment définie : $\tau = 1/(2\pi f_c)$.

L'avantage de cette technique est la fidélité de l'estimation dans les attaques de notes. Or, l'attaque de la note, qui est très importante dans la perception de l'identité des différents instruments de musique, se trouve être la partie de l'enveloppe qui évolue le plus rapidement. Cependant, le résultat présente des oscillations assez fortes qui nécessitent un filtrage passe-bas à f_c.

Bien qu'il soit tout à fait possible de réaliser ces estimations par algorithme (valeur absolue, puis filtrage passe-bas numérique), la technique numérique la plus employée consiste à estimer l'enveloppe par l'énergie à court terme du signal, qui n'est autre qu'une analyse à court terme suivie d'un calcul d'énergie sur chaque trame s_l de signal :

$$e(l) = \sum_{k=0}^{M-1} s_l^2(k) \qquad \text{pour } l = 0, \ldots, L-1$$

d'où est déduite l'estimation de l'amplitude :

$$\hat{A}_l = \sqrt{\frac{e(l)}{M}}$$

Bien entendu, la fréquence d'échantillonnage de $\hat{A}_l$ étant égale à la fréquence d'analyse F_a, il faut suréchantillonner $\hat{A}_l$ d'un facteur D pour l'aligner sur le signal.

Ici, c'est le paramètre de taille de la fenêtre qui détermine l'effet de filtrage. Plus la taille est grande, plus le signal est « moyenné » et donc filtré. En fait, ce calcul peut être interprété comme le filtrage RIF par une fenêtre rectangulaire du signal s_l^2. La largeur totale du lobe principal de la fenêtre rectangulaire étant égale à $2/(MT_e)$, la fréquence de coupure f_c correspond approximativement à $1/(MT_e)$. Il faut remarquer que ce filtrage est très médiocre en termes de performance de filtrage. Ce faisant, l'utilisation d'une fenêtre autre que la fenêtre rectangulaire doit se concevoir comme RI d'un filtre passe-bas, ce qui permet alors de réduire les oscillations résiduelles. Comme dans le cas de la première technique présentée, cette technique ne permet pas d'estimer facilement le gain à appliquer pour que l'estimation « enveloppe » vraiment le signal.

Une dernière technique consiste à estimer l'enveloppe comme le module du signal analytique correspondant à $s(t)$. En effet, le signal analytique est un signal complexe dont la partie réelle est le signal d'entrée et tel que ses parties réelle et imaginaire sont en quadrature. Ainsi, le signal analytique d'un cosinus est l'exponentielle complexe correspondante. De plus, sous l'hypothèse que $A(t) \geqslant 0$ est à bande limitée $[-B, +B]$, le signal analytique de $A(t)\cos(2\pi ft)$, où $f > B$, est le signal $A(t)e^{j2\pi ft}$ dont le module est $A(t)$. Bien entendu, dans notre hypothèse, le cosinus est remplacé par le signal périodique $x_{T_0}(t)$ qui peut s'écrire comme une somme de cosinus et le module du signal analytique est donc $A(t)$ multiplié par le module d'une somme d'exponentielle complexe, qui, cette fois-ci, n'est pas constante. Le résultat présente donc des oscillations aux fréquences du signal $x_{T_0}(t)$ qui peuvent être particulièrement importantes. Un filtrage passe-bas à une fréquence de coupure f_c (dont la valeur a été précédemment discutée) s'avère donc nécessaire. Enfin, le signal analytique se calcule soit en appliquant un filtre RIF passe-tout déphaseur pur dont on peut trouver les caractéristiques dans [RAB 74], soit en annulant les fréquences négatives de la transformée de Fourier.

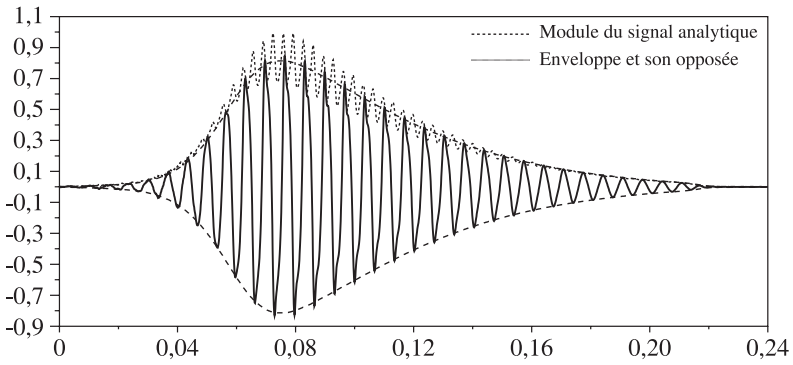

Figure 1.3. *Estimation d'enveloppe temporelle par la méthode du signal analytique. Le signal est un son court de clarinette. Le signal analytique est obtenu par la formule* $sa = \mathrm{TF}^{-1}(\widetilde{sa})$*, où* $\widetilde{sa}(n) = 2\tilde{s}(n)$ *pour* $n < M/2$*,* $\widetilde{sa}(n) = 0$ *sinon. La TF est prise sur* M *points (ici,* $M = 1\,920$*, soit 240 ms). Observez les oscillations importantes du module de* sa *(en pointillés). L'enveloppe (en tirets) a été obtenue par filtrage passe-bas (sans retard de groupe) de* sa *à la fréquence de coupure* $f_c = 15$ *Hz.*

1.3.3. *Segmentation*

Dans l'optique d'une « MIDIfication » du signal acoustique, il faut d'abord segmenter le signal, ce qui revient à détecter le début et la fin de chaque note. Cette détection peut être partiellement résolue en exploitant l'estimation de l'enveloppe temporelle.

La technique la plus directe est de seuiller l'enveloppe : lorsque l'enveloppe dépasse le seuil, cela indique un début de note ; quand elle redescend sous le seuil, cela signale une fin de note.

Malheureusement, cette technique est très sensible à la présence de bruit et est incapable de détecter le début de notes enchaînées dans les cas très courants où l'amplitude ne redescend pas sous le seuil entre deux notes successives.

En fait, l'attaque des notes est plus facilement repérable, car elle est caractérisée par une variation rapide d'amplitude. La technique la plus utilisée est donc de détecter les maximums locaux de la dérivée de l'enveloppe temporelle.

Les difficultés proviennent principalement de la présence de bruit de fond, des modulations du signal et des oscillations résiduelles de la plupart des techniques d'estimation de l'enveloppe temporelle. Ceci peut produire un grand nombre de fausses détections.

Les solutions envisagées sont d'utiliser un seuil global (en général très bas) pour éliminer le bruit et de fixer une durée minimale de la note pour éliminer les détections trop rapprochées.

1.4. Analyse des partiels et du bruit

L'oreille humaine est très sensible au contenu fréquenciel des sons. Le « timbre » des sons est souvent étudié au travers de la répartition spectrale de l'énergie et de son évolution dans le temps. Selon les catégories d'instruments de musique (à hauteur bien définie ou non, à son entretenu ou à son percussif), ce contenu fréquenciel peut être de différentes natures (partiels sinusoïdaux ou sinusoïdaux amortis, bruits de tous types).

L'analyse des partiels et du bruit repose donc sur une représentation du signal musical $s(t)$ en somme de partiels $s_i(t)$ et de bruit $b(t)$:

$$s(t) = \sum_i s_i(t) + b(t)$$

En pratique, la notion de partiel étant suffisamment large, il est tout à fait possible de représenter une partie du bruit par des partiels. Ceci laisse une certaine liberté quant à l'orientation de l'analyse : McAulay et Quatieri [MCA 86] et Fitz et Haken [FIT 96] représentent tout ce qui peut l'être dans le signal par des partiels, Serra et Smith [SER 90] sélectionnent les partiels représentatifs de la partie déterministe et modélisent le bruit séparément, Yegnanarayana *et al*. [YEG 98] ne retiennent que les partiels harmoniques et considèrent le reste comme le bruit qui contient alors toutes les apériodicités.

C'est l'application qui va orienter l'analyse : pour le codage, la technique de McAulay et Quatieri [MCA 86] convient parfaitement, mais pour l'analyse-synthèse avec modification, il est nécessaire de distinguer la partie pseudo-périodique de la partie bruit.

1.4.1. *Partiels d'un signal*

Par définition, le partiel d'un signal est une composante sinusoïdale de ce signal. Un partiel est donc caractérisé par une fréquence centrale f, une amplitude A et une phase initiale ϕ, et son expression numérique est : $s_i(k) = A\cos(2\pi fkT_e + \phi)$, $k = 0, \ldots, K-1$. Le signal complexe correspondant est $A\,e^{j(2\pi fkT_e+\phi)}$. Une définition plus générale est le partiel sinusoïdal amorti $A\,e^{-\alpha k+j(2\pi fkT_e+\phi)}$, où $\alpha > 0$ est un coefficient d'amortissement constant.

Un signal composé d'une somme de partiels a un spectre de raies. Cependant, l'observation pratique de ce signal nécessite de le tronquer, ce qui produit dans le

spectre un élargissement des raies et des oscillations. Cet effet spectral de la troncature (appelé phénomène de Gibbs) s'explique par le fait que multiplier le signal par une fenêtre de pondération (troncature temporelle) est équivalent à convoluer le spectre du signal par le spectre de la fenêtre. Ainsi, chaque raie spectrale sera remplacée, dans le spectre du signal observé, par l'image du spectre de la fenêtre à la position de la raie.

La technique d'analyse spectrale utilisée dans la suite de ce paragraphe est la TFCT. Des techniques plus précises existent mais imposent des contraintes de stationnarité qui ne sont pas respectées par le signal musical.

Les problèmes à résoudre sont alors :

1) la détection d'un partiel dans du bruit,

2) la séparation de plusieurs partiels.

Le choix des paramètres d'analyse spectrale (voir paragraphe 1.2.3 pour une discussion) doit assurer que chaque partiel présent dans le signal provoque l'apparition d'un pic sur le spectre (sinon, il sera impossible de le détecter) et que la résolution spectrale sera suffisante pour séparer deux partiels quelconques. Il faut donc que l'atténuation des lobes secondaires de la fenêtre de pondération soit supérieure au plus grand écart d'amplitude entre deux partiels et que la largeur à -6 dB du lobe principal soit inférieure au plus petit écart de fréquence entre deux partiels (voir tableau 1.1).

1.4.2. *Détection et sélection de pics*

Le principe est simple : il s'agit de repérer tous les pics du spectre d'amplitude (c'est-à-dire les maximums locaux) et de considérer qu'un pic correspond à un partiel. Bien sûr, le signal évoluant dans le temps, cette opération s'effectue à court terme, c'est-à-dire sur chaque trame de la TFCT.

L'algorithme le plus simple est donc :

Pour chaque trame $l = 0, \ldots, L-1$
  $i \leftarrow 0$
  Pour $n \leftarrow 1$ à $N-2$
    Si $|\widetilde{s_l}(n)| > |\widetilde{s_l}(n-1)|$ et $|\widetilde{s_l}(n)| > |\widetilde{s_l}(n+1)|$
      $i \leftarrow i+1$
      $h_l(i) \leftarrow n$
  $I_l \leftarrow i$

Le nombre de pics de la trame l est I_l et ces pics correspondent aux points du spectre d'indices $h_l(i)$, $i = 1, \ldots, I_l$. La fréquence en hertz du pic numéro i est $f_{l,i} = h_l(i)F_e/N$ et son amplitude complexe est $A_{l,i} = \widetilde{s_l}(h_l(i))$.

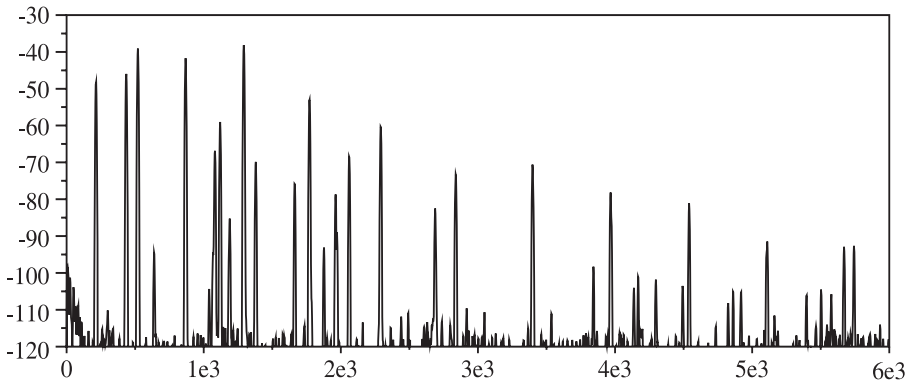

Figure 1.4. *Exemple de spectre de son de cloche. La plupart des pics représentent un partiel dont la fréquence est en abscisse et l'amplitude en ordonnée. Les partiels les plus importants ont comme fréquence 520 Hz, quelques multiples de 440 Hz et, dans les hautes fréquences, quelques multiples de 570 Hz.*

Cet algorithme extrait tous les pics du spectre. Si l'analyse prévoit une sélection, la question qui se pose est de déterminer, parmi les pics trouvés, ceux qui correspondent effectivement à un partiel et ceux qui sont des pics parasites.

Pour éliminer les pics parasites correspondant au bruit de fond, un seuil absolu (très bas ou estimé sur du « silence ») peut être utilisé.

L'atténuation des lobes secondaires de la fenêtre de pondération étant connue, tout pic dont l'amplitude relative à celle du pic le plus fort est inférieure à cette atténuation doit être rejeté. Par exemple, en utilisant la fenêtre de Hamming, tout pic dont l'amplitude est inférieure à celle du pic le plus fort moins 42 dB est à rejeter. En toute rigueur, il faudrait tenir compte de la pente de décroissance des lobes secondaires, mais la pratique est presque toujours de seuiller l'amplitude des pics relativement au pic le plus fort en choisissant ce seuil selon le type de fenêtre utilisé (voir tableau 1.1).

La sélection peut porter sur la « qualité » du pic : dans [SER 90], les pics sont sélectionnés par leur hauteur, c'est-à-dire par l'écart entre l'amplitude du pic et l'amplitude des vallées de chaque côté du pic.

Fitz et Haken interprètent leur sélection en termes de masquage psychoacoustique [FIT 96]. Pratiquement, ils partagent le spectre en bandes de fréquences logarithmiques et appliquent un seuil relatif sur chaque bande. Ceci est spécialement intéressant pour les analyses à grande F_e. Enfin, l'utilisation d'une bande de fréquences utiles est aussi une sélection.

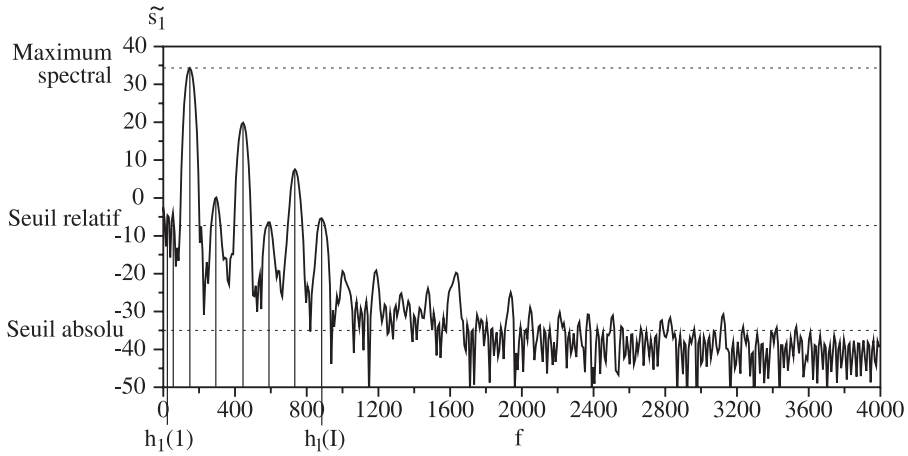

Figure 1.5. *Détection et sélection de pics spectraux. Le seuil absolu (ici -35 dB) élimine les pics correspondant au bruit de fond. Le seuil relatif au maximum spectral (ici $\max - 42$ dB) élimine les lobes secondaires (dont un exemple est visible à environ 200 Hz) et les pics trop faibles. Il reste ici $I = 8$ pics dont six sont des partiels et deux (les deux premiers) pourraient être éliminés pour leur mauvaise « qualité » (voir texte). Le spectre est calculé sur le signal de la figure 1.3.*

1.4.3. *Interpolation spectrale des pics*

L'algorithme précédent fournit des indices de la TFD comme position en fréquence des pics. Dans la plupart des cas, la précision sur les valeurs de fréquences F_e/N est très insuffisante (par exemple, $F_e = 44\,100$ Hz, $N = 1\,024$, précision $= 43$ Hz entre deux indices). La solution consistant à suréchantillonner le spectre en complétant la fenêtre de signal par des zéros lors du calcul de la TFCT (voir paragraphe 1.2.3) est inapplicable car elle augmente démesurément la complexité pour le calcul de la TFCT, mais aussi pour l'algorithme de détection des pics.

La solution qui lui est préférée consiste à interpoler le spectre uniquement au voisinage des pics. Le point de vue le plus direct effectue une interpolation polynômiale de degré 2 ou 3, utilisant donc trois ou quatre points répartis autour du pic à estimer, et renvoie la valeur de fréquence et d'amplitude du maximum du polynôme. Les formules pour l'interpolation parabolique sont les suivantes : soit le i-ième pic correspondant au n-ième point du spectre ($n = h_l(i)$), alors une valeur plus précise de la fréquence de ce pic $\hat{f}_{l,i} = \hat{n}_{l,i} F_e/N$ ($\hat{n}_{l,i}$ n'étant plus entier) et de son amplitude $\hat{A}_{l,i}$ peut être obtenue par une interpolation parabolique :

$$\hat{n}_{l,i} = n + \frac{1}{2}\frac{A_1 - A_{-1}}{2A_0 - A_{-1} - A_1}$$

et :

$$\hat{A}_{l,i} = A_0 + \frac{1}{4}(A_1 - A_{-1})(\hat{n}_{l,i} - n)$$

où, dans cette formule, $A_0 = |\widetilde{s_l}(n)|$, $A_{\pm 1} = |\widetilde{s_l}(n \pm 1)|$. Il faut noter que la précision est meilleure si l'interpolation de la fréquence et de l'amplitude est effectuée sur le spectre en dB [SER 89] (il suffit de remplacer A_0 et $A_{\pm 1}$ par $20 \log_{10} |\widetilde{s_l}(n)|$ et $20 \log_{10} |\widetilde{s_l}(n \pm 1)|$, puis de repasser le résultat en linéaire). Pour obtenir la phase interpolée, il faut effectuer l'interpolation de $\hat{n}_{\,l,i}$ comme précédemment, remplacer uniquement dans la formule donnant $\hat{A}_{l,i}$ les amplitudes A_0 et $A_{\pm 1}$ par les amplitudes complexes $\widetilde{s_l}(n)$ et $\widetilde{s_l}(n \pm 1)$ et prendre l'argument du résultat. Notons que la phase ne peut être estimée qu'à un multiple de 2π près.

Enfin, pour que l'amplitude estimée d'un pic corresponde à son amplitude effective dans le signal, il faut normaliser (diviser) l'amplitude précédente par le poids de la fenêtre sur chaque pic, $\sum_{k=0}^{M-1} w(k)/2$.

1.4.4. *Suivi de partiels*

Déterminer les pics du signal à chaque trame en se fondant uniquement sur les données de cette trame conduit à des listes indépendantes de pics non connectés. Or, un partiel naît avec la note, vit en fonction des variations de la note et meurt avant ou à la fin de la note. Sa durée de vie s'étale donc sur plusieurs trames. Il importe de relier les détections de pics effectuées sur chaque trame pour créer des « lignes » (*tracks* en anglais) de fréquence/amplitude traduisant l'évolution du partiel dans le temps. Dans la suite de ce paragraphe, la notion de partiel correspond donc à une suite de pics spectraux connectés. Le suivi est notamment indispensable pour effectuer une modification cohérente ou simplement une resynthèse.

Bien entendu, la technique qui consiste à associer les pics de deux trames successives ayant le même numéro ne peut pas fonctionner à cause de la présence de pics parasites et de la naissance ou de la mort de partiels de fréquence intermédiaire.

La technique décrite dans [MCA 86] consiste à déterminer de proche en proche les partiels par « continuité » : un partiel déterminé à la trame l est prolongé à la trame $l + 1$ par le pic dont la fréquence est la plus proche (à condition qu'elle appartienne à un intervalle centré sur la fréquence du partiel). Si un tel pic n'existe pas, le partiel « meurt ». De plus, un même pic ne peut appartenir qu'à un seul partiel, ce qui nécessite une technique de résolution de conflit lorsqu'un pic est le plus proche de deux partiels différents. Une fois tous les partiels traités, les pics restants sont considérés comme des partiels naissants. Lors de la naissance (respectivement de la mort) d'un partiel, il faut interpoler l'amplitude à zéro sur la trame précédente (respectivement suivante).

L'algorithme de McAulay et Quatieri est le suivant :

Pour $l \leftarrow 0$ à $L-2$ (pour chaque trame)
- Les fréquences $(f_{l,i})_{i=1,I_l}$ sont supposées triées par ordre croissant.
- Déclarer tous les partiels des trames l et $l+1$ comme non appariés
- Pour $i \leftarrow 1$ à I_l (pour chaque partiel de la trame l)
  - Trouver les pics non appariés de la trame $l+1$
  - dont la fréquence est dans l'intervalle $[f_{l,i}-\Delta, f_{l,i}+\Delta]$
  - Si aucun pic ne correspond
    - $f_{l,i}$ est déclaré mort et marqué comme apparié
  - Sinon
    - Soit $f_{l+1,j}$ la fréquence du pic le plus proche
    - Si le partiel non apparié de la trame l le plus proche de $f_{l+1,j}$ est $f_{l,i}$
      - Apparier $f_{l,i}$ et $f_{l+1,j}$: $suiv_l(i) = j$
      - Marquer $f_{l,i}$ et $f_{l+1,j}$ comme appariés
    - Sinon
      - Si le pic non apparié de la trame $l+1$ de fréquence $f_{l+1,j'}$
      - juste inférieure à $f_{l+1,j}$ est dans l'intervalle $[f_{l,i}-\Delta, f_{l,i}+\Delta]$
        - Apparier $f_{l,i}$ et $f_{l+1,j'}$: $suiv_l(i) = j'$
        - Marquer $f_{l,i}$ et $f_{l+1,j'}$ comme appariés
      - Sinon
        - $f_{l,i}$ est déclaré mort et marqué comme apparié
- Pour tous les pics j non appariés de la trame $l+1$
  - $f_{l+1,j}$ est déclaré naissant

Des améliorations ont été proposées à cet algorithme pour tenir compte d'effets à plus long terme que deux trames successives. En effet, il apparaît souvent qu'un partiel meurt et qu'un autre renaît quelques trames plus loin à la même fréquence, ce qui provient du fait que l'amplitude du partiel passe temporairement sous le seuil de rejet. Pour corriger cet inconvénient, Serra et Smith [SER 90] proposent de définir un état de « veille », retardant la mort des partiels inactifs et permettant leur réactivation s'ils peuvent s'apparier avec un pic avant un nombre déterminé de trames. Cependant, l'amplitude des partiels en état de veille étant nulle, ceci peut s'entendre comme un artefact lors de la resynthèse. Fitz et Haken [FIT 96] proposent de corriger ceci en intégrant les procédures de sélection dans celles de suivi de partiels pour créer un hystérésis dans le seuil de sélection : un partiel peut naître s'il est au-dessus du seuil de montée, mais ne peut mourir que s'il passe sous le seuil de descente qui est strictement inférieur au seuil de montée.

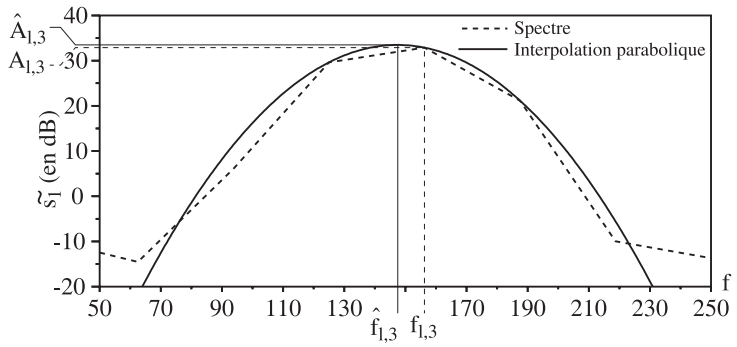

Figure 1.6. *Interpolation parabolique. En agrandissant le pic numéro 3 du spectre de la figure 1.5, on voit apparaître la discrétisation du spectre (en pointillé). La précision sur la fréquence du maximum local, $f_{l,3}$, peut être améliorée par une interpolation parabolique (en trait plein) fournissant la fréquence $\hat{f}_{l,3}$. L'interpolation a été effectuée sur les amplitudes en décibels.*

Enfin, d'autres critères sont utilisés tels qu'un nombre minimal de trames pour qu'un partiel soit valide ou qu'un nombre maximal autorisé de partiels simultanés. De même, lorsque le signal est supposé pseudo-périodique, l'algorithme se simplifie car il suffit alors de déterminer un partiel par pseudo-harmonique.

Le problème de suivi peut être aussi résolu par alignement temporel dynamique : apparier deux vecteurs de fréquences ordonnées revient à minimiser un critère de coût global de l'appariement, ce coût étant égal à la somme des écarts en fréquence de tous les partiels appariés. Cependant, l'effort pour formaliser ce problème peut être poussé un peu plus loin en tenant compte aussi de l'hypothèse de régularité de l'évolution des partiels. Ainsi, une technique de suivi par modèles de Markov cachés est décrite dans [DEP 93].

1.4.5. *Estimation de la partie bruit*

Nous avons vu, lors de la sélection des pics spectraux, des procédures destinées à éliminer les pics « parasites ». Il faut cependant se garder de penser que ce « bruit » n'est qu'une information parasite. En effet, les cas sont nombreux où ce « bruit » est une information primordiale du signal musical : le souffle d'un son de flûte, les transitoires d'attaque des tuyaux d'orgue, l'attaque des sons percussifs, le bruit des marteaux du piano, le frottement de l'archet du violon, les bruits d'aspiration, d'explosion et de frication dans la voix, etc.

La plupart des méthodes utilisées pour estimer la partie « bruit » du signal musical reposent sur un modèle de bruit additif : $s = x + b$, où x est un signal déterministe

(par exemple, pseudo-périodique) et b un signal aléatoire. Ce dernier signal est décrit par sa densité spectrale de puissance et il est souvent modélisé par un bruit blanc filtré. L'estimation de b consiste donc à estimer sa densité spectrale de puissance, soit en la décrivant explicitement en fonction de la fréquence, soit en la résumant dans le jeu de coefficients du filtre. Il faut noter que la phase ne joue aucun rôle et que l'estimation peut donc être réalisée sur le module du spectre.

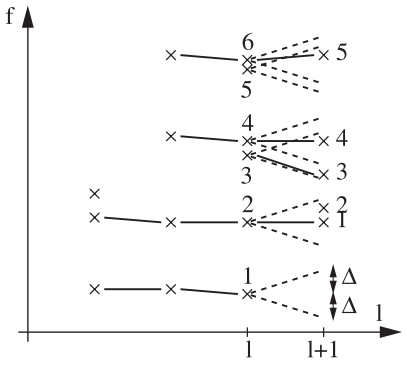

Figure 1.7. *Suivi de partiels. En déroulant l'algorithme sur les trames l et $l+1$, on obtient successivement les appariements suivants : $(l,1)$ meurt ; puis $(l,2)$ est apparié à $(l+1,1)$; ensuite $(l,3)$ est associé à $(l+1,4)$, mais comme ce dernier est plus proche de $(l,4)$, $(l,3)$ est apparié à $(l+1,3)$; puis $(l,4)$ est apparié à $(l+1,4)$; puis $(l,5)$ est associé à $(l+1,5)$, mais comme ce dernier est plus proche de $(l,6)$, $(l,5)$ meurt et c'est $(l,6)$ qui est apparié à $(l+1,5)$; finalement, $(l+1,2)$ est déclaré naissant.*

Néanmoins, en pratique, l'estimation du « bruit » n'est qu'un sous-produit de l'estimation de la partie « déterministe », obtenu par soustraction. La première étape est donc de calculer le signal de synthèse à partir des partiels pour pouvoir le retrancher au signal d'origine. Voici la formule de resynthèse décrite dans [MCA 86] : ayant estimé les amplitudes $\hat{A}_{l,i}$, fréquences $\hat{f}_{l,i}$ (les fréquences relatives sont donc $\hat{\nu}_{l,i} = \hat{f}_{l,i}/F_e$) et phases $\hat{\theta}_{l,i}$ de chaque partiel i pour chaque trame l, le signal de synthèse s'écrit comme une somme de sinusoïdes : $y_l(k) = \sum_{i=1}^{I_l} \hat{A}_{l,i}(k)\cos(\hat{\theta}_{l,i}(k))$ pour $k = 0, \ldots, D-1$ (les trames de synthèse sont donc de taille D et sont juxtaposées), où les amplitudes sont interpolées linéairement et les phases par un polynôme de degré 3 (assurant la continuité de la phase et de sa dérivée) :

$$\hat{A}_{l,i}(k) = \hat{A}_{l,i} + \frac{\hat{A}_{l+1,i} - \hat{A}_{l,i}}{D}k \qquad \text{pour } k = 0, \ldots, D-1$$

$$\hat{\theta}_{l,i}(k) = \hat{\theta}_{l,i} + 2\pi\hat{\nu}_{l,i}k + \left(\frac{\pi}{D}(\hat{\nu}_{l+1,i} - \hat{\nu}_{l,i}) + \frac{6\pi}{D^2}(\hat{M} - M_{opt})\right)k^2 - \frac{4\pi}{D^3}(\hat{M} - M_{opt})k^3$$

où la phase a été déroulée du facteur entier $\hat{M}$ (c'est-à-dire $\hat{\theta}_{l,i}(D) = \hat{\theta}_{l+1,i} + 2\pi\hat{M}$) le plus proche du réel $M_{\ opt}$ qui rend la courbe de phase la plus régulière [3] :

$$M_{opt} = \frac{1}{2\pi}(\hat{\theta}_{l,i} - \hat{\theta}_{l+1,i}) + \frac{D}{2}(\hat{\nu}_{l,i} + \hat{\nu}_{l+1,i})$$

La partie bruit est la différence entre ce signal synthétique reconstruit et le signal d'origine. Les spectres de bruit obtenus présentent alors des « trous » à la position des partiels. Or, ces « trous » étant répartis régulièrement, l'écoute d'un signal reconstruit à partir d'un tel spectre fait apparaître très clairement une hauteur (qui est celle du signal de départ). Cet artefact est supprimé en remplissant les « trous » laissés par les partiels. Plusieurs techniques sont utilisées selon les cas et la précision souhaitée de l'estimation. Serra et Smith [SER 90] utilisent l'interpolation linéaire pour calculer une enveloppe spectrale simplifiée du bruit : ils considèrent Q points régulièrement espacés en fréquence et affectent à chaque point l'amplitude maximale dans une bande centrée en ce point. Ensuite, la reconstruction du spectre est effectuée par interpolation linéaire entre les amplitudes de ces points.

Une autre méthode, fondée sur le modèle de production source/filtre, aboutit à une résolution par LPC (voir section 1.6). L'enveloppe est alors la réponse en fréquence d'un filtre tout-pôle ajusté par minimisation d'une erreur quadratique.

Remarquons que les techniques précédentes ne fournissent pas une décomposition exacte du signal en somme d'une partie « déterministe » et d'une partie « bruit », à cause de l'approximation effectuée sur la partie bruit. Dans certaines applications où la décomposition doit être exacte, il faut que la partie du bruit reconstruite à la position des partiels soit en quelque sorte ponctionnée sur ces partiels. C'est ce que proposent d'Alessandro *et al*. et Yegnanarayana *et al*. [DAL 98, YEG 98] dans un algorithme itératif de décomposition en parties périodique et apériodique.

1.5. Analyse du fondamental

En musique, la notion de fréquence fondamentale est liée à celle de hauteur mélodique. Sa connaissance est donc utile dans de nombreux cas, pour l'étude des échelles musicales en musicologie, du vibrato instrumental ou vocal, le suivi d'instruments, la transcription automatique de partitions, mais aussi comme information nécessaire à d'autres analyses comme l'analyse/synthèse synchrone au fondamental, certaines analyses de la qualité vocale, la décomposition périodique/apériodique, etc.

Les besoins des utilisateurs d'un système d'analyse du fondamental sont donc variés, allant de l'étiquetage du signal période par période à la détermination de la

3. Au sens du minimum du critère $\int_0^{DT_e} (\frac{d^2}{dt^2}\theta(t))^2\, dt$.

mélodie en termes de notes, en passant par l'estimation de la fréquence fondamentale à des instants régulièrement répartis.

Un grand nombre de techniques ont été développées à l'origine pour la parole [HES 83], donc pour un signal monophonique, variant sans cesse et à l'étendue relativement restreinte (de l'ordre de deux octaves). Les adapter au signal musical est possible à condition qu'elles puissent gérer de grandes étendues de fréquences fondamentales (jusqu'à six ou sept octaves). Cependant, la plus grosse difficulté est la polyphonie : lorsque plusieurs musiciens jouent ensemble ou lorsqu'il s'agit d'un instrument polyphonique, le son produit est beaucoup trop complexe pour qu'un système puisse, en toute généralité, en extraire les différentes notes. Les techniques les plus avancées permettent de séparer deux notes simultanées dans les bons cas [CHA 85, CHE 99, MAH 89, MAH 90].

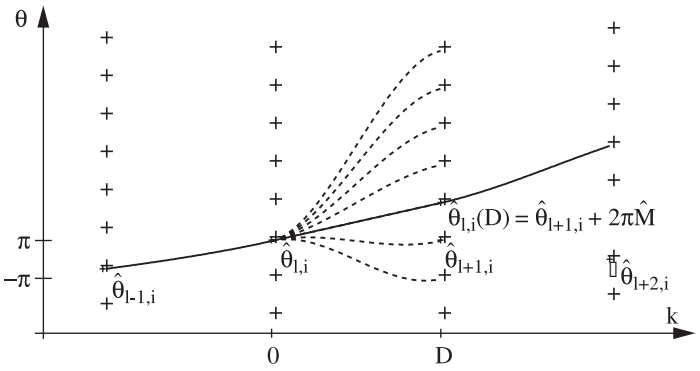

Figure 1.8. *Interpolation cubique de la phase. La phase interpolée est en trait plein. Puisque la phase est définie à un multiple de 2π près, la phase interpolée est la courbe la plus régulière passant par les croix. Par exemple, entre la trame l et la trame $l+1$, les courbes en pointillés représentent différentes possibilités d'interpolation à $2\pi M$ près, pour $M = -1, \ldots, 5$, l'optimum étant ici atteint pour $\hat{M} = 1$.*

1.5.1. *Fréquence fondamentale*

La notion de fréquence fondamentale est liée à la notion de périodicité. C'est donc une propriété du signal, à ne pas confondre avec la notion de « hauteur » qui en est le corrélat perceptif.

Un signal $s(t)$ est périodique s'il existe $T > 0$ tel que $s(t+T) = s(t)$, $\forall t$. Dans le cas où $s(t)$ n'est pas constant, il existe une plus petite valeur T_0 vérifiant cette propriété. La valeur T_0 est appelée période fondamentale et $F_0 = 1/T_0$ fréquence fondamentale ou « fondamental » tout court. Cependant, l'ensemble des valeurs vérifiant

la propriété de périodicité est l'ensemble des multiples de T_0 : $\forall k$, kT_0 est une période du signal (F_0/k est parfois appelée sous-harmonique). La période fondamentale est donc la plus petite des périodes.

C'est de cette définition que viennent toutes les difficultés liées à l'estimation du fondamental. En effet, une variation infime peut changer la période fondamentale en son double (ou en tout multiple k) : il suffit de modifier une valeur toutes les deux (ou k) périodes. Il en résulte une ambiguïté essentielle, se traduisant par des erreurs d'octave lors de la détermination du fondamental.

De plus, les signaux réels ne sont jamais strictement périodiques, donc la détermination du fondamental ne peut pas être fondée uniquement sur cette définition. Les source de variations sont nombreuses : variations de la fréquence fondamentale elle-même (vibrato, glissando), d'amplitude et tout autre paramètre de production (par exemple pour la voix, le conduit vocal et la qualité de la source sont très variables). Il en résulte l'impossibilité de définir le fondamental de façon unique.

Les algorithmes vont donc tenter d'élargir la notion de périodicité stricte, en exploitant les notions proches : la ressemblance entre signaux, l'harmonicité, etc. Cependant, il est très courant que les idées développées bouclent : elles nécessitent la connaissance de la fréquence fondamentale pour l'estimer...

Une dernière remarque est que si l'on peut faire l'hypothèse que le fondamental est situé dans un intervalle d'octave, n'importe quelle technique élémentaire parviendra à l'estimer.

1.5.2. *Concevoir un algorithme de détermination du fondamental*

Le schéma général se décompose en trois étapes : prétraitement, traitement principal et post-traitement. Le prétraitement consiste à mettre en forme le signal acoustique, c'est-à-dire le numériser et lui appliquer certaines transformations globales de type filtrage (passe-bas, filtrage inverse, *clipping*, etc.). Puis, dans le cas de méthodes à court terme, une représentation de ce signal est obtenue par application d'une transformation : représentation temporelle à court terme (extraction de trames successives avec recouvrement et fenêtrage éventuels), représentation fréquencielle ou cepstrale à court terme (transformation temps-fréquence). Le traitement principal peut être vu en général comme le calcul d'une fonction de périodicité $FP(f_0)$ utilisant les informations fournies par le prétraitement, suivi d'une décision sur le meilleur candidat. La fonction de périodicité mesure pour chaque fondamental candidat f_0 à quel point il « explique » la périodicité du signal. La décision devrait se résumer à la détermination d'un maximum : $\widehat{f_0} = \max_{f_0} FP(f_0)$. Cependant, la plupart des méthodes conservent l'ambiguïté d'octave entre les différents candidats. Par exemple, l'autocorrélation d'un signal périodique est périodique de même période. Ceci conduit presque toujours à

corriger *a posteriori* les résultats de $FP(f_0)$ ou à prendre une décision tenant compte à la fois de $FP(f_0)$ et de f_0. Enfin, le post-traitement sert à rendre les courbes de f_0 plus lisses grâce à des techniques de lissage (par exemple filtrage médian) ou de suivi (programmation dynamique ou modèles de Markov cachés).

Quel que soit l'algorithme choisi pour déterminer le fondamental, un certain nombre de paramètres sont déterminants. Ce sont :

1) $F_{0\min}$: la plus petite fréquence fondamentale explorée. C'est le paramètre le plus critique. Il influence le taux d'erreurs grossières par sous-octaviation et doit donc être choisi « au plus juste » ;

2) $F_{0\max}$: la plus grande fréquence fondamentale explorée ;

3) $F_{\max}$: la plus haute fréquence utile du signal. Le nombre maximal d'harmoniques prises en considération dépendra donc du fondamental et sera égal à $F_{\max}/F_{0\min}$ pour $F_0 = F_{0\min}$ et à $F_{\max}/F_{0\max}$ pour $F_0 = F_{0\max}$.

Pour évaluer les tendances d'un algorithme à l'octaviation, il suffit de considérer un signal parfaitement périodique de fondamental F_0 et d'examiner les réponses de l'algorithme aux fréquences multiples et sous-multiples de F_0. La grande majorité des algorithmes donnera, à la sortie du traitement principal, une réponse identique (ou presque) pour le vrai fondamental et pour ses sous-multiples. Les algorithmes qui traitent ce problème dans leur principe sont parmi les plus robustes.

1.5.3. *Prétraitements*

Le but du prétraitement est de mettre en évidence la structure périodique du signal par l'intermédiaire d'une représentation bien choisie, en supprimant autant que possible l'information « inutile » et en renforçant (ou en extrayant) l'information « utile ».

La première idée qui vient à l'esprit est donc l'utilisation d'une procédure de décomposition du signal en parties périodique et apériodique qui permettrait de se concentrer uniquement sur la partie périodique. Malheureusement, ces procédures nécessitent en général l'information du fondamental pour fonctionner et ne sont donc pas utilisables pour la déterminer.

En fait, l'essentiel de l'information concernant le fondamental étant contenu dans les basses fréquences (ou du moins dans les premières harmoniques), le prétraitement le plus utilisé est un filtrage passe-bas souvent associé à un sous-échantillonnage. Ceci se justifie par le fait que l'amplitude des harmoniques suit une tendance décroissante avec le numéro d'harmonique et que, parallèlement, les composantes de bruit ont une répartition spectrale plutôt située du médium à l'aigu. La fréquence de coupure du filtre passe-bas est choisie en fonction de la plus haute fréquence fondamentale du

signal analysé (elle doit au moins lui être supérieure) et du nombre d'harmoniques « utiles » (une heuristique consiste à prendre $F_{\max} = 3$ à 5 fois $F_{0\max}$).

Pour certains algorithmes fonctionnant sur la représentation temporelle du signal, il peut être très utile de supprimer la « composante continue ». Pour ce faire, il suffit d'appliquer sur le signal un filtrage passe-haut avec une faible fréquence de coupure (inférieure à $F_{0\min}$).

Dans certains cas, le modèle de production source/filtre peut s'appliquer. Ce modèle indique que le signal sonore est produit par filtrage (le résonateur) d'un signal de source (l'excitation) ; dans ce cas, l'information relative au fondamental est portée par le signal de source. Il est donc plus intéressant d'estimer le fondamental sur le signal de source, dont une estimation peut être obtenue à partir du signal sonore par des procédures de « filtrage inverse ». Une telle procédure est décrite au paragraphe 1.6.2. Le modèle de production source/filtre est très bien adapté aux signaux de parole et la grande majorité des algorithmes de détermination du fondamental utilise un filtrage inverse comme prétraitement. Il convient aussi à la voix chantée tant que le fondamental n'est pas trop aigu. Pour les signaux instrumentaux, ce prétraitement ne présente d'intérêt que si la source est riche en harmoniques. Le choix de l'ordre du modèle est critique : pour la voix, il peut être choisi égal à deux fois la fréquence maximale exprimée en kHz plus 2 (par exemple, si la fréquence d'échantillonnage est de 16 kHz, choisir un ordre de 18) [4].

Enfin, pour éliminer l'influence des variations d'amplitude sur le signal, il est possible d'effectuer un contrôle automatique de gain, par exemple en multipliant le signal par l'inverse de son enveloppe temporelle (voir section 1.3).

Bien entendu, ces différents prétraitements peuvent être utilisés simultanément.

1.5.4. *Méthodes temporelles*

Puisque la fréquence fondamentale d'un son périodique est celle de son harmonique 1, la première idée qui vient à l'esprit est d'extraire la fréquence de cette harmonique, par exemple par simple filtrage passe-bas. Bien entendu, le choix de la fréquence de coupure est critique et nécessite, pour être optimal, de connaître approximativement la fréquence fondamentale...

Une idée très répandue est de compter directement sur le signal des événements qui se répètent comme les passages par zéro ou par un (ou deux) seuils adaptatifs (en comptant uniquement les fronts montants pour éliminer les effets de composante

4. Cela correspond à associer une paire de pôles de la fonction de transfert à chaque formant, en supposant qu'il y a un formant tous les 1 000 Hz.

continue). Cela ne fonctionne que si le signal analysé est presque sinusoïdal, par exemple s'il ne contient essentiellement que l'harmonique 1. L'idée développée par Kuhn [KUH 90] est de décomposer le signal sonore en le passant dans un banc de filtres en octave (après un contrôle automatique de gain). Sur chaque sortie de filtre est mesurée l'amplitude $A(i)$ et la période $T(i)$ par un simple comptage des passages par zéro. Il reste alors à prendre la décision définitive du fondamental en fonction du résultat des amplitudes et périodes indiquées par chaque filtre.

L'algorithme de décision est le suivant :

$A_{\max} \leftarrow \max_i\big(A(i)\big)$
Si $A_{\max} < seuil\_silence$
  $T_0 \leftarrow -1$
Sinon
  $seuil \leftarrow A_{\max}/4$
  $T_0 \leftarrow -1$
  $convient \leftarrow FAUX$
  $filtre \leftarrow 0$
  Tant que $filtre < n$ et $convient = FAUX$
    Si $A(filtre) > seuil$
      Si $T(filtre)$ est raisonnable pour $filtre$
        $T_0 \leftarrow T(filtre)$
      Sinon si $filtre < n-1$ et $T(filtre+1)$ raisonnable pour $filtre+1$
        $T_0 \leftarrow T(filtre+1)$
      Sinon
        $T_0 \leftarrow -1$
      $convient \leftarrow VRAI$
    Sinon
      $filtre \leftarrow filtre + 1$

Cet algorithme révèle clairement toute l'ambiguïté de la définition de la période fondamentale : il choisit la plus petite période (la première ou éventuellement la deuxième) dont l'amplitude est parmi les plus grandes.

Certaines méthodes analysent des points remarquables du signal (minimums, maximums, passages par zéro) [COO 96, GOL 69].

L'inconvénient majeur des méthodes temporelles est leur mauvaise robustesse aux bruits. Leur avantage est d'être très rapides.

1.5.5. *Méthodes temporelles à court terme*

L'idée sous-jacente des méthodes temporelles à court terme est d'exploiter la ressemblance entre deux tranches de signal séparées par une période. La période estimée sera définie par l'écart de temps entre ces deux tranches. Le prétraitement consiste donc au moins à faire une analyse à court terme du signal, produisant les trames s_l sur lesquelles sera estimée une valeur de fondamental par trame.

Une mesure courante de ressemblance est l'autocorrélation du signal. L'algorithme le plus simple estime le fondamental comme le maximum de la fonction de périodicité suivante :

$$FP_{Autoc,l}(\tau) = \frac{1}{Norme} \sum_{k=0}^{M-1-\tau} s_l(k) s_l(k+\tau)$$

et :

$$f_0 = F_e/\tau$$

où la normalisation par $Norme = (\sum_{k=0}^{M-1-\tau} s_l^2(k) \sum_{k=0}^{M-1-\tau} s_l^2(k+\tau))^{1/2}$ permet de comparer la fonction de périodicité à un seuil absolu pour décider si la trame correspond à un signal périodique ou non. Comme nous l'avons déjà mentionné, cette fonction de périodicité étant périodique de même période que le signal, les deux approches – qui sont d'extraire simplement le maximum ou de choisir la période du premier maximum local de cette fonction – donnent rarement le bon résultat. La plupart des implémentations de cet algorithme proposent soit un mélange de ces deux approches, soit de pondérer la fonction de périodicité (par une exponentielle décroissante, par exemple) pour défavoriser les multiples de la période fondamentale.

Une approche un peu différente et qui donne en général de meilleurs résultats est de choisir le minimum de l'AMDF ou *average magnitude difference function*, qui consiste à faire la différence entre une trame du signal et une version décalée de cette trame :

$$FP_{AMDF,l}(\tau) = \frac{1}{Norme} \sum_{k=0}^{M-1} |s_l(k) - s_l(k+\tau)|$$

et :

$$f_0 = F_e/\tau$$

avec par exemple $Norme = \sum_{k=0}^{M-1} |s_l(k)|$. En théorie, il y aura un minimum à tous les multiples de la période fondamentale. Mais lorsque le signal n'est pas tout à fait périodique, plus le décalage est grand, moins la trame décalée ressemblera à la trame

de départ et plus l'AMDF sera grande. Ainsi, l'algorithme défavorise de façon implicite les périodes multiples. Cependant, des heuristiques sont presque toujours utilisées pour défavoriser les grandes périodes [CHE 91].

Une extension intéressante des fonctions précédentes consiste à calculer la ressemblance sur une fenêtre de taille variable, correspondant par exemple à la période candidate. C'est ce que propose l'algorithme SRPD (*super resolution pitch determination*) [MED 91], fondé sur une intercorrélation entre deux fenêtres successives de taille égale à la période candidate, normalisée pour tenir compte de la variation du nombre de points :

$$FP_{SRPD,l}(\tau) = \frac{\sum_{k=0}^{\tau-1} s_l(k)s_l(k+\tau)}{\left(\sum_{k=0}^{\tau-1} s_l^2(k)\sum_{k=0}^{\tau-1} s_l^2(k+\tau)\right)^{\frac{1}{2}}}$$

et :

$$f_0 = F_e/\tau$$

Quelle que soit la méthode choisie, un des problèmes liés à la discrétisation est la précision de la mesure : la période est mesurée par un nombre de points de décalage, donc la précision est limitée par la période d'échantillonnage du signal. Par exemple, à $F_e = 16\,000$ Hz sur un Do aigu ($F_0 = 1\,046$ Hz), la précision sur la période étant d'un échantillon, soit $0{,}0625$ ms, la fréquence sera déterminée au mieux à ± 33 Hz, donc à un demi-ton près. Une solution est d'interpoler la fonction de périodicité au voisinage du maximum (/minimum) correspondant à la période estimée [GEO 96, MED 91] ou d'utiliser plusieurs fenêtres successives [BRO 89].

1.5.6. *Méthodes fréquencielles à court terme*

Fréquenciellement, la périodicité se traduit par le fait que les partiels du signal sont harmoniques. Un grand nombre de méthodes exploitent cette information en travaillant sur le spectre du signal.

Le principe de ces méthodes est donc d'effectuer une analyse spectrale à court terme (TFCT, voir paragraphe 1.2.3) fournissant un spectre à court terme $\widetilde{s}_l$ par trame $l = 0, \ldots, L-1$ et, pour chaque trame, de calculer une fonction de périodicité $FP(f_0)$ sur $\widetilde{s}_l$ et d'en déduire l'estimation d'une fréquence fondamentale (par exemple par l'indice du maximum de $FP(f_0)$).

Le spectre peut être exploité directement ou n'être qu'une étape pour l'estimation des partiels (voir section 1.4). Un certain nombre de prétraitements du signal et du spectre sont souvent utilisés pour renforcer, isoler ou égaliser les harmoniques.

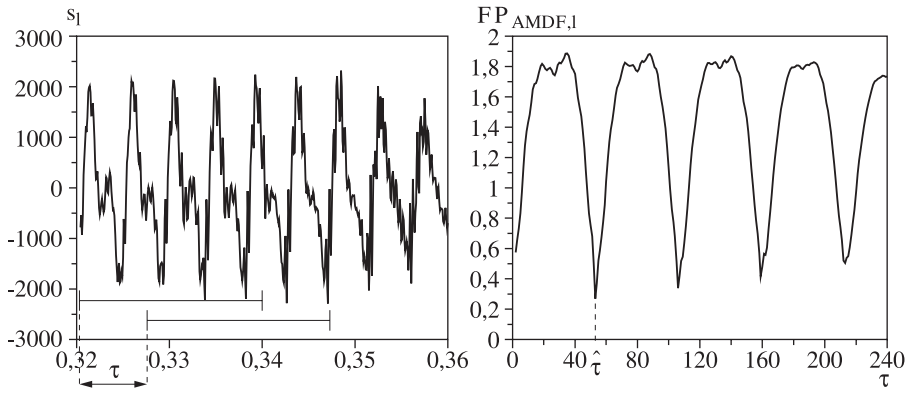

Figure 1.9. *Estimation de la fréquence fondamentale par AMDF. Une trame de signal (à gauche) est comparée avec une version décalée de τ pour fournir un point de la fonction de périodicité (à droite). Le décalage qui provoque la plus petite valeur d'AMDF est considéré comme la période du signal. Ici, il s'agit d'un signal de parole à 12 kHz, analysé pour des périodes allant de 30 à 400 points (correspondant à des f_0 de 50 Hz à 400 Hz). La période estimée est ici de $\hat{\tau} = 53$, soit $\widehat{f_0} = 226\,Hz$ +/− $4\,Hz$. Remarquez la présence de minimums locaux importants à tous les multiples de $\hat{\tau}$.*

Le principe de la compression spectrale consiste à faire la somme du spectre et de toutes les versions « compressées » d'un facteur entier de ce spectre. Ainsi, toutes les contributions du spectre aux positions des harmoniques se retrouvent sommées à la position du fondamental.

Quant à l'appariement d'harmoniques, son principe est d'apparier les partiels du signal à des valeurs de référence choisies pour être les harmoniques d'un candidat au fondamental. Le spectre de référence a donc la forme d'un peigne. Une mesure pour chaque candidat f_0 est alors déduite de l'appariement effectué pour ce f_0.

L'idée la plus simple consiste, pour un candidat au fondamental f_0 donné, à faire l'intercorrélation entre le spectre et le peigne de fondamental f_0 :

$$FP_l(f_0) = \int_f \perp\!\!\!\perp_{f_0}(f)|\widetilde{s_l}(f)|\,df = \sum_i |\widetilde{s_l}(if_0)|$$

Notons que la somme est souvent remplacée par un produit, ce qui revient à calculer la somme sur le spectre exprimé en dB.

Il est remarquable que l'intercorrélation avec un peigne est équivalente à la somme des spectres compressés. Nous ne développerons que la première.

L'algorithme présenté par Martin [MART 81, MART 82] est fondé sur ce principe d'intercorrélation avec une fonction peigne avec quelques améliorations notables. Tout d'abord, le spectre étant discret, la fonction de périodicité est calculée sur les points du spectre d'indice n_0 entier :

$$FP_{IFP,l}(n_0) = \sum_{i=1}^{I(n_0)} \alpha_i |\widetilde{s_l}(in_0)|$$

le fondamental candidat étant $f_0 = n_0 F_e/N$. Ensuite, comme l'indique la formule précédente, les dents du peigne sont pondérées par les coefficients rapidement décroissants $\alpha_i = i^\gamma$, $\gamma < 0$. Ce point est critique pour la robustesse de l'algorithme. Une valeur de $\gamma = -0{,}5$ convient pour la plupart des applications.

Contrairement à ce que l'on pourrait penser, la décroissance des dents du peigne n'a rien à voir avec la tendance naturelle de décroissance des spectres des signaux musicaux. En fait, si les spectres avaient des amplitudes croissantes, il faudrait quand même prendre un peigne décroissant. La décroissance du peigne permet de défavoriser les sous-multiples du vrai fondamental : le candidat de fréquence moitié du vrai fondamental sommera toutes les harmoniques, mais avec les coefficients α_{2i} qui sont inférieurs aux α_i.

De plus, dans l'algorithme de Martin, seuls les partiels importants du signal sont pris en considération dans le calcul. Pour cela, la méthode proposée est, plutôt que de prendre tous les points du spectre, d'effectuer une sélection des pics du spectre candidats aux harmoniques (voir paragraphe 1.4.2) et de reconstruire un spectre en n'utilisant que la fréquence et l'amplitude de ces pics. Pour tenir compte de l'inharmonicité éventuelle, les pics sont reconstruits avec une forme de parabole. Dans la formule précédente, $\widetilde{s_l}$ doit donc être remplacé par le spectre reconstruit.

Enfin, la précision sur la détermination de la fréquence fondamentale est limitée par l'intervalle de fréquence entre deux points du spectre. Une méthode consiste à suréchantillonner le spectre reconstruit pour obtenir la précision voulue, par exemple en utilisant des splines [HER 88]. De plus, l'appariement d'harmoniques (en associant par exemple l'harmonique k avec le pic le plus proche de $k\,\widehat{f_0}$) permet de réestimer le fondamental par régression sur les fréquences des harmoniques pondérées par leurs amplitudes :

$$\widehat{f}_{0\text{précis}} = \frac{\sum_k \hat{A}_{l,i(k)} \hat{f}_{l,i(k)}/k}{\sum_k \hat{A}_{l,i(k)}}$$

où $i(k)$ désigne le numéro du pic apparié à l'harmonique k.

De nombreuses autres méthodes entrant dans la même catégorie ont été développées, essentiellement pour la parole. Paliwal et Rao [PAL 83] proposent de mesurer la distance entre les partiels du spectre et un peigne dont les amplitudes sont

situées sur une enveloppe spectrale estimée sur ce spectre. La méthode de Duifhuis *et al*. [DUI 82] est fondée sur la théorie perceptive de Goldstein et sur la notion de crible harmonique : c'est une sorte de « tamis » qui rejette tout ce qui n'est pas harmonique à une précision près. Dans le cas de sons riches en harmoniques, le maximum du cepstre dans une zone de valeurs possibles correspond à la période fondamentale [NOL 67]. Il existe aussi des approches utilisant des connaissances plus spécifiques sur les sons étudiés : Doval [DOV 94] propose d'effectuer un apprentissage statistique des caractéristiques des harmoniques (nombre, amplitudes, inharmonicités, ce qui permet d'analyser des spectres à partiels inharmoniques) et du « bruit », Sieger et Tewfik [SIE 98] utilisent un dictionnaire d'harmoniques par fondamental.

1.5.7. *Suivi de la fréquence fondamentale*

Le besoin des utilisateurs étant de disposer d'une valeur de fondamental à chaque instant, la décision habituellement prise est locale et consiste à extraire le maximum de la fonction de périodicité. Les erreurs les plus courantes sont alors des erreurs grossières (d'octave) localisées, typiquement un aller-retour entre le fondamental et sa moitié ou parfois son double. Or, il est clair qu'aucun instrument ne pourra produire une telle séquence de fréquences fondamentales. La présence de telles erreurs est due au fait que l'information sur l'évolution du fondamental dans le temps n'est pas utilisée.

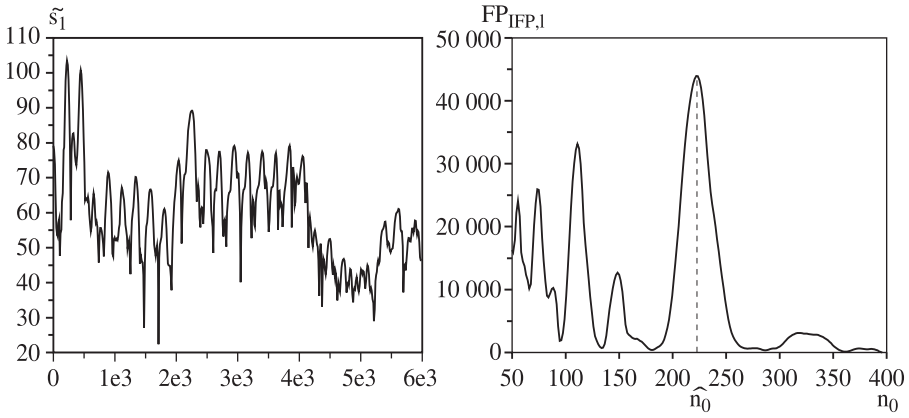

Figure 1.10. *Estimation de la fréquence fondamentale par intercorrélation avec une fonction peigne. Le spectre est corrélé avec un peigne dont les dents sont écartées de n_0 et le résultat fournit un point de la fonction de périodicité. La valeur fournissant la meilleure corrélation ($\widehat{n_0}$) correspond à la fréquence fondamentale. Les conditions sont identiques à celles de la figure 1.9 (même signal) et la fréquence estimée est $\widehat{f_0} = \widehat{n_0} F_e/N = 223$ +/− 1 Hz. Remarquez la présence de maximums locaux atténués aux sous-multiples de $\widehat{f_0}$.*

Le suivi de fréquence fondamentale est alors une technique permettant d'améliorer très nettement les résultats d'estimation. Une des solutions parmi les meilleures est le suivi par programmation dynamique ou, dans sa version mieux formalisée et présentée ici, par modèles de Markov cachés. Ils permettent de prendre une décision globale fournissant une séquence de valeurs du fondamental optimale au sens où elle tient compte de l'évolution possible du fondamental entre deux instants successifs tout en expliquant bien les caractéristiques observées à un instant donné. Une description du fonctionnement et des algorithmes des modèles de Markov cachés peut être trouvée dans [PAP 84, RAB 86].

La structure du modèle de Markov caché (MMC) est un ensemble d'états. Chaque état est associé à un fondamental possible (candidat). On parlera donc de l'état f_0. Tous les états du MMC sont connectés entre eux, ce qui signifie qu'il est *a priori* possible de passer d'un fondamental à n'importe quel autre entre deux instants successifs.

Le fonctionnement direct du MMC est d'émettre le signal observé (celui dont on cherche à connaître la séquence des fréquences fondamentales) trame par trame, de la façon suivante : à l'instant t_l, le MMC se trouve dans l'état f_0 et émet la trame de signal s_l (ou plus exactement sa représentation), puis il effectue une transition vers un autre état (ou vers le même état) et passe à l'instant suivant. Ce processus est itéré jusqu'à obtenir toutes les trames de signal.

Pour pouvoir fonctionner, le MMC a donc besoin, d'une part, de la densité de probabilité d'apparition des observations dans chaque état (la fonction de périodicité $FP_l(f_0)$ est considérée comme une mesure de cette densité) et, d'autre part, de la probabilité de transition d'un état dans un autre, notée $tr(f'_0, f_0)$, qui est justement l'information qui traduit l'évolution du fondamental d'un instant à l'autre. Cette probabilité pourra être apprise (apprentissage statistique paramétrique ou non) ou modélisée *a priori*.

Il apparaît donc que le MMC est un modèle d'émission d'un signal à partir d'une séquence de fréquences fondamentales. Or, le suivi est exactement le problème inverse : celui de retrouver la séquence des fréquences fondamentales connaissant le signal observé. En termes de Markov, le problème est de trouver la séquence optimale des états qui ont pu émettre le signal donné.

L'algorithme permettant de résoudre ce problème est l'algorithme de Viterbi dont nous présentons ici le détail. Considérons le tableau FP dont les abscisses représentent les instants t_l d'analyse des trames de signal, dont les ordonnées représentent les états possibles f_0. Remplissons ce tableau par colonne avec les probabilités d'émission de la trame du temps t_l par l'état f_0 (ce sont les valeurs des fonctions de périodicité de l'instant correspondant, $FP_l(f_0)$). Considérons les chemins qui passent par une et une seule case du tableau à chaque instant et définissons la probabilité d'un tel chemin par le produit de toutes les probabilités d'émission correspondant aux cases traversées

par le chemin et des probabilités de transition entre deux cases successives. Alors, l'algorithme trouve le chemin optimal au sens où il maximise sa probabilité.

Bien entendu, l'algorithme n'explore pas tous les chemins possibles mais construit de proche en proche le tableau noté $\boldsymbol{C}[l, f_0]$ des probabilités de tous les sous-chemins optimaux et choisit celui dont la probabilité finale est maximale. Enfin, pour pouvoir récupérer effectivement la séquence d'états de ce chemin, à chaque transition, l'état précédent est mémorisé dans $\mathbf{prec}[l, f_0]$. En pratique, il y a quatre étapes :

– initialisation : Pour $f_0 \leftarrow F_{0\min}$ à $F_{0\max}$, $\boldsymbol{C}[0, f_0] = FP_0(f_0)$

– propagation : Pour $l \leftarrow 1$ à $L-1$
  Pour $f_0 \leftarrow F_{0\min}$ à $F_{0\max}$
    $\boldsymbol{C}[l, f_0] = \max_{f'_0}\big(\boldsymbol{C}[l-1, f'_0] tr(f'_0, f_0)\big) FP_l(f_0)$
    $\mathbf{prec}[l, f_0] = \arg\max_{f'_0}\big(\boldsymbol{C}[l-1, f'_0] tr(f'_0, f_0)\big)$

– terminaison :
  $proba_{opt} = \max_{f_0}\big(\boldsymbol{C}[L-1, f_0]\big)$
  $\widehat{f_0}[L-1] = \arg\max_{f_0}\big(\boldsymbol{C}[L-1, f_0]\big)$

– recherche arrière : Pour $l \leftarrow L-2$ à 0, $\widehat{f_0}[l] = \mathbf{prec}\big[l+1, \widehat{f_0}[l+1]\big]$

La séquence optimale des valeurs de fréquence fondamentale est alors :

$$\widehat{\boldsymbol{f_0}} = \big(\widehat{f_0}[0], \widehat{f_0}[1], \ldots, \widehat{f_0}[L-1]\big)$$

Il est clair que dans son principe, cette méthode de suivi peut s'adapter à n'importe quel algorithme de détermination du fondamental. Cependant, en toute rigueur, il faut transformer au préalable la fonction de périodicité FP en densité de probabilité. La pratique la plus courante est de pondérer l'importance relative de FP et des probabilités de transition tr pour équilibrer les contraintes de régularité d'évolution et d'adéquation locale.

Un choix standard pour les probabilités tr est une gaussienne sur l'écart relatif des fréquences :

$$tr(f'_0, f_0) = \frac{e^{-0{,}5(\frac{f'_0 - f_0}{0{,}5(f'_0 + f_0)}/\sigma)^2}}{\sigma\sqrt{2\pi}}$$

S'il n'est pas prévu d'apprentissage, une heuristique consiste à choisir σ de l'ordre de T_a ($\sigma = 0{,}4T_a$ convient à l'utilisation directe de l'algorithme décrit au paragraphe 1.5.6).

Le désavantage de cet algorithme est qu'il ne fonctionne pas en temps réel. Il est toujours possible de l'utiliser par bloc, mais des versions sous-optimales en temps réel existent [RAP 99].

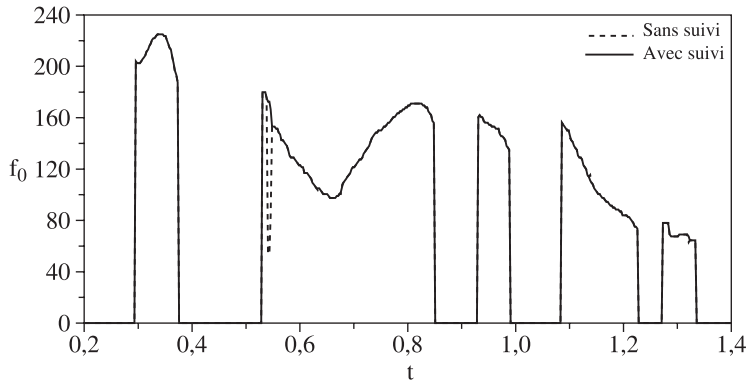

Figure 1.11. *Suivi de fréquence fondamentale. Cette phrase a été analysée avec l'algorithme de Martin qui a fourni une fonction de périodicité pour chaque trame, puis l'algorithme de suivi décrit dans le texte a été appliqué sur les fonctions de périodicité. Remarquez que les erreurs grossières locales sont corrigées (exemple à l'instant $t = 0{,}54\,s$). Sur cet exemple, les parties de signal où la fréquence fondamentale n'est pas définie ont été mises à zéro.*

1.6. Analyse de l'enveloppe spectrale

Pour la plupart des instruments, le son est produit par un excitateur, puis mis en forme par un résonateur. Le signal de l'excitateur (la source) est composé d'harmoniques dont l'amplitude est généralement régulièrement décroissante en fonction du numéro d'harmonique. De son côté, le résonateur agit comme un filtre dont la réponse en fréquence est une fonction continue de la fréquence présentant des maximums et des vallées qui vont « modeler » l'amplitude des harmoniques produites par la source. L'amplitude d'une harmonique est alors la somme (en dB) des amplitudes de la composante correspondante dans la source et de la réponse du filtre.

Le nombre d'harmoniques significatives d'un signal musical peut être parfois très grand et la fidélité de l'analyse nécessite de les estimer à court terme, ce qui produit un nombre énorme de paramètres à conserver pour d'éventuelles transformations ou même juste pour le stockage. L'enveloppe spectrale est alors un moyen économique de codage de l'amplitude des harmoniques en fonction de la fréquence par un jeu de coefficients en nombre très réduit (de l'ordre d'une vingtaine pour des signaux très complexes). Bien entendu, une enveloppe spectrale peut être calculée aussi bien sur des signaux de bruit que sur des signaux pseudo-périodiques.

Les utilisations de l'enveloppe spectrale en musique sont variées [LAN 89] : modification du fondamental d'une voix sans changer la qualité vocalique ni le débit, modification du débit, transformation de la source seule (en voix chuchotée ou

alors parfaitement tonale), synthèse croisée ou encore toute application nécessitant la mesure de distance de « timbre » [DEP 94].

1.6.1. *Enveloppe spectrale*

Une enveloppe spectrale paramétrique peut être vue comme une fonction appartenant à une famille de fonctions (spline, réponse en fréquence de filtres linéaires, cepstres, etc.) décrite par un jeu de paramètres. Une enveloppe est donc associée de façon biunivoque à un vecteur de paramètres.

Le principe général de l'estimation d'enveloppe est alors de déterminer le jeu de paramètres qui minimise un critère d'erreur mesurant l'écart entre l'enveloppe et les points du spectre. D'un point de vue plus pratique, il s'agit de faire passer une courbe régulière par ces points ou proche de ces points.

Un cas plus simple est celui des enveloppes linéaires par morceaux (non paramétriques), définies par un petit nombre de points spectraux et dont l'estimation est directe, puisqu'il suffit de trouver le maximum du spectre par bandes de fréquences [SER 90].

1.6.2. *Estimation par prédiction linéaire (LPC)*

Quand le modèle source/filtre s'applique bien au signal analysé, il est naturel d'estimer l'enveloppe spectrale par la réponse en fréquence de la partie filtre. Généralement, le modèle fait l'hypothèse sur la source qu'elle est composée d'impulsions pseudo-périodiques ou qu'il s'agit de bruit blanc. Son spectre est donc plat (c'est-à-dire que la pente spectrale est de 0 dB/oct). De plus, le filtre est supposé tout-pôle (autorégressif) et sa fonction de transfert est donc de la forme $H(z) = G/\sum_{i=0}^{p} a(i)z^{-i}$, où les $a(i)$ sont les coefficients du filtre, d'ordre p, et où $a(0) = 1$.

Cependant, les sources des instruments de musique ont plutôt des spectres décroissant en $1/f$ (la pente spectrale est de -6 dB/oct) ou même décroissant plus vite. Il est alors utile de « préaccentuer » le signal par une fonction qui ajoute une pente de $+6$ dB/oct pour qu'il satisfasse mieux aux hypothèses du modèle, par exemple par l'application d'un filtre dérivateur du type $x(k) = s(k) - \alpha s(k-1)$, où $\alpha < 1$, par exemple $\alpha = 0{,}976$.

Pour estimer les coefficients du filtre, la technique la plus utilisée, la prédiction linéaire (LPC en anglais), a été développée pour des applications de codage de la parole. C'est une technique des moindres carrés appliquée au critère d'erreur de prédiction des échantillons de signal : $err = \sum_{k=0}^{M-1} (s(k) - \hat{s}(k))^2$, où

$\hat{s}(k) = \sum_{i=1}^{p} a(i)s(k-i)$ est le signal de prédiction. Le lecteur trouvera une description détaillée des développements théoriques et algorithmiques dans [MAR 76]. Cependant, nous donnons ici un algorithme permettant de calculer les coefficients du filtre et son gain. Il s'agit de la méthode d'autocorrélation associée à l'algorithme de Durbin.

Considérons une trame de signal $s_l(k)$, $k = 0, \ldots, M-1$ issue d'une analyse à court terme (voir paragraphe 1.2.3) dont nous souhaitons estimer l'enveloppe spectrale. L'algorithme consiste à estimer les coefficients de corrélation $R(i)$ de cette trame, puis les coefficients du filtre $a(i)$ et son gain G :

– $R(i) = \sum_{k=0}^{M-i-1} s_l(k)s_l(k+i)$, $i = 0, \ldots, p$;

– algorithme de Durbin :

$err \leftarrow R(0)$
$a(0) \leftarrow 1$
Pour $i \leftarrow 1$ à p
  $a(i) \leftarrow \left(-\sum_{j=0}^{i-1} a(j)R(i-j)\right)/err$
  Pour $j \leftarrow 1$ à $i-1$
    $b(j) \leftarrow a(j) + a(i)a(i-j)$
  Pour $j \leftarrow 1$ à $i-1$
    $a(j) \leftarrow b(j)$
  $err \leftarrow \left(1 - a(i)^2\right) err$
Les $a(i)$ sont les coefficients du filtre et err est l'erreur de prédiction

– calcul du gain : $G = \sqrt{\sum_{i=0}^{p} a(i)R(i)}$.

Une fois estimé le filtre, il est simple d'estimer la source par « filtrage inverse » : $e(k) = \frac{1}{G}\sum_{i=0}^{p} a(i)s_l(k-i)$, $k = 0, \ldots, M-1$ en considérant les échantillons d'indice négatif comme nuls.

Si une préaccentuation a été utilisée avant la LPC, il faut maintenant désaccentuer le résultat du filtrage inverse pour obtenir le signal de source estimé.

Le tracé de l'enveloppe sur N points est simple : il s'agit du module de la réponse en fréquence du filtre. Le calcul peut se faire par TFD, car $ENV(n) = |H(z)|_{e^{j2\pi n/N}}| = G/|\sum_{i=0}^{p} a(i)e^{-j2\pi ni/N}|$ où apparaît au dénominateur la TFD d'un signal composé des coefficients du filtre complété par des zéros (pour le choix de N, voir paragraphe 1.2.3).

L'ordre du filtre est un paramètre très important : plus l'ordre est grand, plus l'enveloppe reproduira les petites variations du spectre. Un ordre trop faible risque de ne

pas reproduire certaines résonances, un ordre trop élevé reproduira des détails tels que les harmoniques.

Cette technique est bien adaptée aux signaux de voix graves ou pour tout instrument dont la structure formantique est claire et distincte de la structure harmonique. Quand le fondamental est élevé, la LPC sera donc avantageusement remplacée par le cepstre discret.

1.6.3. *Estimation par cepstre discret*

Le cepstre est la TF^{-1} du logarithme du module du spectre du signal analysé. Il est constitué d'un vecteur de coefficients cepstraux notés $\boldsymbol{c} = \{c(k)\}$ pour $k = 0, \ldots, N-1$. Réciproquement, étant donné un vecteur $\boldsymbol{c}$ de coefficients $c(k)_{k=0,\ldots,p}$, l'enveloppe cepstrale correspondante est définie par :

$$ENV(f;\boldsymbol{c}) = c(0) + 2\sum_{k=1}^{p} c(k)\cos(2\pi k f/F_e)$$

pour $f \in [0, F_e/2]$.

Elle correspond au logarithme d'un spectre, l'expression $20ENV(f;\boldsymbol{c})/\log(10)$ est donc comparable à un spectre d'énergie en dB : $20\log_{10}|\tilde{s}|$. Le tracé sur N points s'effectue en remplaçant f par nF_e/N pour $n = 0, \ldots, N-1$.

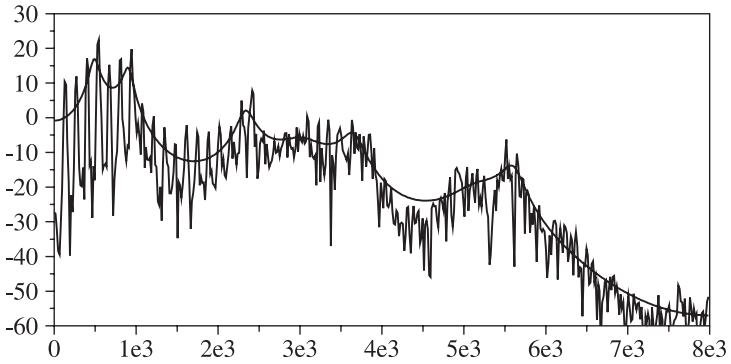

Figure 1.12. *Enveloppe spectrale LPC. Le son d'origine étant une voyelle [o], la figure représente l'enveloppe spectrale superposée à son spectre. Cette enveloppe a été obtenue par l'algorithme de Durbin, le signal ayant été au préalable préaccentué et multiplié par une fenêtre de pondération de Hanning pour le calcul des coefficients d'autocorrélation.*

Pour obtenir l'enveloppe cepstrale d'un signal, il suffit de la calculer sur les premiers coefficients du cepstre : moins il y aura de coefficients, plus l'enveloppe sera

lisse. Cependant, pour des signaux présentant des partiels bien définis, il est préférable d'utiliser le cepstre discret, car il est calculé uniquement sur ces partiels.

L'algorithme est le suivant. Soit les pics spectraux de fréquence $\hat{f}_{l,i}$ et d'amplitude $\hat{A}_{l,i}$, l'enveloppe cepstrale est estimée par minimisation du critère quadratique $err = \sum_{i=1}^{I_l} |\log(\hat{A}_{l,i}) - ENV(\hat{f}_{l,i}; \boldsymbol{c})|^2$ dont la solution directe, obtenue en annulant la différentielle de err par rapport au vecteur $\boldsymbol{c}$, est $\hat{\boldsymbol{c}} = (\boldsymbol{M}^t\boldsymbol{M})^{-1}\boldsymbol{M}^t\boldsymbol{A}$ où $\boldsymbol{A}$ est le vecteur colonne du logarithme des amplitudes $\log(\hat{A}_{l,i})$, $i = 1, \ldots, I_l$ et où $\boldsymbol{M}$ est la matrice à I_l lignes et $p+1$ colonnes de coefficients $\boldsymbol{M}(i,k) = 2\cos(2\pi k \hat{f}_{l,i}/F_e)$, pour $k = 1, \ldots, p$ et $i = 1, \ldots, I_l$ et $\boldsymbol{M}(i,0) = 1$ pour $i = 1, \ldots, I_l$.

La solution obtenue présente en général des oscillations et l'extrapolation (la forme de l'enveloppe en dehors des partiels) est mauvaise. Des techniques de régularisation permettent de supprimer ces inconvénients [CAP 97]. Cette technique présente l'avantage de pouvoir pondérer facilement les différents partiels et comparer différentes enveloppes par distance euclidienne sur les vecteurs de coefficients cepstraux.

1.6.4. *Formants*

Dans le modèle source/filtre, les maximums présentés par l'enveloppe spectrale correspondent à des résonances dues à la partie filtre. La position de ces maximums caractérise en partie le timbre des sons. En particulier pour la voix, ces maximums, appelés formants, caractérisent la voyelle prononcée mais peuvent aussi traduire certains aspects de la qualité de voix comme la perception d'une voix tendue ou relâchée ou certaines techniques de chant comme l'apparition du « formant du chanteur » sur les voix lyriques.

L'estimation de la position des maximums de l'enveloppe spectrale peut être réalisée par une technique de détection de pics analogue à celles développées au paragraphe 1.4.2 pour les partiels. Cependant, lorsque l'enveloppe est déterminée par les coefficients d'un filtre comme dans le cas de la prédiction linéaire, il est possible d'en déduire les pôles (ce sont les zéros du polynôme $\sum_{i=0}^{p} a(i)z^{-i}$) et une approximation de la fréquence des formants grâce au résultat suivant : la fréquence de résonance f_r d'un filtre d'ordre 2 correspondant aux pôles $\rho e^{\pm j\theta}$ vérifie la formule $\cos(2\pi f_r/F_e) = \frac{1}{2}(\rho + 1/\rho)\cos(\theta)$.

1.7. Suivi de notes et applications

Le suivi de notes consiste à transformer le signal acoustique en une représentation musicale du type partition ou plus simplement en signal MIDI. Les informations à extraire sont donc les instants de début et de fin de notes, leurs fréquences et leurs amplitudes.

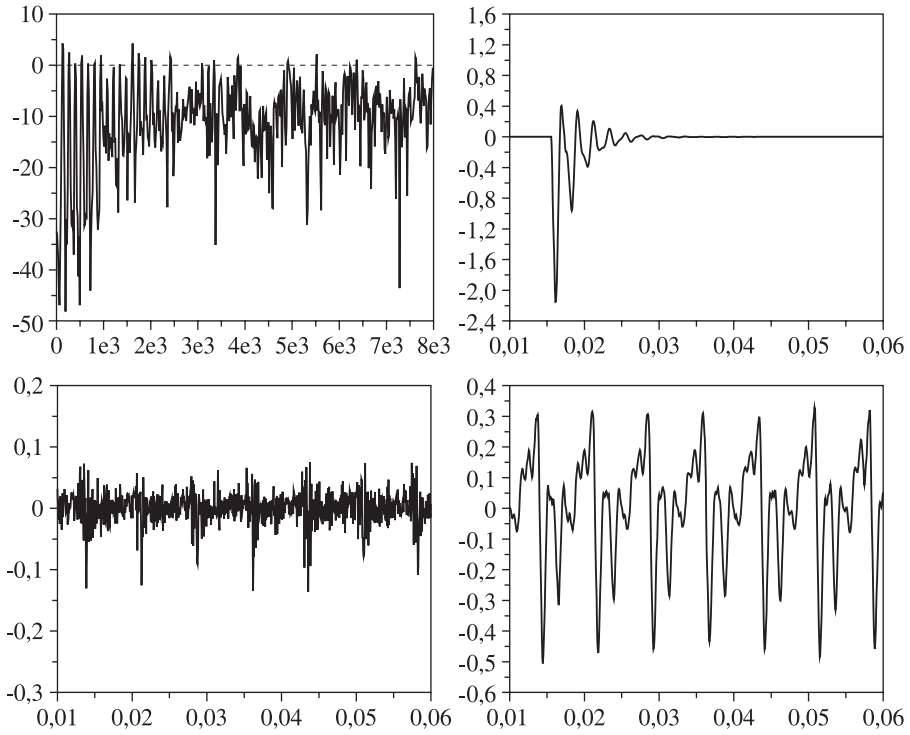

Figure 1.13. *Calcul du résiduel. Sur l'exemple de la figure 1.12, le signal a été filtré par le filtre inverse (correspondant à l'inverse de l'enveloppe spectrale) fournissant le résiduel (en bas à gauche) dont le spectre est plat (en haut à gauche) et qui correspond, en première approximation, à la source sonore. En convoluant ce résiduel par la réponse impulsionnelle du filtre (en haut à droite), on obtient le signal original (en bas à droite). Remarquez que la forme de la réponse impulsionnelle se retrouve sur le signal original à chaque instant où le résiduel contient une impulsion.*

En toute généralité, ce problème est très difficile : comment séparer les différentes notes d'une masse orchestrale ou même d'un petit ensemble ? Comment reconnaître les sons de percussion ? Et même pour un instrument monophonique, la musique contemporaine exploite les possibilités des instruments de telle manière que c'est la notion même de note qui doit être remise en cause. Pour être utilisable, le suivi de note nécessite donc des hypothèses fortes.

Enfin, une difficulté supplémentaire apparaît : la segmentation du signal ne peut pas se faire sur l'amplitude seule à cause des sons filés ou des legatos très lisses, ni sur la fréquence seule à cause des notes répétées. Les algorithmes devront donc combiner les deux sources d'information.

Aux sections 1.3 et 1.5, nous avons présenté des techniques d'estimation de l'amplitude et du fondamental qui fournissent une valeur réelle à chaque instant. Une note MIDI, quant à elle, est décrite par une seule valeur quantifiée d'amplitude (la vélocité) et de fréquence (la hauteur MIDI) pour la durée de la note (les écarts et les variations sont traités à part). Un système de suivi de notes devra donc quantifier la fréquence fondamentale et l'amplitude, et détecter le début et la fin des notes.

1.7.1. *Quantification de la fréquence fondamentale*

La hauteur de note MIDI étant quantifiée par demi-tons, elle s'exprime en fonction de la fréquence en hertz par la formule $q = q_{ref} + Arr(12 \log_2(f/f_{ref}))$, où f_{ref} et q_{ref} sont la fréquence et la hauteur MIDI de référence ($q_{ref} = 69$ pour le diapason standard $f_{ref} = 440$ Hz) et où $Arr(x)$ est l'entier le plus proche de x.

1.7.2. *Détection de début et de fin de notes*

Dans un système développé à l'IRCAM en 1990, l'information des fréquences fondamentales est analysée à court terme ($T_a = 10$ ms) pour déterminer les zones stables, où une zone stable est une suite d'au moins n_0 fréquences fondamentales dont les hauteurs MIDI sont identiques (par défaut $n_0 = 2$). D'autre part, l'information d'amplitude est analysée pour déterminer les attaques et les relâchements. Une attaque est une suite d'amplitudes croissantes supérieures, d'une part, à un seuil de montée relatif à l'amplitude du minimum précédent (d'un facteur $1{,}8$) et, d'autre part, au seuil absolu de bruit (environ -50 dB du maximum). Un relâchement est une suite d'amplitudes décroissantes ayant une vitesse de décroissance supérieure à un seuil de descente (en fait, l'amplitude du point courant doit être inférieure à $0{,}25$ fois le maximum des quatre derniers points) ou ayant une amplitude inférieure au seuil de bruit. L'algorithme prévoit d'envoyer un événement MIDI « note *on* » au début et un « note *off* » à la fin de toute zone d'intersection entre une zone stable et une zone se trouvant entre une attaque et un relâchement. Cet algorithme permet donc de traiter les cas de notes répétées grâce à l'information de changement d'amplitude et aussi les cas de notes legato grâce à l'information de fréquence fondamentale.

Un problème couramment observé est l'apparition de fausses détections dues à la présence de bruit, aux variations internes de fréquences ou d'amplitudes (par exemple le vibrato, difficile à distinguer du trille instrumental) ou même, pour les systèmes multiphoniques, le battement entre des fréquences proches [BOB 98]. Le système décrit ci-dessus prévoit de ne pas répéter de notes si une nouvelle attaque n'est pas générée.

D'autres techniques tentent de segmenter les notes en utilisant plus directement les informations conjointes d'énergie et de fréquence. Bobrek et Koch [BOB 98] proposent un système fondé sur une décomposition du signal en sous-bandes de fréquences. Pour détecter les attaques, ils comptent le nombre de sous-bandes dépassant un seuil adaptatif et décident qu'une note débute à chaque maximum local de ce compteur. La fréquence et l'amplitude de la note ou des notes est obtenue par comparaison entre le contenu fréquenciel des sous-bandes et des valeurs de référence. La fin des notes est effectuée par seuillage absolu. Ce système a été appliqué avec succès sur le suivi de notes de piano.

Goto et Muraoka [GOT 99] détectent les composantes d'attaque par bande de fréquences : en notant $p(l,n) = \frac{1}{N}|\widetilde{s_l}(n)|^2$ la densité spectrale de puissance du signal, si $\min(p(l,n), p(l+1,n)) > \max(p(l-1,n), p(l-1,n \pm 1))$ alors $p(l,n)$ est une composante d'attaque. Les temps d'attaque sont alors calculés comme les pics de la fonction $D(l) = \sum_n d(l,n)$, où les $d(l,n) = \max(p(l,n), p(l+1,n)) - \max(p(l-1,n), p(l-1,n \pm 1))$ sont les vitesses de montée. Cette procédure s'insère dans un système de suivi de pulsation rythmique de musique sans percussion.

Sieger et Tewfik [SIE 98] décomposent le signal en partiels sinusoïdaux selon la méthode de McAulay et Quatieri [MCA 86] (voir section 1.4) et effectuent une reconstruction de la partie sinusoïdale et de la partie bruit. La détection de début et de fin de note est effectuée, d'une part, sur les partiels par regroupement de tous les partiels se superposant, d'autre part, sur la partie bruit par une technique plus classique de seuillage de la dérivée de l'énergie à court terme.

Raphaël [RAP 99] propose de segmenter le signal musical par l'utilisation de modèles de Markov cachés sur l'ensemble du signal, ce qui rend la décision globale et donc permet la prise en considération de la probabilité de transition entre les notes.

1.7.3. *Suivi de partition*

Une des nombreuses applications du suivi de notes est le suivi de partition. Il s'agit de suivre en temps réel sur une partition l'interprétation qu'en donne un musicien. De ce fait, il est possible de programmer le système pour qu'il réponde au musicien : pour jouer un accompagnement tel qu'il est écrit ou en l'adaptant à l'interprétation du musicien, ou encore pour déclencher n'importe quel événement utilisant ou non les sons produits par le musicien.

Dans son principe, le suivi de partition est un problème d'alignement temporel : savoir où le musicien en est dans la partition revient à aligner les notes jouées avec celles qui sont écrites.

La plupart des approches sont fondées principalement sur la mesure de la hauteur de chaque note et seulement en second lieu sur leur durée et leur écart en temps [BAI 93, DAN 84, VER 84]. Certains vont même jusqu'à utiliser l'apparition d'une nouvelle note pour avancer dans la partition, sans tenir compte de l'instant précis où elle devrait être jouée selon cette partition [PUC 92]. Cette approche laisse une grande liberté à l'interprète, mais peut perdre le suivi lors de passages à nombre variable de notes comme dans les trilles ou les notes répétées. A l'opposé, d'autres approches sont fondées essentiellement sur les figures rythmiques et le respect du tempo [VAN 95]. Le suivi simultané des notes jouées et du tempo autorise la prédiction des instants où seront jouées les prochaines notes ; la mesure de l'écart entre la prédiction et la note effective permet de différencier s'il s'agit d'une note de l'accord en cours ou d'une nouvelle note et de recalculer le tempo si nécessaire.

Les difficultés proviennent des erreurs de l'interprète (fausses notes, notes ou groupe de notes pas jouées, notes corrigées après coup), mais aussi des erreurs dues au système de suivi de notes (fausses détections, omissions). La plupart des systèmes prévoient donc des mécanismes de rattrapage pour ces cas d'erreurs. L'utilisation répétée de systèmes de suivi en situation de concert a poussé certains auteurs à développer des mécanismes d'« apprentissage » de certaines caractéristiques de l'interprétation du musicien [VER 85] ou de l'instrument lui-même en contexte musical [KAS 99].

Plus récemment, des systèmes plus performants mais aussi plus complexes de suivi de partitions ont été développés notamment à l'IRCAM [LEM 03, ORI 01a, ORI 01b]. Ils sont construits sur des modèles de Markov cachés à deux niveaux, un niveau représentant la succession des notes de la partition et un niveau représentant l'évolution interne de chaque note. Ces systèmes présentent les avantages de ne plus nécessiter l'estimation directe de paramètres peu robustes (comme le F_0) et d'offrir des possibilités d'apprentissage de l'interprétation musicale.

1.8. Bibliographie

[ALL 77] ALLEN J.B., « Short-term spectral analysis, synthesis and modification by discrete Fourier transform », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, p. 235-238, 1977.

[BAI 93] BAIRD B., BLEVINS D., ZAHLER N., « Artificial intelligence and music: Implementing an interactive computer performer », *Computer Music Journal*, vol. 17, n° 2, p. 73-79, 1993.

[BLA 98] BLANCHET G., CHARBIT M., *Traitement numérique du signal : simulation sous Matlab*, Hermès, Paris, 1998.

[BOB 98] BOBREK M., KOCH D.B., « Music signal segmentation using tree-structured filter banks », *Journal of the Audio Engineering Society*, vol. 46, n° 5, p. 412-427, 1998.

[BRO 89] BROWN J.C., PUCKETTE M.S., « Calculation of a "narrowed" autocorrelation function », *JASA*, vol. 85, p. 1595-1601, avril 1989.

[CAP 97] CAPPÉ O., OUDOT M., MOULINES E., « Spectral envelope estimation using a penalized likelihood criterion », dans *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, octobre 1997.

[CHA 85] CHAFE C., JAFFE D., KASHIMA K., MONT-REYNAUD B., SMITH III J.O., « Techniques for note identification in polyphonic music », dans *International Computer Music Conference*, 1985.

[CHE 91] DE CHEVEIGNÉ A., « A mixed speech F0 estimation algorithm », *Eurospeech*, vol. 2, p. 445-448, 1991.

[CHE 99] DE CHEVEIGNÉ A., KAWAHARA H., « Multiple period estimation and pitch perception model », *Speech Communication*, vol. 27, n° 3-4, p. 175-185, 1999.

[COO 96] COOPER D., KIA C.N., « A monophonic pitch-tracking algorithm based on waveform periodicity determinations using landmark points », *Computer Music Journal*, vol. 20, n° 3, p. 70-78, 1996.

[DAL 98] D'ALESSANDRO C., DARSINOS V., YEGNANARAYANA B., « Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources », *IEEE Transactions on Speech and Audio Processing*, vol. 6, n° 1, p. 12-23, janvier 1998.

[DAN 84] DANNENBERG R., « An on-line algorithm for real-time accompaniment », dans *International Computer Music Conference*, San Francisco, Californie, p. 193-198, 1984.

[DEP 93] DEPALLE P., GARCÍA G., RODET X., « Tracking of partials for additive sound synthesis using hidden Markov models », *ICASSP*, vol. 1, p. 225-228, avril 1993.

[DEP 94] DEPALLE P., GARCÍA G., RODET X., « A virtual castrato (!?) », dans *International Computer Music Conference*, Aarhus, Danemark, 1994.

[DOV 94] DOVAL B., Estimation de la fréquence fondamentale des signaux sonores, Thèse de doctorat, Université Paris VI, 1994.

[DUI 82] DUIFHUIS H., WILLEMS L.F., SLYUTER R.J., « Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception », *JASA*, vol. 71, n° 6, p. 1568-1580, juin 1982.

[FIT 96] FITZ K., HAKEN L., « Sinusoidal modeling and manipulation using Lemur », *Computer Music Journal*, vol. 20, n° 4, p. 44-59, 1996.

[GEO 96] GEOFFROIS E., « The multi-lag-window method for robust extended-range F0 determination », dans *International Conference on Speech and Language Processing*, 1996.

[GOL 69] GOLD B., RABINER L.R., « Parallel processing techniques for estimating pitch periods of speech in the time-domain », *JASA*, vol. 46, p. 442-448, 1969.

[GOT 99] GOTO M., MURAOKA Y., « Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions », *Speech Communication*, vol. 27, n° 3-4, p. 311-335, 1999.

[HAR 78] HARRIS F.J., « On the use of windows for harmonic analysis with the discrete Fourier transform », *Proceedings of the IEEE*, vol. 66, n° 1, p. 51-83, 1978.

[HER 88] HERMES D., « Measurement of pitch by subharmonic summation », *JASA*, vol. 83, n° 1, p. 257-264, janvier 1988.

[HES 83] HESS W., *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin, 1983.

[KAS 99] KASHINO K., MURASE H., « A sound source identification system for ensemble music based on template adaptation and music stream extraction », *Speech Communication*, vol. 27, n° 3-4, p. 337-349, 1999.

[KUH 90] KUHN W.B., « A real-time pitch recognition algorithm for music applications », *Computer Music Journal*, vol. 14, n° 3, p. 60-71, 1990.

[LAN 89] LANSKY P., *Compositional applications of linear predictive coding*, dans Mathews M., Pierce J. (dir.), *Current Directions in Computer Music Research*, MIT Press, Cambridge, Massachusetts, 1989.

[LEM 03] LEMOUTON S., SCHWARZ D., ORIO N., « Score following: State of the art and beyond », dans *Conference on New Instruments for Musical Expression (NIME'03)*, 2003.

[MAH 89] MAHER R.C., An approach for the separation of voices in composite musical signals, Thèse de doctorat, Université de l'Illinois, Urbana-Champain, 1989.

[MAH 90] MAHER R.C., « Evaluation of a method for separating digitized duet signals », *JAES*, vol. 38, n° 12, p. 956-979, décembre 1990.

[MAR 76] MARKEL J.D., GRAY A.H., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[MART 81] MARTIN P., « Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne », dans *Douzièmes journées d'études sur la parole*, p. 223-232, mai 1981.

[MART 82] MARTIN P., « Comparison of pitch detection by cepstrum and spectral comb analysis », *ICASSP*, vol. 1, p. 180-183, 1982.

[MCA 86] MCAULAY R.J., QUATIERI T.F., « Speech analysis/synthesis based on a sinusoidal representation », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, p. 744-754, 1986.

[MED 91] MEDAN Y., YAIR E., DAN C., « Super resolution pitch determination of speech signals », *IEEE ASSP*, vol. 39, n° 1, p. 40-48, janvier 1991.

[NOL 67] NOLL A.M., « Cepstrum pitch determination », *JASA*, vol. 41, n° 2, p. 293-309, 1967.

[ORI 01a] ORIO N., DÉCHELLE F., « Score following using spectral analysis and hidden Markov models », dans *ICMC*, ICMA, La Havane, Cuba, 2001.

[ORI 01b] ORIO N., SCHWARZ D., « Alignment of monophonic and polypophonic music to a score », dans *ICMC*, La Havane, Cuba, 2001.

[PAL 83] PALIWAL K.K., RAO P.V.S., « A synthesis-based method for pitch extraction », *Speech Communications*, vol. 2, 1983.

[PAP 84] PAPOULIS A., *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1984.

[PIC 98] PICINBONO B., *Signaux et systèmes linéaires*, Ellipses, Paris, 1998.

[POR 80] PORTNOFF M.R., « Discrete time signals and systems based on short-time Fourier analysis », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n° 1, p. 55-64, février 1980.

[PUC 92] PUCKETTE M., LIPPE C., « Score following in practice », dans *International Computer Music Conference*, San Francisco, Californie, p. 182-185, 1992.

[RAB 74] RABINER L.R., SCHAFER R.W., « On the behavior of minimax FIR digital Hilbert transformers », *The Bell System Technical Journal*, vol. 53, n° 2, février 1974.

[RAB 75] RABINER L.R., GOLD B., *Theory and application of digital signal processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[RAB 86] RABINER L.R., JUANG B.H., « An introduction to hidden Markov models », *IEEE ASSP Magazine*, p. 4-16, janvier 1986.

[RAP 99] RAPHAEL C., « Automatic segmentation of acoustic musical signals using hidden Markov models », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, n° 4, p. 360-370, 1999.

[SER 89] SERRA X., A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition, Thèse de doctorat, CCRMA, Université de Standford, 1989.

[SER 90] SERRA X., SMITH III J.O., « Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition », *Computer Music Journal*, vol. 14, n° 4, p. 12-24, 1990.

[SIE 98] SIEGER N.J., TEWFIK A.H., « Audio coding for representation in MIDI via pitch detection using harmonic dictionaries », *Journal of VLSI Signal Processing Systems for Signal Image and Video Technology*, vol. 20, n° 1-2, p. 45-59, 1998.

[VAN 95] VANTOMME J.D., « Score following by temporal pattern », *Computer Music Journal*, vol. 19, n° 3, p. 50-59, 1995.

[VER 84] VERCOE B., « The synthetic performer in the context of live performance », dans *International Computer Music Conference*, San Francisco, Californie, p. 199-200, 1984.

[VER 85] VERCOE B., PUCKETTE M., « Synthetic rehearsal: Training the synthetic performer », dans *International Computer Music Conference*, San Francisco, Californie, p. 275-278, 1985.

[YEG 98] YEGNANARAYANA B., D'ALESSANDRO C., DARSINOS V., « An iterative algorithm for decomposition of speech signals into periodic and aperiodic components », *IEEE Transactions on Speech and Audio Processing*, vol. 6, n° 1, p. 1-11, janvier 1998.