Effets et transformations sonores

support de cours

Technologie Multimédia

ENSEIRB

Vincent Verfaille
CNRS – LMA
31, chemin Joseph Aiguier
13402 Marseille Cedex 20
verfaille@lma.cnrs-mrs.fr

Janvier 2003

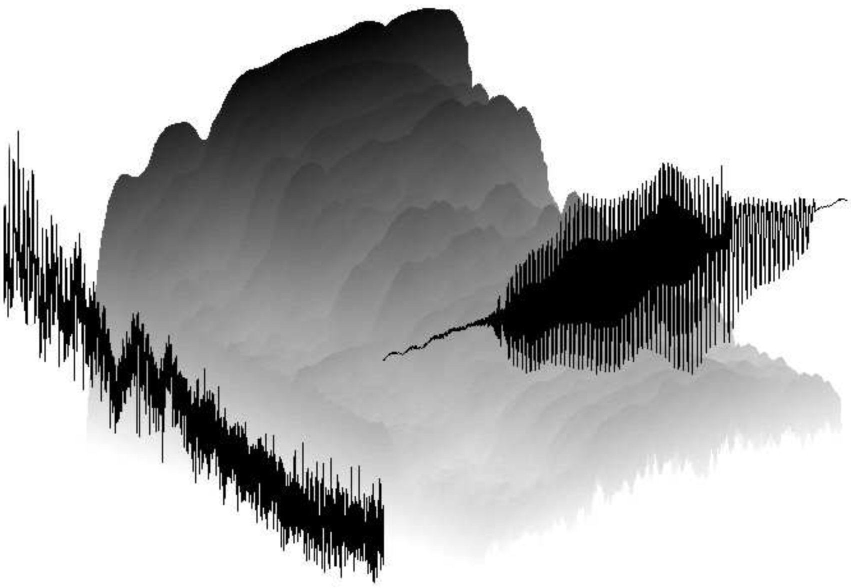

Table des matières

Ce document est le support de cours la partie "**traitements sonores**" de l'UV3 de Technique Multimédia enseignée à l'ENSEIRB en janvier 2003. Il porte sur les effets et transformations sonores appliquées sur le signal numérique. Le document comporte quatre parties principales :

- **chapitre 1** : des rappels et des bases portant sur le filtrage, les lignes à retard, les modulateurs et démodulateurs, ainsi que les méthodes d'analyse–transformation–synthèse temporelles (mémoire circulaire et TD-PSOLA), temps-fréquence (vocodeur de phase) et fréquentielles (somme de sinusoïdes + bruit + résidu) ;
- **chapitre 2** : une liste non exhaustive d'effets et transformations, classés selon la taxonomie perceptive, à savoir portant sur la dynamique, la hauteur, l'échelle temporelle, la spatialisation, le timbre ;
- **chapitre 3** : des notions sur l'implémentation et le contrôle des effets ;
- **chapitre 4** : les énoncés des travaux pratiques, pour moitié réalisés avec **Matlab** et pour moitié dans l'environnement temps-réel **Pure Data**.

ainsi qu'une bibliographie sommaire (pour compléter certaines notions).

Pour les curieux, les insatisfaits, ceux dont la lecture de ce document laisserait légitimement sur leur faim, voici trois ouvrages complémentaires, ayant fortement inspiré la rédaction de ce document :

- Jean Laroche,
  Traitement des Signaux Audio-Fréquences,
  département TSI, Sup'Télécom Paris,
  support de cours,
  février 1995 ;
- Curtis Roads,
  L'audionumérique,
  éditions Dunod, ISBN 2-10-004136-3
  1998 ;
- Udo Zölzer,
  DAFx: Digital Audio Effects,
  éditions Wiley, ISBN 0-471-49078-4
  2002.

Ce dernier livre présente l'intérêt d'être très complet, de disposer d'un site web où la plupart des codes **Matlab)** sont téléchargeable pour tester les effets, à l'adresse suivante : `http://www.dafx.de/`

Je suis disponible par Mél. pour toute question, toute correction (de quelque ordre que ce soit) correspondant à des erreurs que vous pourriez relever, n'étant moi-même pas sûr de ne pas en avoir laisser s'immiscer.

Sur ce, bonne lecture !

Chapitre 1

Rappels, bases

Nous proposons tout d'abord un chapitre introductif donnant quelques bases, sous forme de rappels. Ces bases seront utilisées pour tout le chapitre 2, aussi il est fortement conseillé de s'y apesantir le temps nécessaire.

1.1 Traitements, effets, transformations

Commençons par définir les terme que nous utilisons.
Traitement : calcul appliqué à un signal, dans notre cas un signal numérique sonore. Les traitements sonores sont des calculs mathématiques, basées sur les opérations mathématiques algébriques : additions, multiplications.
Effet : i) perception d'une transformation caractéristique du son (ce que l'on perçoit et uniquement cela) ; ii) méthode de traitement en vue de la modification d'un signal (changement en surface, par ajout ou soustraction). Un effet, dans son acceptation "traitement de signal", en tant que traitement du son, est considéré comme un traitement de bas niveau ou de surface.
Transformation : dans le cas du son, une transformation consiste à appliquer un traitement qui modifie le son en profondeur (une connaissance de la structure est sous-jacente). Une transformation est considérée comme un traitement de haut niveau.

A la suite de ces définitions, il apparaît clair que si l'on veut parler aussi bien des effets que des transformations sonores, l'expression "traitements sonores" est la plus appropriée, ceci afin de se placer du côté de la personne qui conçoit et applique les traitements.

1.2 Notions d'harmoniques, d'enveloppe spectrale et de formants

1.2.1 Sinusoïde, partiel, bruit:indexbruit

On appelle **partiel** une sinusoïde, que l'on reconnaît dans le spectre à court-terme d'un signal par une raie ou un pic. Les **bruits** que l'on considère sont des signaux aléatoires. Les pics qu'ils peuvent présenter dans des paniers de fréquences d'une transformée de Fourier à court-terme ont ceci de différent avec les partiels que leur amplitude est souvent plus petite, les maxima locaux voisins sont plus nombreux et de plus grande amplitude, et d'une transformée à la suivante (dans le temps), ils apparaissent et disparaissent, alors que les partiels demeurent présent et évoluent plus doucement (en fréquence et en amplitude).

1.2.2 Les sons harmoniques

Les sons harmoniques correspondent à des sons constitués d'un ensemble de raies spectrales (ou partiels) tous multiples d'une fréquence fondamentale (cf. fig.1.1). Ce sont principalement des sons instrumentaux (voix parlée ou chantée, instruments à vent ou à cordes, cuivres). Les sons bruités, quant à eux, correspondent à des sons sans hauteur prédominante, ou avec plusieurs hauteurs audibles. Leur spectre ne présente pas du tout la même allure (cf. fig.1.1).

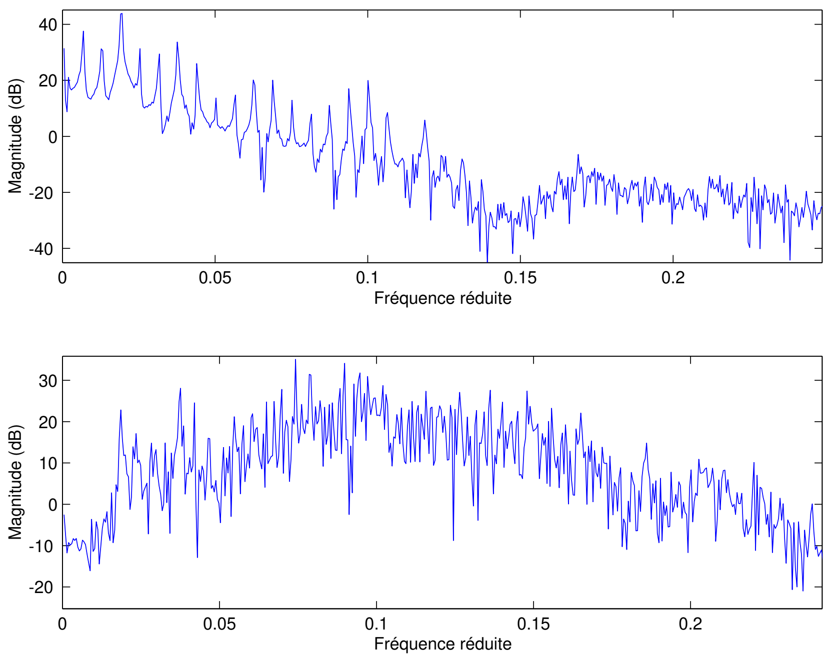

Fig. 1.1: *Transformée de Fourier à court-terme d'un son instrumental (fig. haut) et d'un bruit (fig. bas).*

Entre ces deux extrêmes se situent :

- les sons quasi-harmoniques : les partiels voient leur fréquence dévier petit à petit du multiple de la fréquence fondamentale, au fur et à mesure que l'on monte en fréquence et en numéro de partiel. C'est le cas de sons instrumentaux tels que ceux des instruments à cordes ou certains instruments de percussion (ex : marimba, xylophone).
- les sons inharmoniques : les partiels ne sont pas multiples d'une fondamentale ; cependant, on peut entendre une hauteur, voire plusieurs selon l'émergence de chaque partiel (ex : les sons de cloches, cf. fig.1.2).
- les sons filtrés présentant un phénomène d'émergence : il s'agit de bruits filtrés passe-bande étroit. Selon la structure du ou des filtres, une hauteur peut apparaître lors de leur audition (par exemple avec un filtre résonant).

On remarque que les sons harmoniques, quasi-harmoniques et inharmoniques sont constitués de partiels mais aussi de bruit, intrinsèquement lié aux partiels.

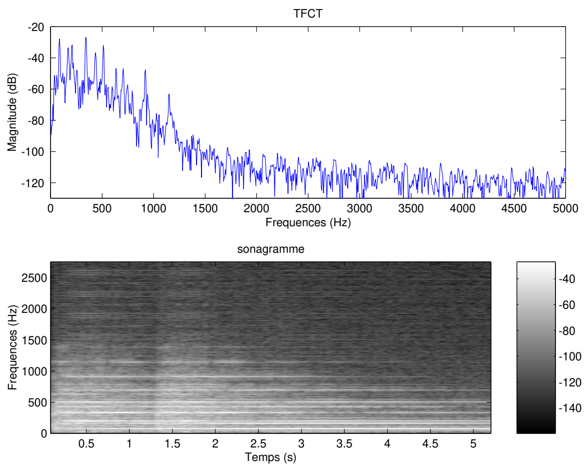

Fig. 1.2: *Transformée de Fourier à court-terme (fig. haut) et spectrogramme (fig. bas) d'un son de cloche.*

1.2.3 L'enveloppe spectrale et les formantsormant

Si l'on continue d'observer le spectre d'amplitude de signaux musicaux, cette fois-ci en regardant leur évolution dans le temps, nous remarquons que des formes plus générales que les partiels apparaissent. Il s'agit de l'enveloppe spectrale, définie comme l'enveloppe supérieure (une forme reliant les pics ou partiels), souvent approchée par un lissage du spectre. Cette enveloppe est formée de bosses appelées formants, et de creux. La figure 1.3 montre bien la coexistence de ces deux niveaux de structure que sont les formants (fig. haut) et les partiels (fig. bas).

1.2.4 Perception

Nous faisons ici quelques rappels succints sur la perception des sons (se reporter au cours de psycoacoustique pour plus de détails).

La **hauteur** perçue est la hauteur que l'on entend d'un son ; elle peut être distincte de la fréquence fondamentale du son. De même, des sons bruités peuvent présenter une hauteur ou plusieurs hauteurs simultanées.

La **sonie** est la perception d'intensité sonore. Elle correspond à un codage effectué par le système auditif, prenant en compte des informations temporelles (enveloppe) et fréquentielles (énergie dans des sous-bandes du spectre d'amplitude).

La **localisation** dans l'espace se fait, en azimuth, principalement à partir des différences de temps d'arrivée du signal acoustique aux deux oreilles et des différences de niveau de ces signaux. La localisation en élévation se fait à partir du filtrage procédé par le torse, le cou et la tête. La distance quant à elle est donnée en espace clos par la réverbération (son niveau par rapport au niveau du son direct), et en espace ouvert par le niveau du signal et son filtrage effectué par l'air.

Le **timbre** est codé principalement (pour les sons instrumentaux) par les trois dimensions suivantes :

1. la brillance, caractéristique de la distribution de l'énergie spectrale, et corrélée au centroïde (ou centre de gravité spectrale) [9, 12, 11] ;

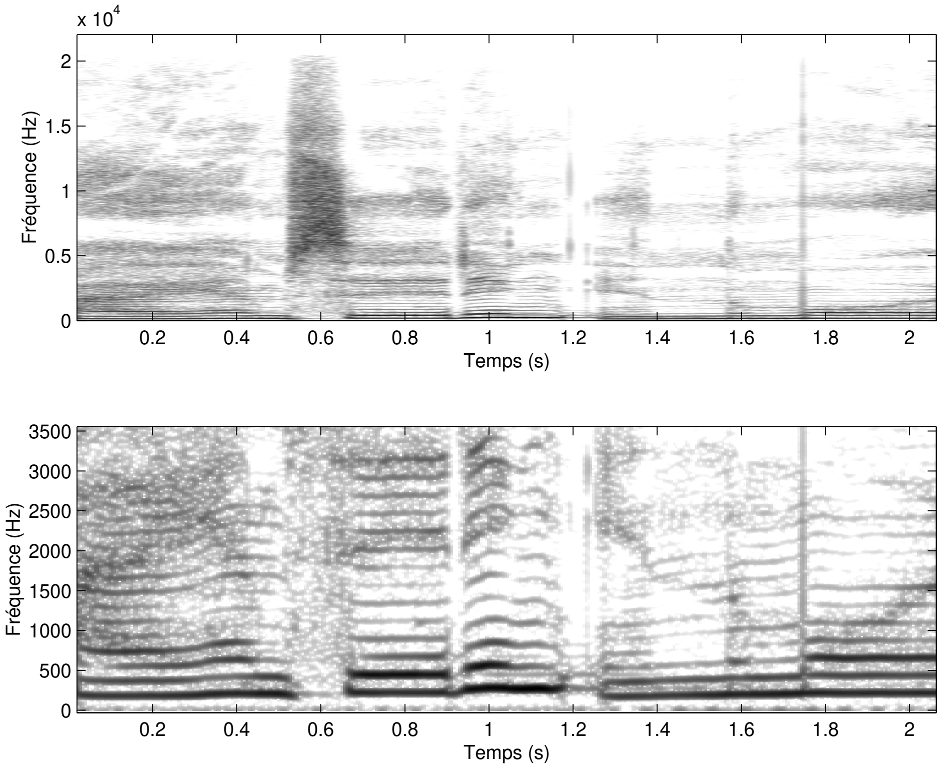

Fig. 1.3: *Sonagramme (spectrogramme d'amplitude) d'une voix chantée : ensemble du spectre (fig. haut) et basses fréquences (fig. bas).*

2. les propriétés spectrales fines, à savoir la cohérence des micro-variations des composantes spectrales, mesurées par le flux spectral et la synchronicité des attaques et décroissances des harmoniques [12] (la synchronicité en question est statique pour les instruments à vents et dynamique pour les cuivres) ;

3. le transitoire d'attaque qui confère le caractère d'explosivité ou non de l'attaque d'un son ; il est décrit par la synchronicité et l'harmonicité des partiels dans différents parties du spectre.

1.3 Filtrages

1.3.1 Filtre à réponse impulsionnelle finie (RIF)

Un filtre à réponse impulsionnelle finie est un filtre dont la réponse à un signal impulsionnel est de longueur finie. Un échantillon en sortie d'un tel filtre s'obtient par une somme pondérée des valeurs des N derniers échantillons entrants (cf. eq. 1.1). N est la longueur de la réponse impulsionnelle du filtre.

$$y(n) = a_0\ x(n) + \sum_{i=1}^{N} a_i\ x(n-i) \tag{1.1}$$

1.3.2 Filtre à réponse impulsionnelle infinie (RII)

Un filtre à réponse impulsionnelle infinie est un filtre dont la réponse à un signal impulsionnel peut être de longueur infinie. Un échantillon en sortie d'un tel filtre s'obtient par une somme pondérée des valeurs de l'échantillon entrant et des N échantillons sortants (cf. eq. 1.2). Un bouclage a donc lieu entre la sortie et l'entrée du filtre RII.

$$y(n) = a_0\ x(n) + \sum_{i=1}^{N} b_i\ y(n-i) \tag{1.2}$$

1.3.3 Filtre à étage

Un filtre à étage est représenté par l'un des graphiques de la figure 1.4. Il s'agit soit d'un filtre passe-bas (qui ne laisse passer que les fréquences en dessous de la fréquence de coupure f_c), soit d'un filtre passe-haut (qui ne laisse passer que les fréquences au-dessous de la fréquence de coupure).

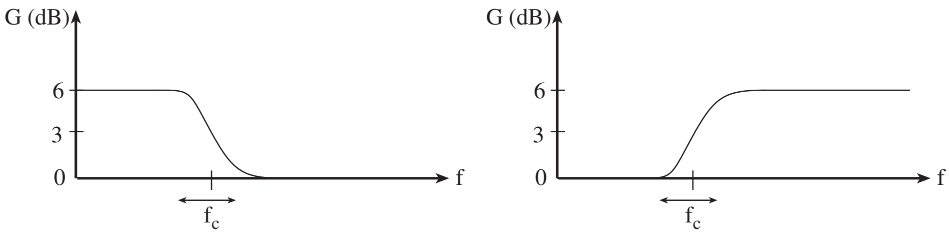

Fig. 1.4: *Filtre passe-bas (à gauche) et passe-haut (à droite).*

1.3.4 Filtre passe-tout

Filtre à étage du premier ordre

Un filtre passe-bas du premier ordre est décrit par sa fonction de transfert :

$$H(z) = 1 + \frac{H_0}{2}\ [1 + A(z)] \tag{1.3}$$

tandis qu'un filtre passe-haut du premier ordre est décrit par sa fonction de transfert :

$$H(z) = 1 + \frac{H_0}{2}\ [1 - A(z)] \tag{1.4}$$

avec le filtre passe-tout du premier ordre :

$$A(z) = \frac{z^{-1} + a}{1 + az^{-1}} \tag{1.5}$$

et a la fréquence de coupure. Sa valeur diffère selon que l'on veut augmenter, *booster* la bande de fréquence (a_B) ou l'atténuer (a_C), avec F_e la fréquence d'échantillonnage et f_c la fréquence de coupure du filtre :

$$a_B \quad = \quad \frac{\tan(\pi f_c/F_e) - 1}{\tan(\pi f_c/F_e) + 1} \tag{1.6}$$

$$a_C = \frac{\tan(\pi f_c/F_e) - V_0}{\tan(\pi f_c/F_e) + V_0} \quad (1.7)$$

Le facteur gain G en décibels peut être ajusté à l'aide du paramètre :

$$H_0 = V_0 - 1 \ \ avec \ \ V_0 = 10^{G/20} \quad (1.8)$$

La pente de ces filtres est de 6 dB par octave. Du fait que le filtre à étage du premier ordre possède un pôle et un zéro, sa réponse en fréquence possède un plateau (une tangente horizontale) au gain G désiré pour la fréquence nulle et un autre à 0 dB pour la fréquence maximale. La fréquence de coupure indique la zone où la réponse en fréquence chute vers 0.

Filtre à étage du second ordre

Les filtres à étage du second ordre sont à la fois passe-haut et passe-pas, du fait qu'ils contiennent deux pôles et deux zéros : ils sont définis par deux fréquences de coupure. Ils permettent notamment d'obtenir des filtres de pente plus élevée que celle des filtres du premier ordre.

1.3.5 Filtre en pic

Un filtre en pic (ou passe-bande) est donné par sa réponse en fréquence :

$$H(z) = 1 + \frac{H_0}{2}\left[1 - A_2(z)\right] \quad (1.9)$$

avec le filtre passe-tout d'ordre deux :

$$A_2(z) = \frac{-a + (d - d\,a)\ z^{-1} + z^{-2}}{1 + (d - d\,a)\ z^{-1} - a\ z^{-2}} \quad (1.10)$$

Le paramètre d relatif à la fréquence centrale du filtre et le coefficient H_0 sont donnés par :

$$d = -\cos(2\pi\ f_c/F_e) \quad (1.11)$$

$$V_0 = 10^{G/20} \quad (1.12)$$

$$H_0 = V_0 - 1 \quad (1.13)$$

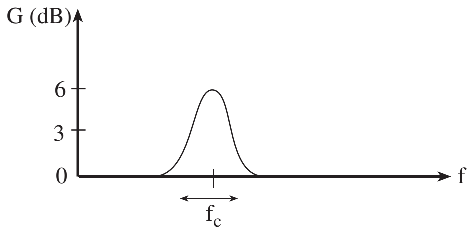

Fig. 1.5: *Filtre passe-bande.*

La largeur de bande est ajustée à l'aide des paramètres a_B pour augmenter le gain et a_C pour diminuer le gain :

$$a_B = \frac{\tan(\pi f_b/F_e) - 1}{\tan(\pi f_b/F_e) + 1} \quad (1.14)$$

$$a_C = \frac{\tan(\pi f_b/F_e) - V_0}{\tan(\pi f_b/F_e) + V_0} \quad (1.15)$$

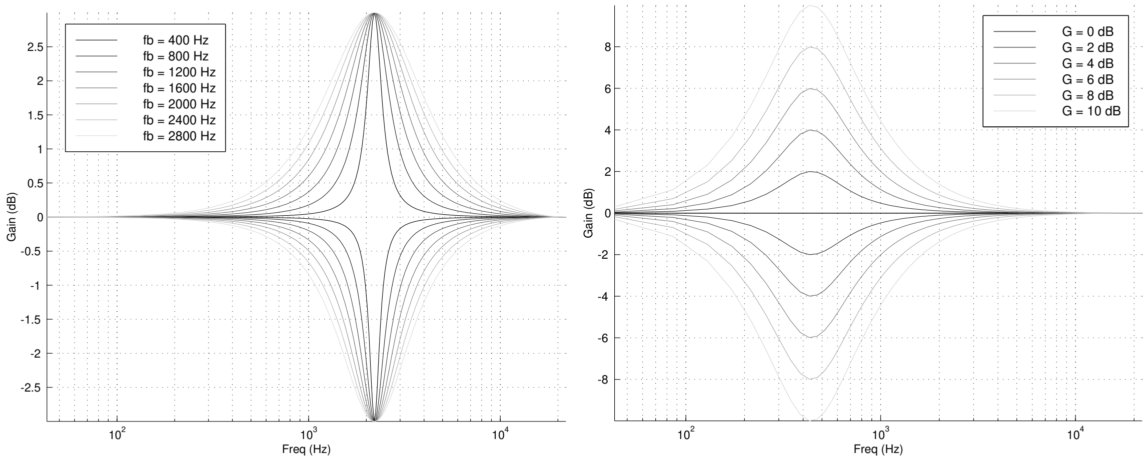

Fig. 1.6: *Filtre passe-bande pour différentes largeurs de bandes (différentes valeurs de Q) à gain constant (fig. gauche) et pour différents gains à Q constant (fig. droite).*

1.4 Lignes à retard

1.4.1 Principe

Un ligne à retard est une structure tampon dans laquelle les échantillons de sortie sont identiques aux échantillons qui y sont entrés, avec un temps de retard dépendant de la longueur de la ligne à retard (en nombre d'échantillons).

1.4.2 Filtrage à l'aide de lignes à retard

Filtre en peigne avec filtrage à réponse impulsionnelle finie (RIF)

Soit $x(n)$ un signal entrant dans une ligne à retard de longueur L. Si l'on ajoute le signal en entrée de la ligne à retard au signal en sortie, avec g_x le gain, on obtient le signal :

$$y(n) = x(n) + g_x\ x(n - L) \quad (1.16)$$

correspondant au filtrage en peigne à réponse impulsionnelle finie (RIF), de fonction de transfert :

$$H(z) = 1 + g_x\ z^{-L} \quad (1.17)$$

Pour une fréquence d'échantillonnage F_e, le délai temporel (ou temps de retard) induit par le passage à travers la ligne à retard est alors $\tau = \frac{L}{F_e}$.

Dans le cas où le gain g_x est positif, le filtre amplifie toutes les fréquences multiples de $\frac{1}{\tau} = \frac{F_e}{L}$ et atténue toutes les fréquences entre ces multiples. Le gain varie en amplitude entre

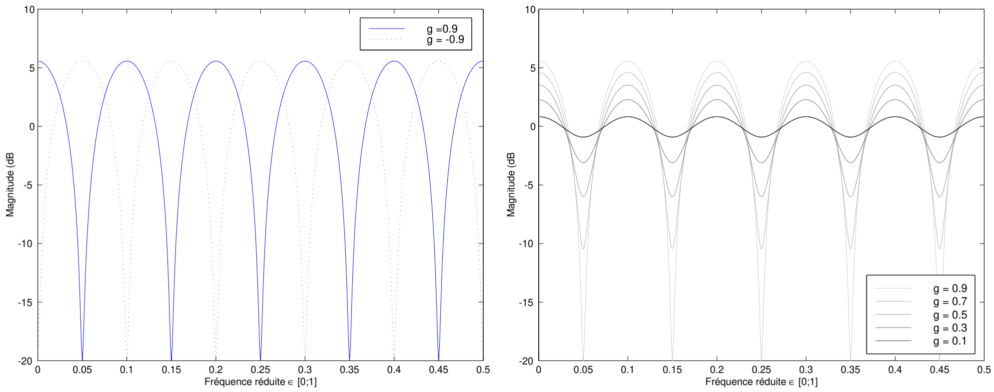

Fig. 1.7: *Réponse en fréquence d'un filtre en peigne FIR : à gauche pour $g_x = 0.9$ et $g_x = -0.9$, à droite pour $g_x = 0.1$, $g_x = 0.3$, $g_x = 0.5$, $g_x = 0.7$, $g_x = 0.9$.*

$1 - g_x$ et $1 + g_x$, et en décibels entre $20\log_{10}(1 - g_x)$ et $20\log_{10}(1 + g_x)$ dB. La réponse en fréquence de ce filtre ressemble à un peigne, les trous dans le spectre sont de plus en plus prononcés lorsque le gain $|g_x|$ augmente en valeur absolue (cf. fig. 1.7).

Selon le délai du filtre, le son produit sera perçu comme une coloration (transformation du timbre, pour des τ petits) ou comme un écho (τ grand).

Synthèse de sons (percussifs et entrenus)

La ligne à retard est l'équivalent numérique du guide d'onde physique parfait, ce qui explique son utilisation très fréquente pour la modélisation de sons instrumentaux. Les cordes vibrantes et les colonnes d'air sont des exemples de guides d'ondes. On retrouve ces idées de modélisation depuis Karplus-Strong ; l'utilisation qui en est faite aujourd'hui dépasse les simples guides d'ondes tubulaires, puisque les équipes de Rochesso en Italie et Smith aux Etats-Unis d'Amérique travaillent sur ces structures pour modéliser des plaques vibrantes (table d'harmonie de piano ou de guitare) par exemple.

1.4.3 Retard fractionnaire

Pour certains effets utilisant des lignes à retard, la précision temporelle donnée par l'échantillonnage n'est pas suffisante. Ainsi, pour des lignes à retard dont la longueur peut varier, avec une fréquence d'échantillonnage de $44,1$ kH, un temps de retard de 10 ms sera modélisé par une ligne à retard de $44,1$ échantillons, ce qui n'est pas réalisable avec des retards entiers. Dans ce cas, On utilise des lignes à retard plus grandes avant de procéder à une interpolation.

Soit $d \in \mathbb{R}$ le délai en nombre d'échantillons. Il peut se décomposer en une partie entier $M \in \mathbb{N}$ et une partie réelle $\gamma \in]0; 1[$. Le retard fractionnaire peut alors être modélisé comme la sortie d'une ligne à retard fractionnaire $y(n) = x(n - d) = x(n - (M + \gamma))$.

Pour calculer la valeur de $x(n - d)$, on procède à une interpolation à partir des valeurs voisines, au minimum à partir de $x(n - M)$ et $x(n - (M + 1))$. Différentes interpolations peuvent être utilisées :

- linéaire : $y(n) = \gamma.x(n - (M + 1)) + (1 - \gamma).x(n - M)$
- passe-tout : $y(n) = \gamma.x(n - (M + 1)) + (1 - \gamma).x(n - M) - (1 - \gamma).y(n - 1)$

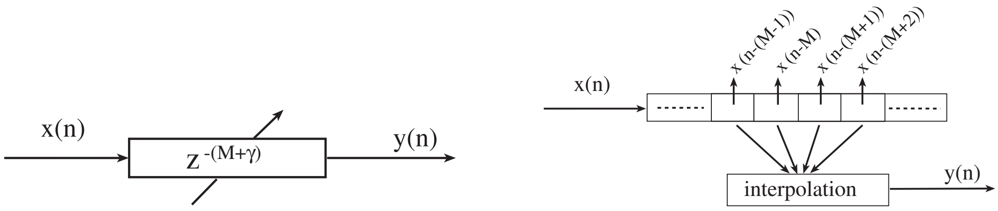

Fig. 1.8: *Diagrammes de la ligne à retard fractionnaire : à gauche, représentation "traitement du signal", à droite, détail de la méthodologie utilisée.*

- splines : $y(n) = \frac{\gamma^3}{6}.x(n-(M+1)) + \frac{(1+\gamma)^3-4\gamma^3}{6}.x(n-M) + \frac{(2-\gamma)^3-4(1-\gamma)^3}{6}.x(n-(M-1)) + \frac{(1-\gamma)^3}{6}.x(n-(M-2))$

1.5 Modulateurs

Pour tous les modulateurs, on considère un signal $x(n)$ dit porteuse, modulé d'une manière (amplitude, phase, fréquence) par une modulante $m(n)$. En télécommunications, la modulante est un signal composé d'une seule sinusoïdale. En tant qu'effet musical, la modulation peut se faire avec pour modulante une somme de sinusoïdes.

1.5.1 Modulateur en anneau

Le modulateur en anneau (*ring modulator*) est un multiplicateur de deux signaux, échantillon par échantillon. Les modulateurs en anneau analogiques étaient réalisés à l'aide d'un circuit de 4 diodes disposées en anneau, d'où le nom. Le diagramme correspondant à la modulation en anneau est donné en fig. 1.9.

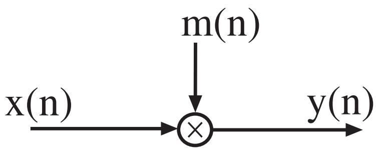

Fig. 1.9: *Diagramme du modulateur en anneau.*

Soient deux signaux sinusoïdaux x et m :

$$\begin{aligned} x(n) &= x_0 \sin(2\,\pi\,f_x\,n+\phi_x) && (1.18)\\ m(n) &= m_0 \sin(2\,\pi\,f_m\,n+\phi_m) && (1.19)\\ y(n) &= x(n).m(n) && (1.20)\\ &= x_0\,m_0 \sin(2\,\pi\,f_x\,n+\phi_x)\;.\;\sin(2\,\pi\,f_m\,n+\phi_m) && (1.21)\\ & && (1.22) \end{aligned}$$

En utilisant la formule trigonométrique usuelle :

$$\sin(a)\sin(b) = \frac{\cos(a-b)-\cos(a+b)}{2} \qquad (1.23)$$

on obtient :

$$y(n) = \frac{x_0\,m_0}{2}\;(-\cos[2\,\pi\,(f_m+f_x)\,n+(\phi_m+\phi_x)]$$

$$+\cos\left[2\,\pi\,\left(f_m - f_x\right)\,n + \left(\phi_m - \phi_x\right)\right]\big) \tag{1.24}$$

c'est-à-dire la somme de deux signaux sinusoïdaux de fréquences $f_m + f_x$ et $f_m - f_x$ (cf. ex. fig.1.10).

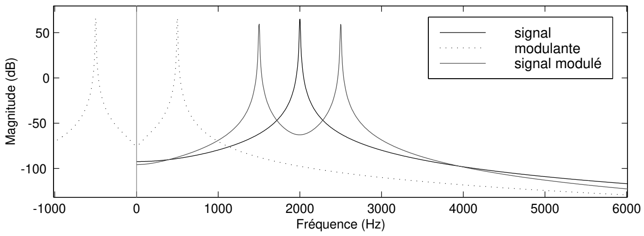

Fig. 1.10: *Modulateur en anneau, avec une fréquence de* 500 Hz *pour la modulante, appliqué à un signal contenant une sinusoïde à* 2000 Hz.

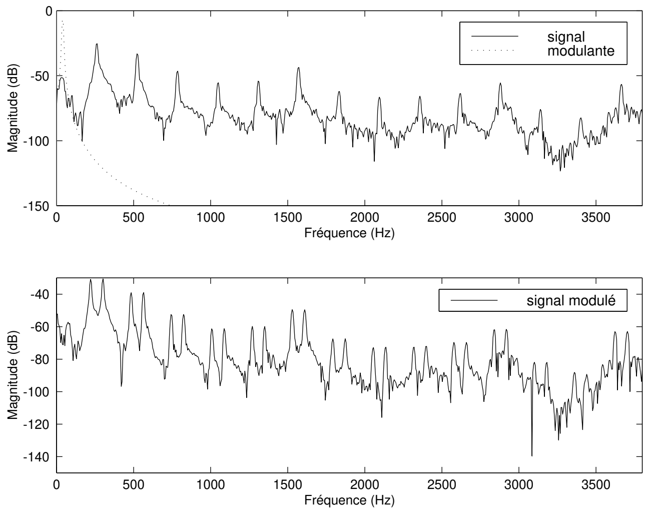

Fig. 1.11: *Modulateur en anneau, avec une fréquence de* 40 Hz *pour la modulante, appliqué à un signal réel harmonique de fondamentale* 200 Hz.

Dans le cas où le signal x est harmonique, la lecture du spectre devient moins aisée, comme en atteste la figure 1.11. Prenons maintenant le cas où les signaux x et m sont harmoniques :

$$x(n) = \sum_{k=1}^{H_x} x_k\,\sin(2\,\pi\,k\,f_x\,n + \phi_{x,k}) \tag{1.25}$$

$$m(n) = \sum_{l=1}^{H_m} m_l\,\sin(2\,\pi\,l\,f_m\,n + \phi_{m,l}) \tag{1.26}$$

$$y(n) = \sum_{k=1}^{H_x}\sum_{l=1}^{H_m} x_k\,m_l\,\sin(2\,\pi\,k\,f_x\,n + \phi_{x,k})\,.\,\sin(2\,\pi\,l\,f_m\,n + \phi_{m,l}) \tag{1.27}$$

$$= \sum_{k=1}^{H_x}\sum_{l=1}^{H_m} x_k\,m_l\,\big(-\cos\left[2\,\pi\,\left(l\,f_m + k\,f_x\right)\,n + \left(\phi_{m,l} + \phi_{x,k}\right)\right] \tag{1.28}$$

$$+\cos\left[2\,\pi\,\left(l\,f_p - k\,f_x\right)\,n + \left(\phi_{m,l} - \phi_{x,k}\right)\right]\big) \tag{1.29}$$

Si le rapport entre les fréquences f_m et f_x est une fraction entière, par exemple $f_m = \frac{M}{N}\ f_x$, alors le signal résultant de la modulation en anneau est harmonique de fréquence fondamentale $\frac{f_x}{N}$:

$$\begin{aligned} y(n) &= \sum_{k=1}^{H_x}\sum_{l=1}^{H_m} x_k\ m_l\ \left(-\cos\left[2\ \pi\ \left(l\frac{M}{N}+k\right)\ f_x\ n + (\phi_{m,l}+\phi_{x,k})\right]\right. && (1.30)\\ &+\ \cos\left.\left[2\ \pi\ \left(l\frac{M}{N}-k\right) f_x\ n + (\phi_{m,l}-\phi_{x,k})\right]\right) && (1.31)\end{aligned}$$

Les figures 1.12 illustrent ceci pour deux signaux harmoniques composés de 10 sinusoïdes d'amplitudes $1/k^2$, k étant le numéro de l'harmonique. Le signal à traiter a pour fondamentale $f_x = 400\ Hz$, la fondamentale de la modulante prenant quant à elle des valeurs $f_m = \frac{P}{4}\ f_x$, multiples du quart de f_x.

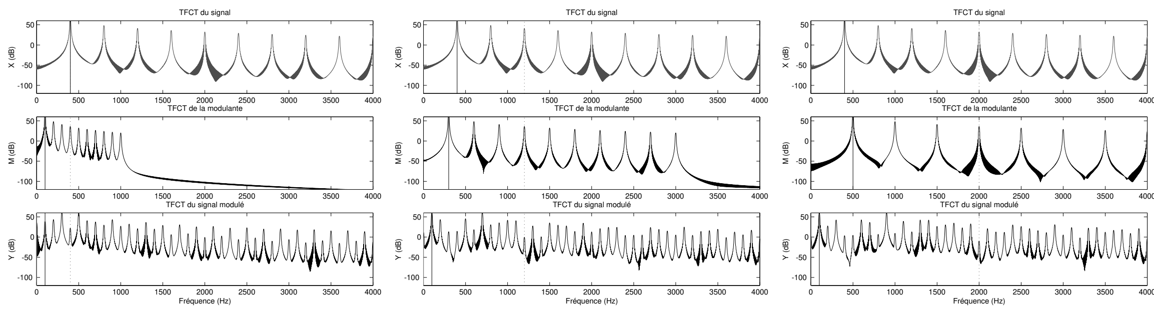

Fig. 1.12: *Modulation en anneau de deux signaux harmoniques de fréquences fondamentales* 400 Hz *et* $\frac{M}{N}$100 Hz *pour la modulante, soit un rapport* $\frac{M}{N} = 1/4$ *à gauche,* $\frac{M}{N} = 3/4$ *au centre et* $\frac{M}{N} = 5/4$ *à droite.*

La fréquence fondamentale des trois signaux obtenus dans l'exemple donné fig.1.12 est identique ; cependant, les spectres sont différents, du fait que les amplitudes des pics ne sont pas les mêmes. Il en résulte des timbres différents, pouvant aller jusqu'à des sons dont la fondamentale ne s'entend plus et dont un groupe d'harmoniques est prédominant, ce qui donne une hauteur perçue différente de la fondamentale. Aussi, selon l'effet que l'on veut rendre (choisir la fondamentale comme hauteur perçue, ou non), on choisira avec parcimonie le ratio $\frac{M}{N}$.

1.5.2 Modulateur d'amplitude

La modulation d'amplitude (*AM: amplitude modulation*) consiste à multiplier échantillon par échantillon un signal $x(n)$ (appelé porteuse) par un signal basse fréquence $m(n)$, appelé modulante, additionné au signal de départ.

$$y(n) = [1 + \alpha\ m(n)]\,.x(n) \tag{1.32}$$

Le signal de modulation $m(n)$ a pour amplitude maximale 1 ; α est le coefficient déterminant l'amplitude de la modulation.

Pour des fréquences de modulation en dessous de 20 Hz, la modulation sera entendue dans le domaine temporel (ex : trémolo), alors qu'au dessus de 20 Hz, elle sera entendue dans le domaine fréquentiel comme la somme de trois signaux : la porteuse $x(n)$, le signal composé

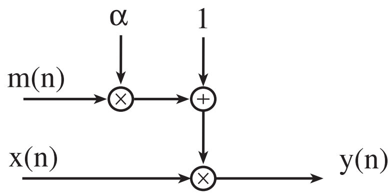

Fig. 1.13: *Diagramme du modulateur d'amplitude.*

des sommes et celui composé des différences de fréquences. Si par exemple la porteuse est une sinusoïde de fréquence f_x et la modulation une sinusoïde de fréquence f_m, le signal résultant de la modulation sera la somme des sinusoïdes de fréquences f_x, $f_m - f_x$ et $f_m + f_x$. Cette différence de perception provient du temps d'intégration de l'information de notre système auditif.

1.5.3 Modulateur BLU ou à bande latérale unique *(SSB, Single Side Band)*

Dans le cas des signaux sonores (ie. réels), les fréquences positives et négatives du spectre transportent la même information, bien qu'organisée différemment. En effet, soit $x(n) \in \mathbb{R}$, avec $X(k)$ la transformée de Fourier à temps discret (définie en (1.56) avec le vocodeur de phase) pour $k = 0, \ldots, N-1$. On a l'égalité $X(k) = \overline{X}(N-k)$ (avec $\overline{X}$ le signe conjugué). L'idée utilisée pour le modulateur SSB était de ne transmettre que la moitié pertinente de l'information.

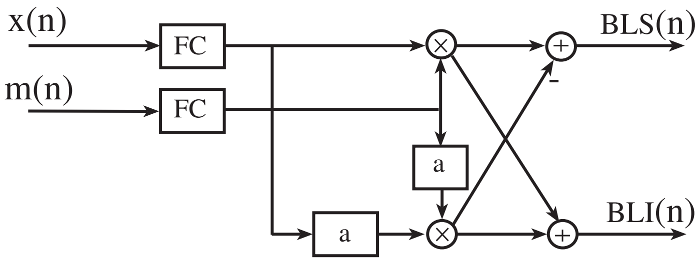

Fig. 1.14: *Diagramme du modulateur BLU.*

Tout d'abord, on déphase de 90° les phases du signal. Ceci se fait à l'aide de la Transformée de Hilbert à temps discret. Elle se modélise par un filtre RIF, de réponse impulsionnelle :

$$a(n) = \frac{1 - \cos(\pi\, n)}{2} \tag{1.33}$$

$$= \begin{cases} \frac{2}{\pi\, n} \text{ si n est impair} \\ 0 \text{ si n est pair} \end{cases} \tag{1.34}$$

$$\overline{a}(n) = a(n).w(n) \tag{1.35}$$

où w est une fenêtre de taille N, par exemple une fenêtre de Hamming. Si l'on note $\hat{x} = x * \overline{a}$ et $\hat{m} = m * \overline{a}$ les signaux dont la phase est décalée de 90°, on obtient l'expression analytique des signaux BLI (bande latérale inférieure, ou *LSB, Lower Side Band*) et BLS (bande latérale supérieure, ou *USB, Upper Side Band*) :

$$BLS(n) = x(n)\, m(n) - \hat{x}(n)\, \hat{m}(n) \tag{1.36}$$

$$BLI(n) \quad = \quad x(n)\ m(n) + \hat{x}(n)\ \hat{m}(n) \tag{1.37}$$

A partir de $N = 60$, les résultats sonores sont de qualité acceptable.

1.5.4 Modulateur de fréquence et de phase

La modulation de fréquence (MF, *FM, Frequency Modulation*) a connu de nombreuses applications dans la diffusion (télé-diffusion, radio-diffusion) ainsi que dans la synthèse sonore, grâce aux travaux de John Chowning dans les années 70 (et son brevet avec Yamaha sur le DX7, premier synthétiseur FM, au début des années 80).

Le principe est le suivant : le signal de modulation est une cosinus de fréquence porteuse f_p et de phase $\Phi(t)$ fonction d'une modulante $m(t)$. La formulation analytique est :

$$x_{MF/MP}(t) = A_p\ \cos\left(2\pi f_p t + \Phi(t)\right) \tag{1.38}$$

avec $\Phi(t) = k_{MP}\ m(t)$ pour la modulation de phase (la phase est directement proportionnelle à la modulante) et $\Phi(t) = 2\pi\ k_{MF} \int_{-\infty}^{t} m(\tau) d\tau$ pour la modulation de fréquence (la phase est proportionnelle à l'intégrale de la modulante).

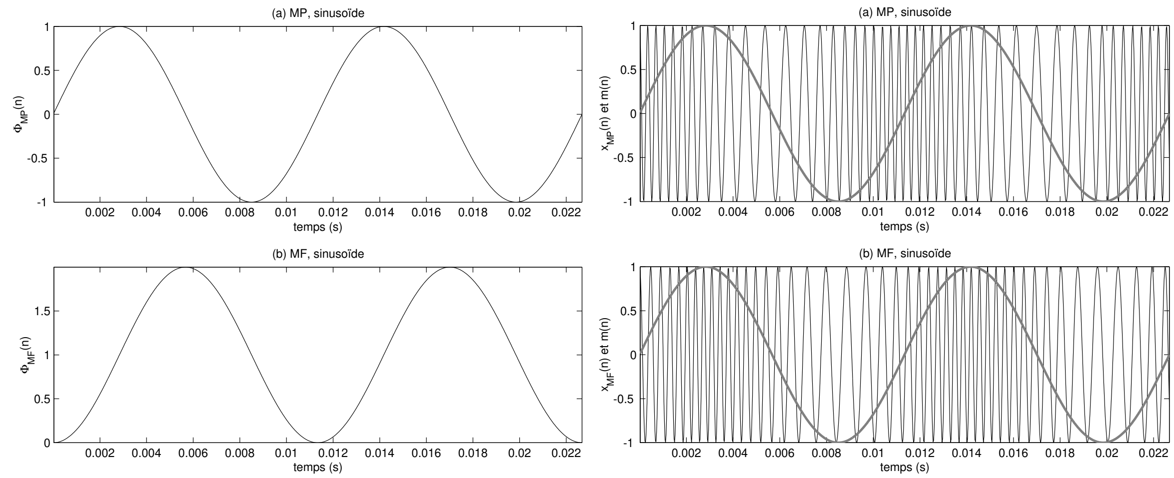

Fig. 1.15: *Résultats d'une modulation en fréquence et en phase pour une modulante sinusoïdale.*

Dans le cas des effets sonores, la modulation d'angle n'est plus tout à fait la même chose. L'idée est toujours d'utiliser une modulante $m(n)$ pour modifier la phase $\Phi(t)$ d'une sinusoïde porteuse, de fréquence f_p. Cette fois-ci, la représentation utilisée est un filtre à réponse impulsionnelle $h(n)$ variant dans le temps, définie par :

$$h(n) = \delta\left(n - m(n)\right) \tag{1.39}$$

Le signal résultant de ce filtrage est un signal à modulation de phase :

$$y(n) = x(n) * h(n) = x(n - m(n)) = x_{MP}(n) \tag{1.40}$$

La transformée de Fourier de ce signal est :

$$Y\left(e^{j\Omega}\right) = X_{MP}\left(e^{j\Omega}\right) = X\left(e^{j\Omega}\right)\ e^{-j\Omega m(n)} \tag{1.41}$$

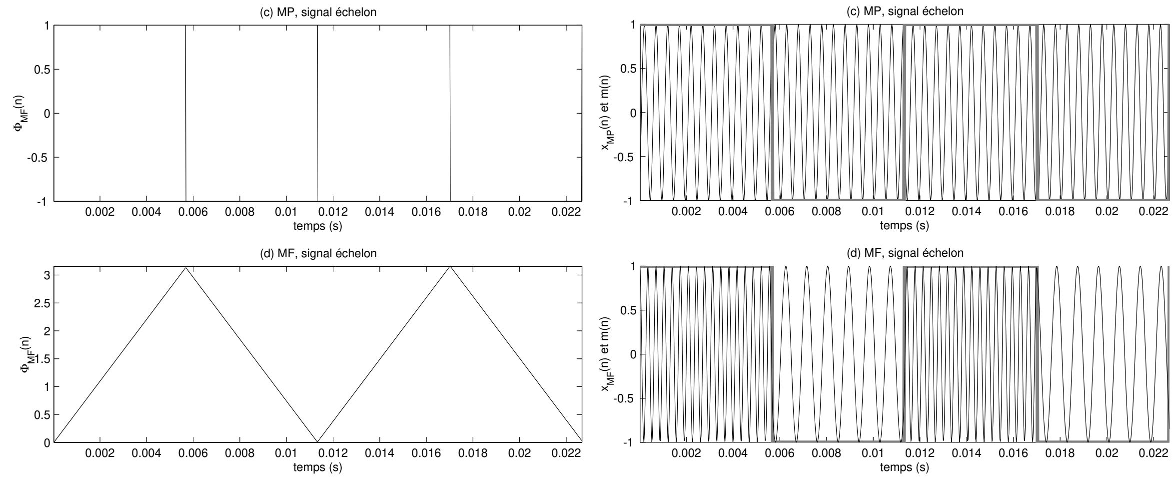

Fig. 1.16: *Résultat d'une modulation en fréquence et en phase pour une modulante "échelon".*

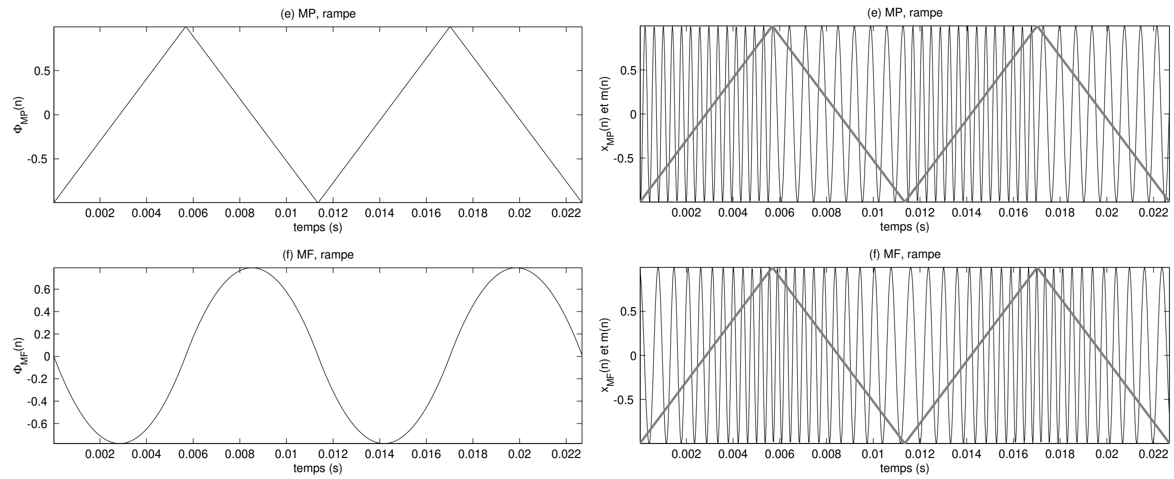

Fig. 1.17: *Résultat d'une modulation en fréquence et en phase pour une modulante "rampe".*

On voit bien apparaître dans le dernier terme la modulation sur la phase. La variable $m(n)$ est continue, et peut donc se décomposer en une partie entière M et une partie fractionnaire γ (cf. les délais fractionnaires 1.4.3).

Le fait que le temps de délai varie implique que la partie de signal dupliquée est lue à une distance plus où moins grande, donc à un taux d'échantillonnage différent de celui d'écriture ; il en résulte une différence de hauteur perçue : c'est l'effet **Munchkin**.

Modulante sinusoïdale Dans le cas de modulante sinusoïdale (utilisée pour réaliser des vibratos), le signal de modulation s'écrit :

$$m(n) = M + d\ \sin(\omega_M nT) \tag{1.42}$$

avec d la profondeur de la modulation, ω_M la fréquence de la modulation (en radians). Pour un signal d'entrée sinusoïdal, le facteur de ré-échantillonnage s'obtient ainsi :

$$\alpha(n) = \frac{\omega_I}{\omega} = 1 - \gamma\ \omega_M\ T\ \cos(\omega_M\ n\ T) \tag{1.43}$$

avec ω_I la fréquence instantanée du signal sortant (en radians), ω la fréquence de la sinusoïde entrant (en radians). La valeur moyenne de $\alpha(n)$ est $< \alpha(n) >= 1$, ce qui signifie qu'en moyenne, le signal de sortie sera de même longueur, avec une variation de hauteur autour de la fréquence de la sinusoïde entrant (**vibrato**).

Modulante rampe Dans le cas de modulante de type "rampe", le signal de modulation s'écrit :

$$m(n) = M \pm p\, n \tag{1.44}$$

avec p la pente. Le facteur de ré-échantillonnage se dérive ainsi :

$$\alpha(n) = 1 \mp p \tag{1.45}$$

d'où un signal en sortie de longueur $L_{sortie} = \frac{L_{entree}}{\alpha}$, correspondant à la **transposition** d'un facteur α (ie. $F_{sortie} = \alpha F_{entree}$).

1.6 Démodulateurs

A chaque méthode de modulation est associée une méthode de démodulation. Un démodulateur est constitué d'un détecteur, d'un moyenneur et d'un élément de mise à l'échelle (*scaler*), cf. fig.1.18. Ce dernier est très souvent partie intégrante du matériel recevant le signal démodulé.

Fig. 1.18: *Diagramme du démodulateur.*

Les détecteurs et moyenneurs sont quant à eux utiles pour différents effets ou méthodes d'analyse-synthèse, aussi nous allons détailler un peu plus ce dont il s'agit.

1.6.1 Détecteurs

Soit $x(n)$ un signal et $\hat{x}(n)$ sa Transformée de Hilbert à temps discret. Quatre principaux détecteurs sont utilisés :

$$\begin{aligned} d_h(n) &= max\,(0, x(n)) && (1.46)\\ d_f(n) &= |x(n)| && (1.47)\\ d_r(n) &= x^2(n) && (1.48)\\ d_i^2(n) &= x^2(n) + \hat{x}^2(n) && (1.49)\\ & && (1.50) \end{aligned}$$

Le détecteur $d_h(n)$ est appelé rectificateur demi-forme d'onde (*half-wave rectifier*), $d_f(n)$ est le rectificateur (*full-wave rectifier*). Le détecteur $d_r(n)$ correspond au carré (*squarer*), et est notamment utilisé pour le calcul du RMS. Enfin, $d_i^2(n)$ est le détecteur d'enveloppe instantanée.

1.6.2 Moyenneurs

Un moyenneur est un filtre du premier ordre. A l'origine, il s'agissait d'un réseau Résistance-Capacité en circuit analogique.

Soit $g = \exp(-1/f_e\tau)$ le gain et $d(n)$ la sortie d'un détecteur. Le moyenneur le plus simple est donné par :

$$y(n) = (1-g)\ d(n) + g\ y(n-1) \tag{1.51}$$

f_e correspond à la fréquence d'échantillonnage, et τ au temps de montée du moyenneur. Dans le cas général, ce moyenneur n'est pas suffisant, du fait qu'aucune distinction ne soit faite entre la montée et la descente en énergie du signal. Aussi, un second moyenneur, appelé moyenneur AR (auto-régressif) à deux constantes de temps a été développé. Pour la montée ou l'attaque, on utilise $g = \exp(-1/(f_e\tau_a))$, avec $\tau_a \approx 5\ ms$. Pour la descente ou relaxation, on utilise $g = \exp(-1/(f_e\tau_r))$, avec $\tau_r >> \tau_a$.

$$\begin{aligned} y_{ar}(n) &= (1-g)\ d(n) + g\ y_{ar}(n-1) && (1.52)\\ g &= g_a \text{ si } y_{ar}(n-1) < d(n) && (1.53)\\ &= g_r \text{ sinon} && (1.54) \end{aligned}$$

1.6.3 Applications

Diverses applications utilisent ces moyenneurs et détecteurs, à savoir :

application	**détecteur**	**moyenneur**
détecteur AM	d_h	y
VU-mètre	d_h	
Peak-meter	d_f	y_{ar} avec $\tau_a = 10\ ms$, $\tau_r = 1000\ ms$
RMS	d_r	y
vocodeur, *live electronics*	d_f	y_{ar}
suivi d'enveloppe (HR)	d_i^2	y ou y_{ar}

1.7 Systèmes d'analyse – modification – synthèse

Contrairement aux traitements utilisant les techniques présentées jusqu'ici (filtrage, modulation de lignes à retard, traitements temporels) qui ne font aucune hypothèse sur le signal (ie. sur sa nature) et peuvent être utilisés pour tout type de son, les systèmes d'analyse – transformation – synthèse sont fondés sur des modèles de signal. L'hypothèse faite sur le signal permet d'aboutir à des méthodes de modification mieux adaptées, et souvent plus performantes.

Quatre modèles sont présentés :

- les méthodes segment-temporelles, ne faisant pas d'hypothèse sur le signal (méthode non paramétrique), utilisées uniquement pour des changements d'échelle temporelle ou fréquentielle ;
- le vocodeur de phase (méthode temps-fréquence), dont l'hypothèse est que le signal est d'énergie finie, à variations lentes et décomposable en grains de sons stationnaires (modèle non-paramétrique) ;

- le modèle additif (ou spectral), où le signal est modélisé comme une somme de sinusoïdes de modules et fréquences variant lentement dans le temps, plus une composante transitoire d'attaque et une composante bruitée résiduelle pour le modèle le plus raffiné (méthode paramétrique) ;
- le modèle soustractif (ou soruce-filtre), où le signal provient d'une source (ou excitation) riche en harmoniques, filtrée par un filtre, encore appelé résonance (méthode paramétrique).

Nous verrons au chapitre 2 les utilisations que l'on peut en faire en vue de transformations sonores.

1.7.1 Méthodes segment-temporelles

Nous présentons deux techniques ou méthodes segment-temporelles : celle de la mémoire circulaire, la plus ancienne (dérivant d'un système analogique des années 50), et celle de l'ajout-superposition temporel synchrone à la hauteur (TD-PSOLA).

Technique de la mémoire circulaire

La technique de la mémoire circulaire est la plus simple et la plus ancienne des techniques de modifications de l'échelle temporelle et fréquentielle. Il s'agit d'une méthode fonctionnant dans le domaine temporel.

L'origine analogique Cette technique dérive d'un système analogique proposé dans les années 50 par Fairbanks. Elle consiste à utiliser un magnétophone muni d'une tête rotative. La bande en boucle fermée s'enroule sur la moitié du cylindre (comme pour les magnétoscopes ou les DAT) et défile à vitesse constante. Le cylindre est muni de deux têtes de lectures diamétralement opposées dont les signaux sont mélangés avec un gain identique. Il est possible de contrôler le sens de la rotation et la vitesse du cylindre.

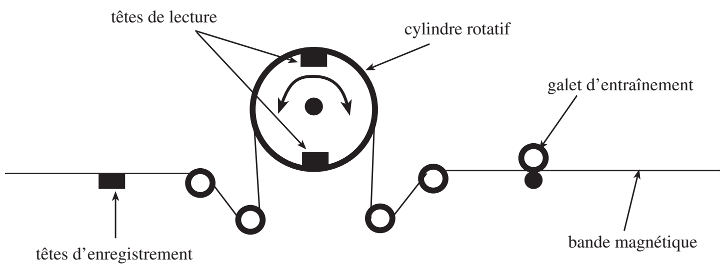

Fig. 1.19: *Système analogique avec tête rotative*

Lorsque le cylindre est immobile, la bande défile de façon identique devant la tête d'enregistrement et devant l'une des têtes de lecture. Le signal lu est donc identique au signal enregistré, aux erreurs d'enregistrement près.

Lorsque le cylindre tourne en sens inverse du défilement de la bande, la vitesse relative V_r de défilement de la bande par rapport à la tête de lecture est supérieure à sa vitesse de défilement absolue V_a. Pendant la durée du contact entre la tête de lecture et la bande, le signal est donc lu plus rapidement qu'il n'a été enregistré, ce qui correspond à une dilatation de l'axe des fréquences (transposition vers le haut). La présence des deux têtes assure la continuité grâce à un *cross-fade* naturel (lorsqu'une tête quitte la bande, l'autre s'en rapproche, de sorte que l'amplitude du signal total ne diminue pas). On remarque que certaines portions du signal peuvent être lues deux ou plusieurs fois, en fonction de la vitesse de rotation de la tête. C'est cette relecture qui permet de conserver la durée du signal original.

A l'inverse, lorsque le cylindre tourne dans le sens de défilement de la bande, le contenu en fréquence du signal est contracté vers l'origine (transposition vers le bas) puisque la bande est lue à une vitesse moindre qu'elle n'est enregistrée. Dans ce cas, certaines portions du signal peuvent ne pas être lues du tout.

Le rapport de l'homothétie en fréquences s'exprime par :

$$\alpha = \frac{V_r}{V_a} = 1 + \frac{R\Omega_{cyl}}{V_a} \tag{1.55}$$

avec V_a la vitesse de défilement de la bande devant la tête d'enregistrement, V_r la vitesse relative de la bande par rapport à la tête de lecture, Ω_{cyl} la vitesse de rotation du cylindre en $rad\ s^{-1}$ et R le rayon du cylindre. Dans tous les cas, l'alternance régulière des deux têtes se traduit par un "bruit" périodique de fréquence $f_{bruit} = \Omega_{cyl}/\pi$.

Les modifications de l'échelle temporelle du signal sont obtenues par exemple en enregistrant le signal une première fois sur la bande, puis en la rejouant avec une vitesse de défilement de bande multipliée par un facteur α. En l'absence de rotation de la tête de lecture, le hauteur du signal est bien sûr multipliée par le facteur α, ce que l'on cherche à éviter. On compense donc le changement de hauteur par une rotation appropriée de la tête de lecture.

Implémentation numérique La plupart des modificateurs de hauteur disponibles dans le commerce sont basés sur une réalisation numérique du système décrit ci-dessus. La bande magnétique est remplacée par une mémoire circulaire dans laquelle on place les échantillons du signal en entrée. Cette mémoire circulaire est lue par deux pointeurs diamétralement opposés.

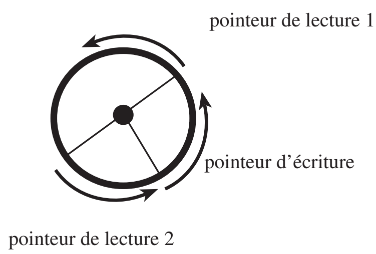

Fig. 1.20: *Système numérique avec mémoire circulaire*

La mémoire circulaire correspond à une ligne à retard. A chaque échantillon écrit dans la mémoire (toutes les Δt secondes), on avance les pointeurs de lecture de $\alpha\Delta t$ secondes, où α est le taux de modification, puis on lit un échantillon dans la mémoire. Pour des valeurs de α non entière, on se retrouve dans la problématique des lignes à retard fractionnaires 1.4.3.

Ainsi, le signal est lu avec une fréquence d'échantillonnage différente de celle à laquelle il a été écrit, ce qui provoque une modification de l'échelle des fréquences de taux α. Un problème se pose quand le pointeur de lecture rattrape (lorsque $\alpha > 1$) ou est rattrapé par (lorsque $\alpha < 1$) le pointeur d'écriture. Comme dans l'équivalent analogique, la continuité est assurée par un mixage des deux pointeurs au moment où se produit la rencontre (*cross-fade* : l'échantillon lu par le pointeur de lecture courant (par exemple le pointeur 1) subit une pondération décroissante tandis que celui lu par l'autre pointeur (le pointeur 2) subit une pondération croissante. Finalement, le second pointeur devient le pointeur courant et garde sa pondération maximale jusqu'à ce que le pointeur d'écriture s'en rapproche.

Implémenté de cette manière, le changement de hauteur par mémoire circulaire a un comportement sensiblement équivalent à celui de son homologue analogique (à ceci près qu'il est plus facilement paramétrable : taille de la mémoire ou de la tête rotative, fonction de pondération, etc). Son implémentation en temps-réel ne pose pas de problème particulier, puisqu'il réclame très peu de calculs. Il produit malheureusement un bruit artificiel qui provient du mixage périodique de deux pointeurs de lecture. Pour tenter d'améliorer la qualité obtenue, on cherche à mieux raccorder les signaux lus par les pointeurs de lecture, de façon similaire à ce qui va être présenté pour la méthode pitch-synchrone. On peut par exemple utiliser la fonction d'autocorrélation du signal pour déterminer l'endroit optimal pour le *cross-fading*.

Méthode temporelle "pitch-synchrone"

La méthode temporelle "pitch-synchrone" (ou *TD-PSOLA, Temporal Domain Pitch Synchronous OverLap Add*) suppose que l'on traite un signal de parole dont on connaît la période (déterminée par exemple par le premier pic de la fonction d'autocorrélation du signal). L'idée est fondée sur l'hypothèse que le signal de parole est constitué d'impulsions glottales filtrées par le conduit vocal. On observe une succession de réponses impulsionnelles, positionnées en des temps multiples de la période (hypothèse du peigne temporel convolué avec une réponse impulsionnelle de conduit vocal). On définit d'abord des marques d'analyse synchrones de la fréquence fondamentale pour les parties voisées, positionnées sur la forme d'onde à chaque période. Les modifications d'échelle temporelle et fréquentielle sont alors effectuées comme suit (nous préférons les exposer directement pour cette méthode, pour sensibiliser le lecteur à ces deux traitements majeurs et aux difficultés qui en découlent).

Modification de l'échelle temporelle Pour modifier la durée d'un signal sans en altérer la fréquence fondamentale, on va simplement dupliquer (étirement temporel) ou éliminer (compression temporelle) des périodes de la forme d'onde, en fonction du taux de modification désiré. On est donc conduit à définir des marques de synthèse également synchrones du fondamental, associées aux marques d'analyse (de façon non-bijective puisque certaines marques sont dupliquées ou éliminées).

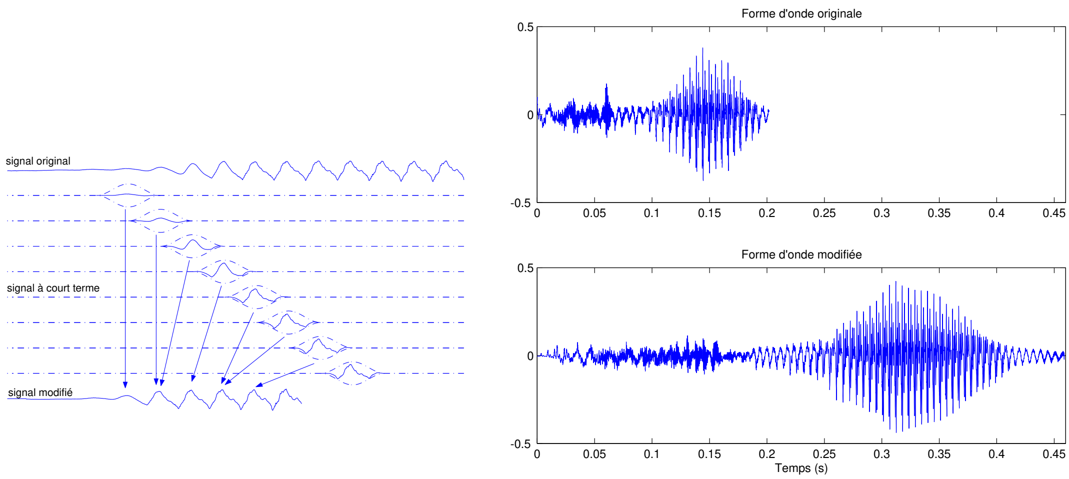

Fig. 1.21: *Changement d'échelle temporelle par TD-PSOLA, gauche, principe pour un facteur 2/3, droite pour un facteur 2, pour la syllabe "si".*

Les signaux à court-terme situés autour de chaque marque d'analyse sont alors extraits (par l'utilisation d'une fenêtre temporelle, par exemple de Hanning, de durée égale à deux périodes centrée sur la marque d'analyse) et recopiés autour des marques de synthèse correspondantes ; le signal modifié est obtenu par une simple méthode de superposition-ajout. La figure 1.21 gauche illustre le principe de cette méthode pour un taux d'étirement local de 2/3 (une contraction).

On voit que deux périodes du signal original ont donné naissance à une période du signal de synthèse sur la figure de gauche, ce qui correspond bien à un étirement temporel, mais la durée de la période n'est pas modifiée (l'écartement des marques de synthèse est le même que celui des marques d'analyse), la fréquence fondamentale du signal est conservée.

La figure 1.21 droite donne un exemple d'application à la syllabe “si”. on remarque la partie non-voisée à gauche de chaque forme d'onde (le son [s]), suivie de la partie voisée [i].

Modification de l'échelle fréquentielle Si l'on est capable de positionner dans le signal les marques d'analyse exactement sur le début de chaque onde glottale (réponse impulsionnelle du conduit vocal se produisant à chaque fermeture glottale), on conçoit que diminuer (respectivement augmenter) l'intervalle de temps séparant deux marques d'analyse consécutives va permettre d'augmenter (respectivement de diminuer) la fréquence du fondamental, sans que les formants soient modifiés (la réponse impulsionnelle n'est pas modifiée, en particulier sa décroissance temporelle et ses fréquences de résonances, ou formants).

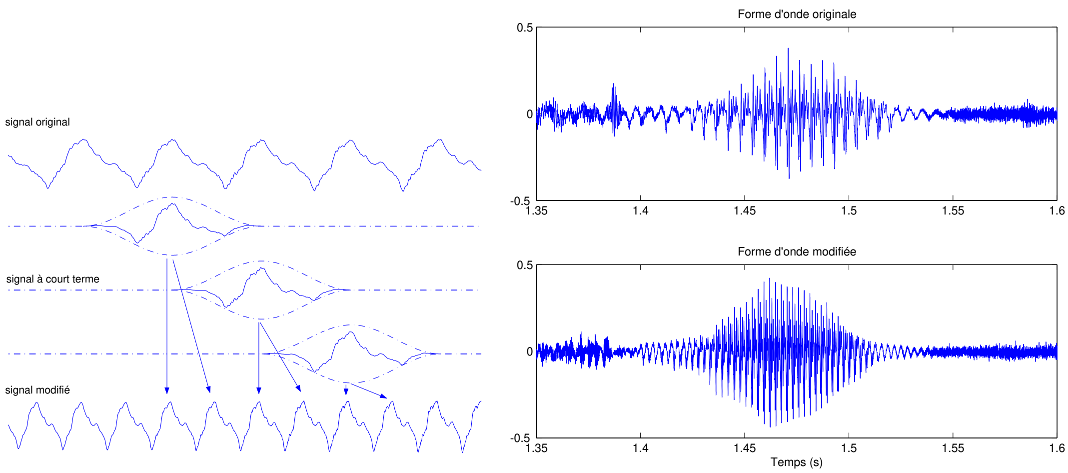

Fig. 1.22: *Changement d'échelle fréquentielle par TD-PSOLA pour un facteur 2 : principe (gauche) et exemple (droite)*

On est ainsi conduit à définir des marques de synthèse correspondant à la valeur modifiée du fondamental, et à les associer aux marques d'analyse comme précédemment. Puisque les marques de synthèse sont plus serrées (élévation du fondamental) ou écartées (abaissement du fondamental) que dans le signal original, il faut dupliquer ou éliminer certains marques si l'on veut conserver la durée du signal. La figure 1.22 illustre le principe de cette méthode. On constate que les marques de synthèse étant plus resserrées que les marques d'analyse, la période du signal est raccourcie. Pour éviter un raccourcissement du signal, il est nécessaire de dupliquer périodiquement certains signaux à court-terme.

Lorsque le signal ne possède plus de fréquence fondamentale bien précise (cas des consonnes), la modification est réalisée de façon non synchrone, jusqu'à ce que l'on retrouve une région où le fondamental est présent.

La méthode décrite ci-dessus est appliquée principalement à la parole, et réalise des modifications de très bonne qualité. Par sa simplicité, elle peut faire l'objet d'une implémentation temps-réel. En revanche, son application à des sons plus complexes (polyphoniques) ou dénués de hauteur (bruits) pose de sérieux problèmes. Les modifications du fondamental sont très sensibles à la position des marques d'analyse. Pour rendre la méthode plus robuste, les modifications de l'échelle fréquentielle peuvent être réalisés dans le domaine des fréquences (*FD-PSOLA, Frequential Domain PSOLA*).

1.7.2 Transformée de Fourier à Court-Terme et Vocodeur de Phase

Transformée de Fourier à Court-Terme

La Transformée de Fourier à Court-Terme (TFCT), ou représentation de Fourier à court-terme d'un signal discret $x(n)$, est définie par :

$$\begin{aligned} X(n,k) &= \sum_{m=-\infty}^{+\infty} x(m)\, h_a(n-m)\, W_N^{mk} \quad \text{avec } W_N = e^{-2j\pi/N} \quad \text{k=0,...,N-1} && (1.56)\\ &= |X(n,k)| e^{j\phi(n,k)} && (1.57)\\ &= X_R(n,k) + jX_I(n,k) && (1.58) \end{aligned}$$

C'est un nombre complexe représentant la magnitude et la phase d'un spectre variant dans le temps, avec $k = 0, \ldots, N-1$ l'index du panier de fréquence (ou *bin* de fréquence) et n l'index de temps. Pour chaque valeur de n, le signal $x(m)$ est pondéré par une fenêtre d'analyse de longueur finie M : $h_a(n-m)$.

Transformée de Fourier à Court-Terme par bloc

Si l'on écrit maintenant le même traitement effectué par bloc, on obtient :

$$\begin{aligned} X(sR_a,k) &= \sum_{m=-\infty}^{+\infty} x(m)\, h_a(sR_a-m)\, W_N^{mk} && (1.59)\\ &= W_N^{sR_a k} \sum_{m=-\infty}^{+\infty} x(m)\, h_a(sR_a-m)\, W_N^{-(sR_a-m)k} && (1.60)\\ &= W_N^{sR_a k} \tilde{X}(sR_a,k) && (1.61) \end{aligned}$$

La TFCT est échantillonnée tous les R_a échantillons en temps ; s est l'index de temps de la TFCT au taux d'échantillonnage de décimation. L'index de temps est $n = sR_a$ avec R_a le pas d'analyse ; $h_a(n)$ est la fenêtre d'analyse.

Il est important d'avoir l'origine des temps de la TF au centre de la fenêtre d'analyse à la fois à l'analyse et à la synthèse [15]. De cette manière, un signal impulsionnel centré sur la fenêtre temporelle aura ses phases à zéro. Ceci est obtenu par décalage circulaire d'une demi-fenêtre avant application de la TF. Le décalage circulaire d'une demi-fenêtre consiste à prendre la moitié de droite et à l'inverser avec la moitié de gauche. Il est donc identique dans les deux sens. Notons $\hat{x}(n) = x(n - N/2)$ le signal après décalage circulaire. Sa transformée de Fourier discrète est donc $\hat{X}\left(e^{j\Omega}\right) = e^{-j\Omega\frac{N}{2}} X\left(e^{j\Omega}\right)$. En exprimant la phase $\Omega_k = \frac{2\pi}{N}\, k$,

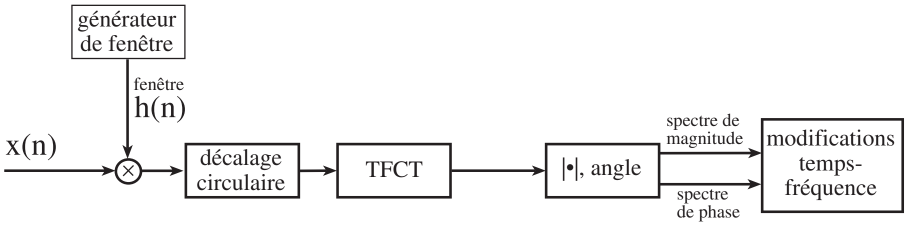

Fig. 1.23: *Diagramme de l'analyse par vocodeur de phase*

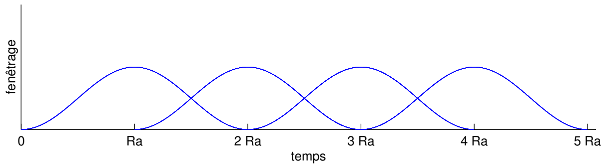

Fig. 1.24: *fenêtrage lors de l'analyse*

on obtient $\hat{X}(k) = e^{-j\frac{2\pi}{N}k\frac{N}{2}}X(k) = (-1)^k X(n)$. Il suffit donc de multiplier chaque valeur de la transformée de Fourier $X(k)$ par $(-1)^k$ pour effectuer le décalage circulaire et mettre les phases à zéros au centre du grain.

Après modification temps-fréquence (sur les modules et phases), la transformée de Fourier $X\left(sR_a, k\right)$ est modifiée en $Y\left(sR_s, k\right)$, avec R_s le pas de synthèse. Le signal temporel est obtenu par convolution de la tranformée de Fourier à court-terme inverse avec la fenêtre de synthèse :

$$y(n) = \sum_{s=-\infty}^{+\infty} y_s(n - sR_s)\, h_s(n - sR_s) \tag{1.62}$$

$$y_s(n - sR_s) = \frac{1}{N}\sum_{k=0}^{N-1}\left(W_N^{-sR_sk}\, Y\left(sR_s, k\right)\right)\, W_N^{-nk} \tag{1.63}$$

Tant que les modifications en module et en phase sont linéaires, le fenêtrage lors de la synthèse n'est pas nécessaire. Cependant, pour toute transformation non linéaire, le fenêtrage devient nécessaire, les grains de synthèse n'ayant plus nécessairement une enveloppe permettant la superposition sans problème de normalisation.

Déroulement de la phase

Lors du processus de modification, il est nécessaire de calculer le déroulement de la phase entre deux fenêtres de synthèse successives (et donc entre deux fenêtres d'analyse), de manière à bien interpoler les phases et calculer les fréquences instantanées de chaque panier de fréquence de la TFCT.

La phase déroulée $\tilde{\phi}(n, k)$ se calcule à partir de la phase $\phi(n, k)$ du grain d'analyse par la formule :

$$\tilde{\phi}(n, k) = \Omega_k n + \phi(n, k) = \frac{2\,\pi\,k}{N} n + \phi(n, k) \tag{1.64}$$

Nous avons besoin d'une fonction qui transforme toute phase en son argument principal dans l'intervalle $]-\pi; \pi]$, définie par :

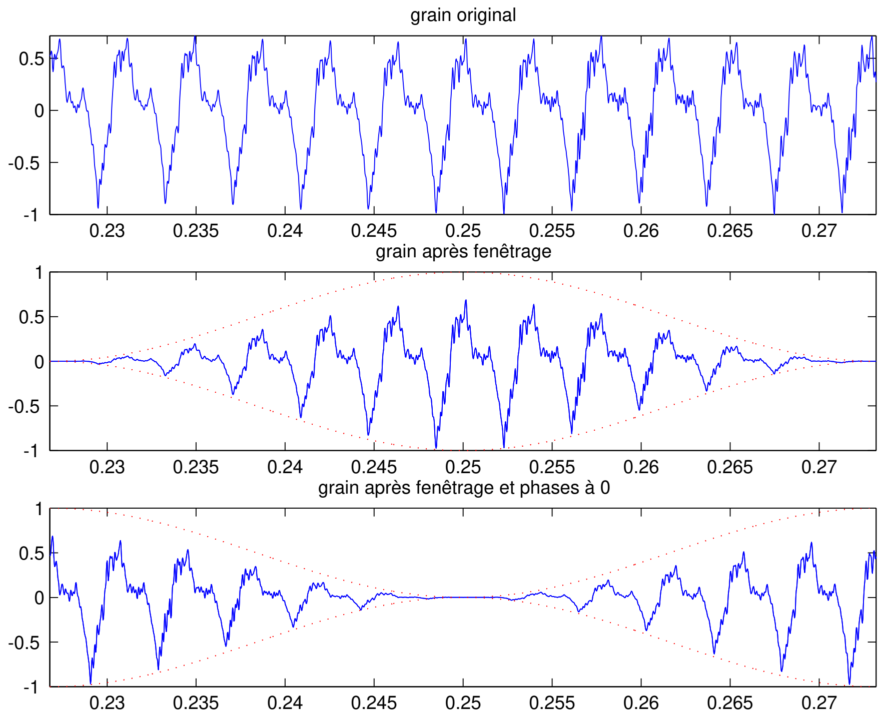

Fig. 1.25: *Illustration du décalage circulaire d'une demi-fenêtre avant application de la TF.*

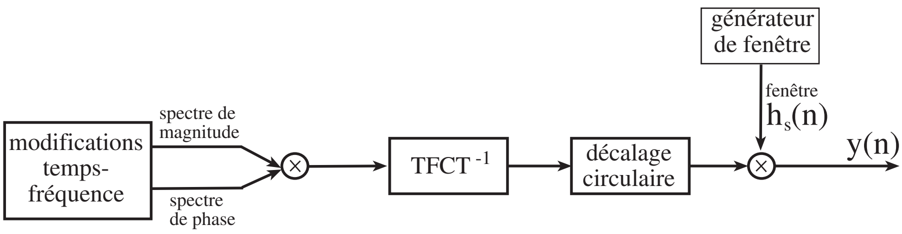

Fig. 1.26: *Diagramme de la synthèse par vocodeur de phase*

$$y = \text{princarg}(2\pi m + \phi_x) = \phi_x \in] - \pi; \pi] \tag{1.65}$$

On peut exprimer cette fonction "princarg" d'après la fonction "modulo" :

$$y = \text{mod}(\phi_y + \pi, -2\pi) + \pi \tag{1.66}$$

Etant données $\tilde{\phi}(sR_a, k)$ et $\tilde{\phi}((s+1)R_a, k)$ les phases de deux trames successives, on a :

$$\tilde{\phi}((s+1)R_a, k) = \Omega_k R_a + \tilde{\phi}(sR_a, k) \tag{1.67}$$

dans le cas où le panier de fréquence numéro k a une sinusoïde (c'est-à-dire un pic) de fréquence Ω_k. La phase déroulée s'exprime alors :

$$\tilde{\phi}_u((s+1)R_a, k) = \tilde{\phi}_c((s+1)R_a, k) + \tilde{\phi}_d((s+1)R_a, k) \tag{1.68}$$

avec $\tilde{\phi}_c((s+1)R_a, k)$ la phase cible et $\tilde{\phi}_d((s+1)R_a, k)$ la déviation par rapport à cette cible. Cette déviation s'exprime comme la différence entre la phase au temps $(s+1)\ R_a$ et sa cible :

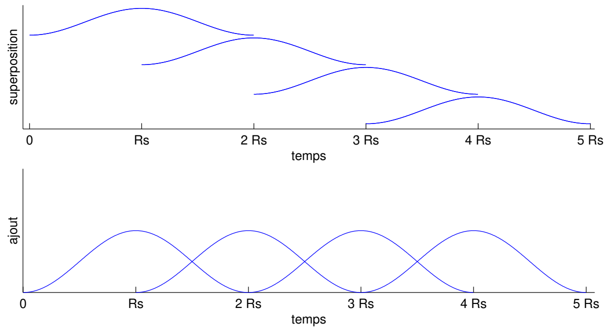

Fig. 1.27: *Ajout et superposition des fenêtres lors de la re-synthèse*

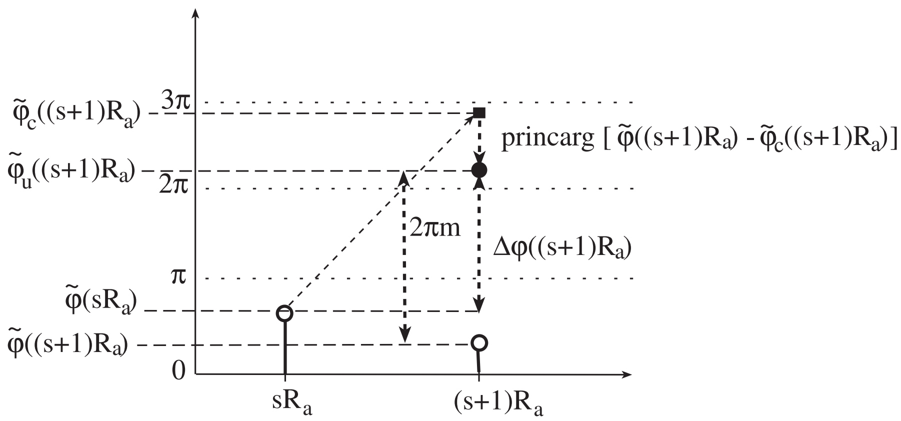

Fig. 1.28: *Déroulement de phase*

$$\tilde{\phi}_d\left((s+1)R_a,k\right) = \text{princarg}\left(\tilde{\phi}\left((s+1)R_a,k\right) - \tilde{\phi}_c\left((s+1)R_a,k\right)\right) \quad (1.69)$$

Remplaçons maintenant $\tilde{\phi}_t\left((s+1)R_a,k\right)$ par sa valeur donnée en (1.67) :

$$\begin{aligned}\tilde{\phi}_u\left((s+1)R_a,k\right) &= \tilde{\phi}\left(sR_a,k\right) + \Omega_k R_a \\ &+ \text{princarg}\left(\tilde{\phi}\left((s+1)R_a,k\right) - \tilde{\phi}\left(sR_a,k\right) - \Omega_k n R_a\right)\end{aligned} \quad (1.70)$$

La différence de phases déroulées est alors :

$$\begin{aligned}\Delta\phi\left((s+1)R_a,k\right) &= \tilde{\phi}_u\left((s+1)R_a,k\right) - \tilde{\phi}\left(sR_a,k\right) & (1.71)\\ &= \Omega_k R_a + \text{princarg}\left(\tilde{\phi}\left((s+1)R_a,k\right) - \tilde{\phi}\left(sR_a,k\right) - \Omega_k n R_a\right) & (1.72)\end{aligned}$$

La fréquence instantanée s'obtient alors aisément :

$$f_i\left((s+1)R_a,k\right) = \frac{1}{2\pi}\,\frac{\Delta\phi\left((s+1)R_a,k\right)}{R_a}\,F_e \quad (1.73)$$

1.7.3 Analyse-synthèse additive

La méthode additive se base sur l'observation suivante : les sons harmoniques peuvent être considérés comme la somme de sinusoïdes, dont l'amplitude et la fréquence évoluent dans le temps. De cette observation découle le modèle de McAulay et Quatieri [6]. Si l'on écoute un signal analysé puis resynthétisé selon cette méthode, on remarque que le bruit caractéristique du signal a disparu, si bien que le son de synthèse ne sonne pas exactement comme l'original : la représentation n'est pas suffisamment complète. Dans ce but, Serra [15] a modélisé l'enveloppe spectrale du résidu. Un raffinement possible consiste à différencier dans le résidu les transitoires d'attaque du bruit résiduel : c'est ce que Verma [13] propose.

Cette représentation du son est particulièrement adaptée à la voix, au sons de cordes (frottées et pincées), de vents (colonnes d'air), de cuivres.

Somme de sinusoïdes (McAulay, Quatieri)

Ce modèle était basé sur l'analyse-synthèse de la voix humaine à l'aide d'une somme de sinusoïdes.

Modèle Les sons quasi-stationnaires peuvent se modéliser comme une somme de sinusoïdes évoluant lentement dans le temps (en fréquence, phase et amplitude). Le modèle est donc le suivant :

$$x(n) = \sum_i A_i(n)\ \cos[\Theta_i(n)] \tag{1.74}$$

avec la phase :

$$\Theta_i(n) = \int_0^{nT} \omega_i(\tau)\ d\tau + \Phi_i \tag{1.75}$$

et avec $A_i(n)$ l'amplitude instantanée (enveloppe variant dans le temps), $\omega_i(n)$ la fréquence instantanée (piste fréquentielle de la i^{eme} sinusoïde, en radians) et $\Theta_i(n)$ la phase instantanée.

Analyse On applique un transformée de Fourier à court-terme. Les phases sont calculées à partir du spectre de phase, aux valeurs données par les pics (amplitude et fréquence) données dans le spectre de magnitude (cf. fig1.29).

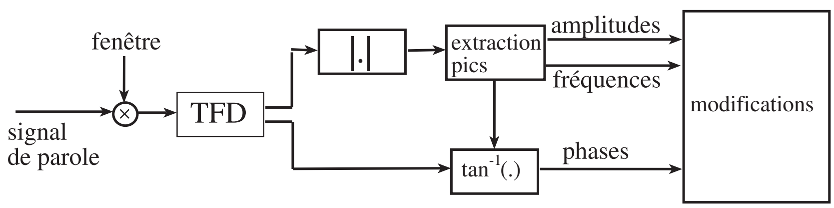

Fig. 1.29: *Diagramme de l'analyse additive MacAulay-Quatieri*

Synthèse La synthèse se fait elle aussi fenêtre par fenêtre. Les magnitudes sont interpolées linéairement, et servent d'enveloppes à des sinusoïdes générées après déroulement de phase et interpolation (linéaire ou non) entre deux phases successives dans le temps. Les sinusoïdes sont ensuite additionnées pour former le signal de synthèse (cf. fig1.30).

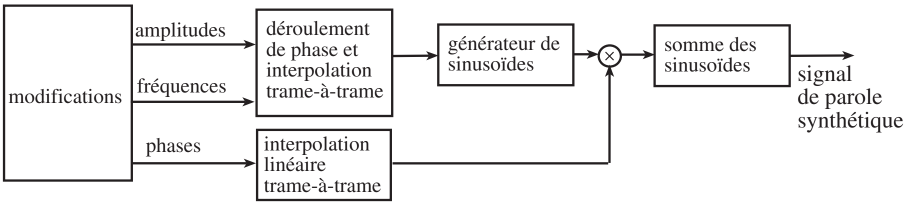

Fig. 1.30: *Diagramme de la synthèse additive MacAulay-Quatieri*

Modifications Les modifications peuvent avoir lieu sur chacun des paramètres extraits par le modèle : enveloppe, trajectoires fréquentielles des partiels. C'est à cet endroit que nous pouvons intervenir.

Somme de sinusoïdes + bruit (Serra, Smith)

Un raffinement du modèle précédent a été proposé en 1989 par Serra. Le signal est modélisé comme la somme de sinusoïdes (partie déterministe) et d'un bruit résiduel (partie stochastique). Ceci permet une meilleur preprésentation des sons non stationnaires. Le modèle s'appelle SMS, pour *Spectral Modelling Synthesis*.

Modèle Les sons peuvent se modéliser comme une somme de sinusoïdes évoluant lentement dans le temps (en fréquence, phase et amplitude) et d'un bruit filtré, dont l'enveloppe spectrale évolue lentement dans le temps. Le modèle est donc le suivant :

$$x(n) = \sum_{i}^{L} A_i(n)\ \cos[\Theta_i(n)] + e(n) \tag{1.76}$$

avec $A_i(n)$ l'amplitude instantanée, $\omega_i(n)$ la fréquence instantanée, $\Theta_i(n)$ la phase instantanée et $e(n)$ le bruit filtré à l'instant n. L'hypothèse de base du modèle est une variation lente en amplitude et en fréquence de la partie déterministe du signal.

Analyse L'analyse se fait par extraction des pics, calcul de la partie déterministe (somme de sinusoïdes) et soustraction au signal (temporelle, échantillon par échantillon), afin d'analyser ensuite la partie stochastique.

Dans chaque fenêtre, on applique une fenêtre de Blackman-Harris, dont la propriété est d'avoir un rapport lobe primaire sur lobes secondaires de 92 dB : la modélisation du premier lobe suffit donc, puisque les contributions des autres lobes est noyée dans le bruit de quantification du signal codé sur 16 bits. De plus, l'interpolation entre paniers de fréquence en décibels se fera par une parabole (propriété de cette fenêtre).

Pour éviter les effets de phase linéaire introduits par la fenêtre, chaque trame temporelle se voit appliquer un centrage autour de l'origine avant l'application de la transformée de Fourier. Après extraction des pics, la fréquence fondamentale est calculée (si elle existe) et sert à adapter la taille de la fenêtre d'analyse (analyse pitch-synchrone). Une procédure de continuation de pics permet de donner les trajectoires sinusoïdales (ie. attribuer à chacun des pics de la trame $n-1$ les bons successeurs, et à chacun des pics de la trame n les bons prédécesseurs) : il s'agit de l'algorithme de Maher et Beauchamp permettant d'estimer la fondamentale tout en effectuant le suivi des partiels [5].

Le résidu est calculé par resynthèse de la partie déterministe et soustraction au signal original dans le domaine temporel. Après application de la transformée de Fourer à court-terme sur le résidu, on le modélise à l'aide de l'une des méthodes suivantes :

- interpolation par des splines ;
- approximation par segments de droites ;
- codage par prédiction linéaire (LPC).

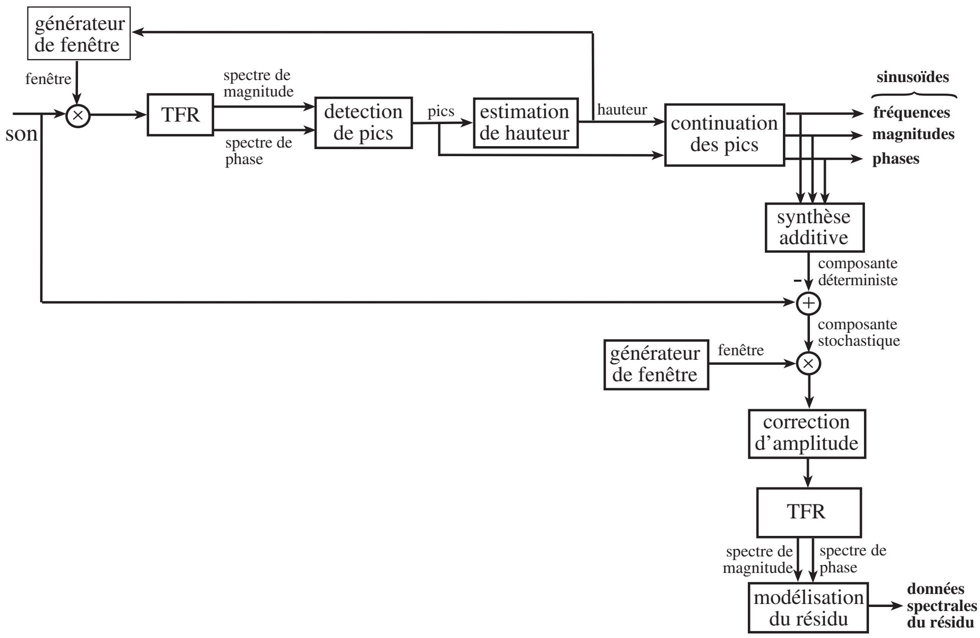

Fig. 1.31: *Diagramme de l'analyse additive + bruit (Serra)*

Synthèse La partie déterministe est générée par une synthèse additive sans information de phase, ce qui permet de synthétiser le signal dans le domaine temporel (synthèse de sinusoïdes) comme dans le domaine fréquentiel, basé sur le transformée de Fourier inverse. En terme de calculs, cette dernière méthode est la moins coûteuse (et donc la plus rapide).

La synthèse de la partie stochastique peut être vue comme un bruit blanc passant dans un filtre variant dans le temps, ce qui s'implémente généralement comme la convolution dans le domaine temporel d'un bruit blanc avec une réponse impulsionnelle de filtre. En pratique, on applique une FFT inverse à un spectre complexe, constitué de la forme spectrale du résidu et de phases aléatoires.

Les deux composantes peuvent s'ajouter dans le domaine fréquentiel (une seule FFT inverse à calculer) ou dans le domaine temporel.

Modifications Les modifications que l'on peut effectuer portent sur la partie stochastique (modifications d'enveloppe temporelle ou fréquentielle) et sur la partie déterministe (identiques à celles du modèle "somme de sinusoïdes").

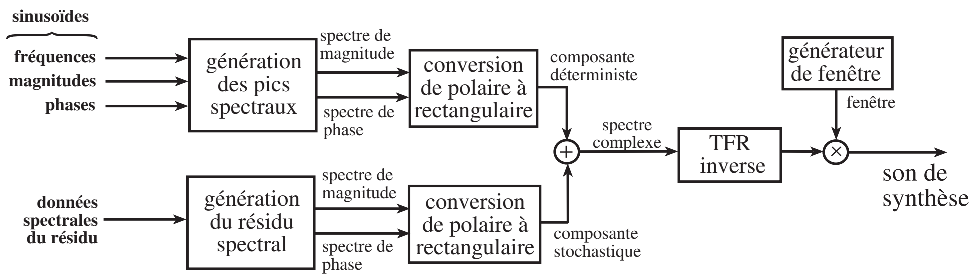

Fig. 1.32: *Diagramme de la synthèse additive + bruit (Serra)*

Remarques Un grand intérêt de cette méthode est que l'analyse est rendue indépendant de la synthèse. Ainsi, les paramètres d'analyse (taille et type de fenêtre, taille de FFT, pas entre deux trames) peuvent être choisis de façon à optimiser l'analyse. Plusieurs implémentations de ce modèle existent, dont **SMS** [3] (renommé récemment CLAM [4]) du Music Technology Group de Serra, **Additive** de l'Ircam, **SAS** développé au Labri par Sylvain Marchand[8].

Somme de sinusoïdes + transitoire + bruit (Verma)

Un extension intéressante du modèle précédent consiste à prendre en compte les transitoires dans la partie résiduelle et de les séparer de celle-ci, en les modélisant par un petit jeu de paramètres. Ainsi, chacune des trois composantes du signal peut faire l'objet de modifications distinctes, indépendantes.

Modèle

$$x(n) = \sum_{i}^{L} A_i(n)\ \cos[\Theta_i(n)] + r_t(n) + r_b(n) \qquad (1.77)$$

avec $r_t(n)$ le résidu concernant la partie transitoire, et $r_b(n)$ la partie bruitée.

Analyse Une fois l'anayse de type SMS effectuée, le premier résidu obtenu est analysé. Un éventuel transitoire est détecté, paramétrisé, synthétisé et soustrait au résidu. Le second résidu obtenu est la composante stochastique de ce modèle. Cette composante peut alors être modélisée à son tour, selon la méthodologie expliquée pour le modèle "somme de sinusoïdes + bruit".

L'algorithme utilisé pour la modélisation du transitoire tient compte de la dualité entre sinusoïdes et transitoire. Une sinusoïde variant lentement dans le domaine temporel correspond à un pic, une impulsion dans le domaine fréquentiel. C'est ainsi qu'on peut la détecter dans une TFCT, à l'aide du spectre de magnitude. Des signaux impulsifs dans le domaine temporel deviennent oscillants dans le domaine fréquentiel.

La première étape consiste donc à mettre en correspondance le domaine temporel des transitoires avec des signaux sinusoïdaux dans un certain domaine fréquentiel. La transformée en cosinus discret (TCD, *DCT : dicrete cosine transform)* permet ce type de correspondance.

La transformée en cosinus discret est définie de la manière suivante :

$$C(k) = \beta(k) \sum_{n=0}^{N-1} x(n)\ \cos\left[\frac{(2n+1)\,\pi}{2\,N}\right], \quad n, k \in \{0, 1, \ldots, N-1\} \qquad (1.78)$$

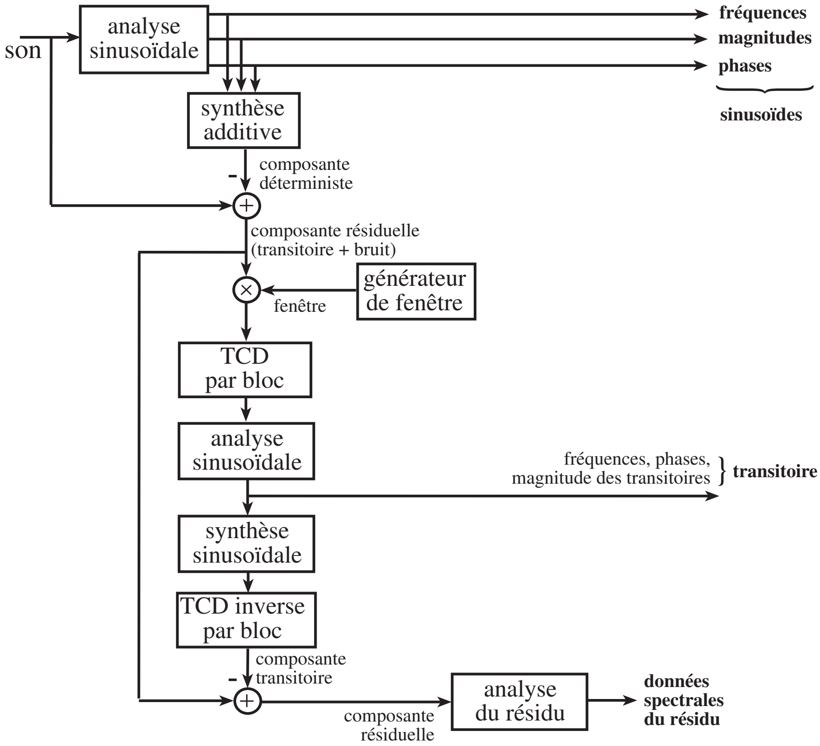

Fig. 1.33: *Diagramme de l'analyse additive + transitoire + résidu (Verma)*

avec $\beta(k) = \sqrt{1/N}$ si $k = 1$ et $\beta(k) = \sqrt{2/N}$ sinon. Une impulsion au début d'une trame temporelle correspondra à la présence d'une cosinus basse fréquence, alors qu'une impulsion à la fin d'une trame temporelle correspondra à la présence d'une cosinus haute fréquence. Inversement, lorsque l'on modife la fréquence de la sinusoïde correspondant à un transitoire, on le déplacera dans le temps. L'algorithme consiste à prendre des blocs qui ne se superposent pas, calculer une TCD sur chaque bloc, extraire les paramètres de la transitoire en appliquant un modèle sinusoïdale sur chaque trame DCT (ie. sur chaque bloc).

Synthèse La synthèse se fait par reconstruction des sinusoïdes du domaine DCT et l'utilisation de la transformée inverse (TCDI). Le transitoire synthétisé est ajouté à la partie déterministe et à la partie stochastique de synthèse.

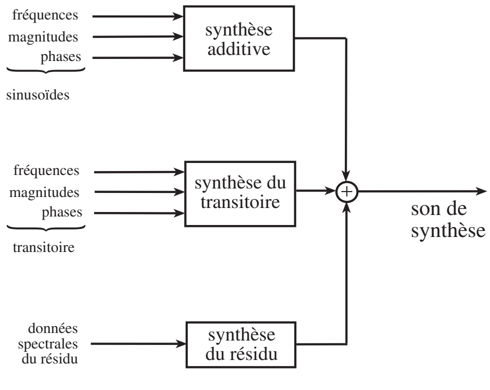

Fig. 1.34: *Diagramme de la synthèse additive + transitoire + résidu (Verma)*

Modifications Elles peuvent ne s'appliquer que sur l'une des composantes, ou plusieurs à la fois. Attention, lors du changement d'échelle temporelle, de bien appliquer le même facteur aux trois composantes, afin de s'assurer que la synchronisation du transitoire avec le reste ait bien lieu !

1.7.4 Soustractive, ou source-filtre

En partant d'un modèle de la voix humaine constitué d'une source riche en harmoniques, filtrée dans plusieurs cavités successives, on obtient un modèle de synthèse soustractive. C'est un modèle très performant pour décrire la voix humaine : en effet, les cordes vocales produisent un train d'onde riche en harmoniques, et le larynx, la fosse nasale et la cavité buccale sont des cavités filtrant le signal, responsables de l'apparition des formants dans le spectre d'un signal. La reconnaissance de la parole (reconnaissance naturelle faite par chacun d'entre nous, ou automatique à l'aide de systèmes informatiques), et particulièrement des voyelles, se fait grâce à ces formants. De ce fait, la reconnaissance d'autres sons (les sons instrumentaux, notamment) se fait aussi, en partie, à l'aide des formants.

Fort de ces constatations, les chercheurs ont utilisé le modèle soustractif ou source-filtre pour la synthèse comme pour la modification de sons. Ainsi, en changeant la source ou les filtres, on modifie le timbre d'un son.

La grande difficulté réside en l'extraction de l'enveloppe spectrale. Pour les sons harmoniques, il s'agit de la courbe passant par les pics spectraux des sinusoïdes présentes dans le signal. Dans ce cas, se pose le problème de connaître précisément l'emplacement de ces pics (l'utilisation de la méthode additive répond à cette question), et de savoir quelle interpolation est la plus censée dès lors que l'on veut connaître la valeur de l'enveloppe spectrale entre les pics. Dans le cas de sons inharmoniques ou bruités, cette définition de l'enveloppe spectrale n'est plus valide, et la notion d'enveloppe spectrale est tatolement dépendante de ce que l'on considère comme source, excitation et de ce que l'on considère comme filtre (ou résonance) dans le signal bruité ou inharmonique.

Une approximation de la notion d'enveloppe spectrale consiste à dire que c'est un lissage du spectre d'amplitude, qui tend à laisser de côté les lignes spectrales tout en conservant la forme générale du spectre.

Nous allons maintenant présenter deux principales manières d'estimer l'enveloppe spectrale : la prédiction linéaire et le cepstre.

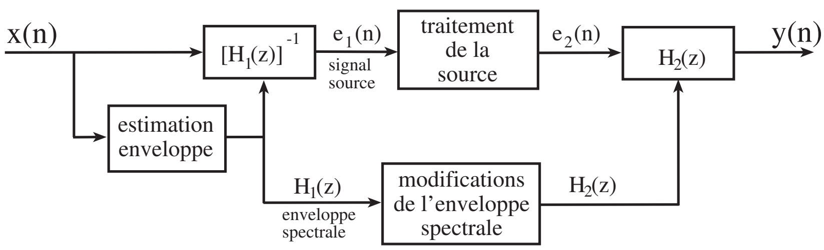

Fig. 1.35: *Diagramme de l'analyse-synthèse soustractive*

Prédiction linéaire (LPC)

La LPC est un modèle source-filtre dont le filtre est tous-pôles (il ne contient que des pôles) et représente l'enveloppe spectrale. Ce modèle a été développé pour la voix, en considérant à juste titre que l'excitation (ou la source) des sons voisés correspondent à un train d'onde de période la fréquence fondamentale, tandis que la source des sons non voisés est un bruit. Dans les deux cas, l'excitation est filtrée par les formants. Le *codage à prédiction linéaire* (ou *LPC, linear prediction coding*) modélise le filtre par un filtre linéaire. Ainsi, pour un signal sonore $x(n)$, le modèle utilisé est :

$$\hat{x}(n) = \sum_{k=1}^{p} a_k \; x(n-k) \tag{1.79}$$

avec p l'ordre du filtre, ou ordre de prédiction, et a_k les coefficients du fitre, ou coefficients de prédiction. La qualité de cette prédiction est évaluée grâce à l'erreur de prédiction, ou résidu :

$$e(n) = x(n) - \hat{x}(n) \tag{1.80}$$

La transformée en z du fitre de prédiction est donnée par :

$$P(z) = \sum_{k=1}^{p} a_k \; z^{-k} \tag{1.81}$$

et l'erreur de prédiction dans le domaine de la transformée en z :

$$E(z) = X(z) - \hat{X}(z) = X(z) \;\; (1 - P(z)) \tag{1.82}$$

Fig. 1.36: *Diagramme de l'analyse et de la syntèse soustractive par LPC*

Comme on peut le voir en figure 1.36, l'analyse LPC correspond à une prédiction "*feed-forward*". Le filtre inverse, ou filtre d'erreur de prédiction est $A(z) = 1 - P(z)$, et l'erreur s'exprime alors :

$$E(z) = X(z) \; A(z) \tag{1.83}$$

Le son est obtenu en bout de chaîne de traitement à l'aide du signal d'excitation (éventuellement modifié) $\tilde{x}(n)$ en entrée du filtre tout pôle $H(z) = \frac{1}{A(z)} = \frac{1}{1-P(z)}$, d'où :

$$Y(z) = \tilde{E}(z) \; H(z) \tag{1.84}$$

avec $H(z)$ obtenu par une boucle de réinjection, cf. figure 1.36 à droite. Le filtre RII (à réponse impulsionnelle infinie) $H(z)$ est appelé filtre LPC ou filtre de synthèse, et représente le modèle spectral (à un facteur gain près) de $x(n)$.

Si le filtre est optimal, l'erreur résiduelle est minimisée. Cette méthode est utilisée en télécommunications, où le résidu quantifié $\tilde{e}(n) = Q(e(n))$ est utilisé comme signal d'excitation du filtre LPC.

Calcul des coefficients Il existe différentes méthodes de calcul des coefficients du filtre LPC ; la plupart des implémentations les calculent à partir d'un grain du signal $x(n)$, et remettent à jour le filtre régulièrement. C'est le cas des méthodes de covariance (filtres non stables), d'autocorrélation et l'algorithme de Burg (ces deux dernières utilisent une modélisation en filtres en treillis, ce qui impliquent l'obtention de filtres toujours stables, ce que l'on recherche avant tout).

Nous présentons ici la méthode d'autocorrélation. L'énergie de l'erreur de prédiction est donnée par l'espérance mathématique de l'erreur : $E_p = E\{e^2(n)\}$. Le filtre est optimal lorsque l'énergie de l'erreur de prédiction est minimal, donc lorsque sa dérivée est nulle. Sa dérivée est données par :

$$\frac{\partial E_p}{\partial a_i} = 2\ E\left\{e(n).\frac{\partial e(n)}{\partial a_i}\right\} = -2E\left\{e(n)\ x(n-i)\right\} \tag{1.85}$$

La condition d'optimalité $\frac{\partial E_p}{\partial a_i} = 0$ s'écrit :

$$-2\ E\left\{\left(x(n) - \sum_{k=1}^{p} a_k\ x(n-k)\right)\ x(n-i)\right\} \quad = \quad 0 \tag{1.86}$$

D'où la formulation des équations normales :

$$\sum_{k=1}^{p} a_k\ E\left\{x(n-k)\ x(n-i)\right\} = E\left\{x(n)\ x(n-i)\right\} \tag{1.87}$$

L'autocorrélation d'une séquence de N valeurs de $u(n)$ est définie par :

$$r_{xx}(i) = \sum_{n=i}^{N-1} u(n)\ u(n-i) \tag{1.88}$$

avec $u(n) = x(n)\ w(n)$ le fenêtrage de x par la fenêtre $w(n), n = 0, \ldots, N-1$ (dans notre cas, on utilise la fenêtre de Hamming). On peut remplacer les espérances par leurs approximations en utilisant l'autocorrélation, d'où la nouvelle formulation des équations normales :

$$\sum_{k=1}^{p} a_k r_{xx}(i-k) = r_{xx}(i) \qquad i = 1, \ldots, p \tag{1.89}$$

Les coefficients a_k sont obtenus par résolution de ces équations normales, à l'aide de la récursion de Levinson-Durbin. Minimiser l'énergie du résidu revient à trouver la meilleur correspondance spectrale dans le domaine fréquentiel, à un facteur gain près. Dans ce cas, le modèle d'entrée $x(n)$ est modélisé par le filtre :

$$H_g(z) = G.H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.90}$$

avec G le facteur gain. Avec ce filtre de synthèse modifié, le signal original est modélisé en utilisant comme signal d'excitation un bruit blanc de variance unité. Pour la méthode d'autocorrélation, le facteur de gain est défini par :

$$G^2 = r_{xx}(o) - \sum_{k=1}^{p} a_k\ r_{xx}(k) \tag{1.91}$$

Remarquons que ce gain dépend de l'énergie de l'erreur de prédiction.

Cepstre

Terme inventé par Noll et Tuckey en 1964, le cepstre[1] est une méthode qui permet l'estimation de l'enveloppe spectrale à l'aide de la FFT d'une trame temporelle du signal fenêtrée. Selon la précision que l'on désire obtenir sur l'enveloppe spectrale, le bourrage de 0 ainsi que les fenêtres de Hanning, Hamming, et de Gauss peuvent être utilisées.

Le cepstre complexe est défini comme la transformée de Fourier à court-terme (TFCT) inverse du logarithme de la TFCT $X(n)$ du signal :

$$c(n) = \mathcal{F}^{-1}\left(\log\left(X\right)\right) \tag{1.92}$$

Le cepstre réel est défini comme la TFCT inverse du logarithme du module $|X|$ de la TFCT du signal :

$$c_R(n) = \mathcal{F}^{-1}\left(\log|X|\right) \tag{1.93}$$

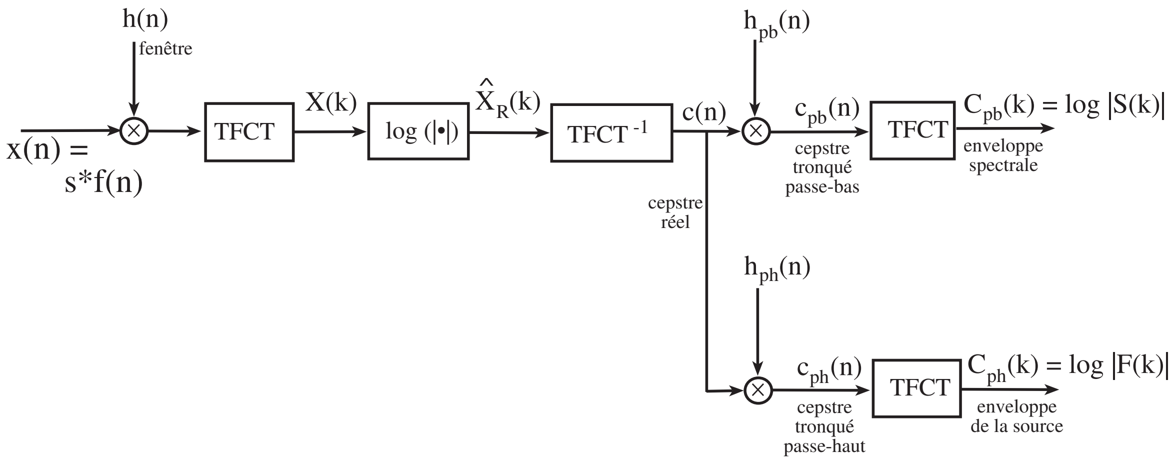

Fig. 1.37: *Diagramme de l'analyse soustractive par cepstre*

Lorsque l'on filtre passe-bas le cepstre, on obtient un lissage du spectre. Si le filtre est bien choisi, le lissage correspond à une bonne approximation de l'enveloppe spectrale. Soit $h_{pb}(n)$ la fenêtre utilisée pour le filtrage passe-bas, définie par :

$$h_{pb}(n) = \begin{cases} 1 & n = 0, N/2 \\ 2 & 1 \leq n < N/2 \\ 0 & N/2 < n \leq N-1 \end{cases} \tag{1.94}$$

Pour la définition de cette fenêtre, on utilise le fait que le cepstre réel est une fonction paire, donc sa TFCT inverse $c(n)$ est liée au cepstre complexe par la formule $c_R(n) = \frac{c(n)+c(-n)}{2}$.

Le filtrage passe-bas $c_{pb}(n)$ du cepstre réel est obtenu par convolution de la fenêtre $h_{pb}(n)$ avec le cepstre réel $c(n)$. L'enveloppe spectrale est obtenue par TFCT de $c_{pb}(n)$:

$$\mathcal{S} = \exp\left(\mathcal{F}\left(c_{pb}\right)\right) \tag{1.95}$$

En utilisant pour cepstre réel $c_R = \mathcal{F}^{-1}\left(20\ \log_{10}\left(|X|\right)\right)$, on obtient l'enveloppe spectrale en décibels sans utiliser la fonction exponentielle : $\mathcal{S} = \mathcal{F}\left(c_{pb}\right)$.

[1]Cepstre : terme obtenu par inversion des 4 premières lettres du mot "spectre".

Si l'on considère le signal $x(n)$ constitué comme la convolution d'une source $s(n)$ et d'un filtre $f(n)$, sa transformée de Fourier est donnée comme le produit des transformées de Fourier de s et de f : $\mathcal{F}(x) = \mathcal{F}(s).\mathcal{F}(f)$, ce qui reste vrai pour les modules :

$$|\mathcal{F}(x)| = |\mathcal{F}(s)|.|\mathcal{F}(f)| \tag{1.96}$$

En échelle logarithmique, un produit devient une somme :

$$\log(|\mathcal{F}(x)|) = \log(|\mathcal{F}(s)|) + \log(|\mathcal{F}(f)|) \tag{1.97}$$

La séparation source-filtre s'obtient alors aisément avec la méthode du cepstre, par l'utilisation de deux fenêtres, l'une h_{pb} passe-bas pour extraire l'enveloppe spectrale, l'autre passe-haut h_{ph} pour extraire la source. Les valeurs du cepstre sont des temps ; aussi, les valeurs de temps basses, appelées "quéfrences", correspondent à des variations lentes du spectre de magnitude (donc à l'enveloppe spectrale), et les hautes quéfrences correspondent aux variations rapides, aux pics du spectre de magnitude (soit à la source).

Afin de trouver un bon seuil de quéfrence pour le calcul de la source et du filtre par la méthode du cepstre, on peut utiliser le fait que les quéfrences sont des variables temporelles, donc si le signal est harmonique, le cepstre sera périodique (la période correspondant à la fondamentale du signal harmonique). Dans ce cas, cette valeur est la quéfrence maximale définissant l'enveloppe. Une valeur inférieure lissera l'enveloppe, une valeur supérieure incluera une partie des pics des harmoniques (ou des partiels) dans l'enveloppe.

Remarques Tout d'abord, notons que la troncature effectuée lors du filtrage passe-bas correspond à un filtrage dans le domaine fréquentiel par $\frac{\sin(f)}{f}$. Une alternative est d'utiliser une fenêtre dont la transition est plus lisse.

Ensuite, l'utilisation du logarithme donne des valeurs variant entre $-\infty$ et $0\ dB$. Pour éviter les problèmes de calculs dûs à des valeurs infinies, la valeur de l'opérande est limitée par l'addition d'une valeur très faible (10^{-5}), afin d'obtenir $-100\ dB$ comme limite inférieure.

Une manière d'améliorer le calcul de l'enveloppe spectrale consiste à utiliser un algorithme itératif qui ne calcule que la différence positive entre le spectre instantané et l'enveloppe spectrale estimée à chaque itération.

Enfin, même si l'on utilise très couramment le spectre réel, le spectre complexe peut lui aussi être utilisé pour réaliser l'estimation de l'enveloppe spectrale. Dans ce cas, l'enveloppe spectrale est définie par une TFCT complexe.

Chapitre 2

Effets et transformations sonores

2.1 Introduction

Nous présentons ici un ensemble assez complet, bien que non exhaustif, des effets et transformations audionumériques possibles. Certains peuvent s'obtenir par plusieurs procédés, pour lesquels des différences audibles de résultat existent. Nous nous limiterons à présenter les familles d'effets, ainsi qu'au moins une des méthodes d'obtention de chaque effet.

Plusieurs approches permettent de présenter les traitements selon différentes classes. Nous en présentons trois faisant sens : la taxonomie méthodologique, la taxonomie cognitive et la taxonomie perceptive.

2.1.1 Taxonomie méthodologique

On peut présenter les traitements en fonction des méthodes d'implémentation ; c'est le choix qui a été fait pour la rédaction du livre "DAFx: Digital Audio Effects" [16]. Les méthodes ainsi mises en avant sont :

- filtrages ;
- lignes à retard ;
- modulateurs et démodulateurs ;
- traitements non-linéaires ;
- effets spatiaux ;
- traitement segment-temporels ;
- traitements temps-fréquence (vocodeur de phase) ;
- traitements source-filtre (cepstre, LPC, vocodeur à canaux) ;
- traitements spectraux (basés sur une analyse additive) ;
- conformation temporelle et fréquentielle.

Cette classification (ou taxonomie) est dite **méthodologique**.

2.1.2 Taxonomie cognitive

Un deuxième classification possible est donnée dans [7], et permet plusieurs angles de lectures. Ce répertoire des effets sonores est le fruit de collaborations et recherches entre ingénieurs, chercheurs, architectes, urbanistes, sociologues, philosophes, géographes, musicologues et chercheurs du Cresson, et porte sur différents niveaux d'interprétation et de compréhension des effets, d'où le nom de taxonomie cognitive. Les catégorie proposées sont :

- effets élémentaires (ex: filtrage, distorsion, résonance, réverbération) : ils concernent la matière sonore (selon la taxonomie perceptive) ou la modalité de propagation. Ils sont enracinés dans la connaissance acoustique contemporaine et sont tous quantifiables ;
- effets de composition (ex: masquage, traînage, coupure, bourdon, effet téléphone) : 'ils concernent des agencements sonores complexes et se sont définis par des caractères remarquables touchant soit à la dimension synchronique, soit à la dimension diachronique du contexte. Tous ces effets dépendants du dispositif spatio-temporel de la propagation sont sujets à évaluation physique pour une part au moins de leurs composantes" ;
- effets liés à l'organisation perceptive (ex: gommage, synecdoque, rémanence, anticipation, métabole) : "ils sont dûs en priorité à l'organisation perceptive et mnémique des individus en situation concrète. On les repère toujours à partir d'une expression ou d'une aperception de la part des entendants. Par ailleurs, les caractères propres à la culture et la sociabilité de références sont partie prenante dans les particularités et la force de l'effet" ;
- effets psychomoteurs (ex: enchaînement, créneau, attraction, effet phonotonique) : "ces effets impliquent l'existence d'une action sonore de l'entendant ou tout au moins d'une esquisse motrice ou d'un schème faisant interagir perception et motricité" ;
- effets sémantiques (ex: décalage, imitation) : "ces effets sonores jouent sur l'écart de sens entre le contexte donné et la signification émergente. Il y a toujours décontextualisation, que ce soit sous la forme de l'imprévu anxiogène, de l'humour, du jeu conscient, ou d'une valeur esthétique ajoutée".

Cette classification présente les effets dans les deux sens du termes, à la fois en tant que traitements sonores et en tant qu'effet perçu lors de l'audition de situations sonores. C'est un cadre trop général pour ce cours, aussi nous nous contentons des traitements sonores.

2.1.3 Taxonomie perceptive

On peut enfin présenter les traitements en fonction des paramètres perceptifs qui seront modifiés dans le son. Nous avons choisi cette dernière classification, dite perceptive, pour présenter le reste de l'exposé, du fait qu'elle se permet de se placer du point de vue (ou plutôt d'écoute) de l'auditeur qui manipule des traitements audionumériques à sa disposition.

Les différents axes perceptifs sont :

- la dynamique : c'est la sensation d'intensité sonore perçue, autrement dit la sonie ;
- la hauteur : on parle ici de la hauteur perçue, et non de la hauteur fondamentale (ou fréquence fondamentale) des sons instrumentaux par exemple, même si dans le cas précis des sons instrumentaux et des autres sons tonaux, les deux sont quasiment identiques ;

- l'espace : il s'agit de la localisation des sources, de leur spatialisation ou de leurs déplacements, et de leur rayonnement ;
- la durée des sons, en secondes ;
- le timbre : il est alors représenté par un espace de timbres, dont les dimensions ne sont par les mêmes selon que l'on décrit un son tonal ou non.

Bien évidemment, on pourra aussi trouver des effets agissant sur plusieurs axes à la fois. Nous pouvons maintenant présenter les effets un par un.

2.2 Traitements modifiant la dynamique

Les effets que nous allons présenter ici sont des processus modifiant la dynamique. Certains sont non linéaires, ou dynamiques (que l'on peut considérer comme des effets adaptatifs). Ils effectuent dans un premier temps une mesure du niveau d'énergie du signal d'entrée, et modifient ensuite le gain appliqué à ce signal.

2.2.1 Amplification

Définition

L'amplification d'un signal consiste à appliquer un gain de façon à augmenter ou diminuer l'amplitude de tous les échantillons. Ceci constitue une amplification de base. En effet, les amplificateurs haute-fidélité utilisent différents étages de traitement, entre autres pour apporter des corrections au spectre (équalisation). Ceci n'est pas abordé dans cette partie.

Fonctionnement

La manière d'effectuer une amplification de base consiste à définir un gain g_{ampli} en décibel, et à multiplier tous les échantillons par ce gain en amplitude. Soit $x(n)$ le signal à amplifier, le signal amplifié est donné par :

$$y(n) = 10^{g_{ampli}/20} x(n) \tag{2.1}$$

2.2.2 Normalisation

Définition

la normalisation consiste à appliquer un gain à un signal de sorte que son niveau maximal soit de 0 dB.

Fonctionnement

La manière d'effectuer la normalisation d'un signal est très simple, et se fait hors temps réel. On calcule le maximum de la valeur absolue du signal, et l'on divise la valeur de chaque échantillon par ce maximum. Soit $x(n)$ le signal à normaliser, le maximum M est donné par :

$$M = \max_n |x(n)| \tag{2.2}$$

Le signal normalisé est donné par :

$$y(n) = \frac{x(n)}{M} \tag{2.3}$$

Intérêt de cet effet

Cet effet sert à utiliser toute la plage de dynamique que le codage en 8, 16 ou 24 bits permet lors de l'écriture en fichier ou le stockage en mémoire.

2.2.3 Expandeur, *noise gate*

Définition

L'expandeur est un traitement non linéaire qui augmente la dynamique d'un signal de façon à ce que les signaux de bas niveau sonore soient atténués tandis que les signaux de haut niveau sonore ne sont ni atténués, ni augmentés. Le *noise gate* correspond à un expandeur poussé à l'extrême : l'entrée de faible niveau est atténuée fortement, voire éliminée et remplacée par du silence.

Fonctionnement

L'expandeur est un amplificateur à gain variable. Le gain ne dépasse jamais 1 et est contrôlé par le niveau d'énergie du signal. Lorsque ce niveau est élevé, l'expandeur a un gain unitaire, tandis que lorsque le niveau diminue, le gain décroît de même, rendant le niveau du signal plus bas. La figure 2.1 montre la structure d'un expandeur.

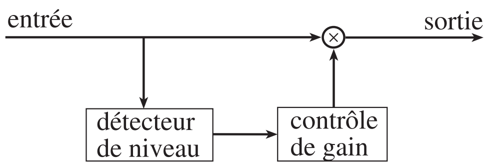

Fig. 2.1: *Diagramme de l'expandeur et du compresseur/limiteur*

Le niveau d'énergie correspond à une mesure pondérée du signal. Par exemple, le RMS est un bon indicateur du niveau d'énergie d'un signal.

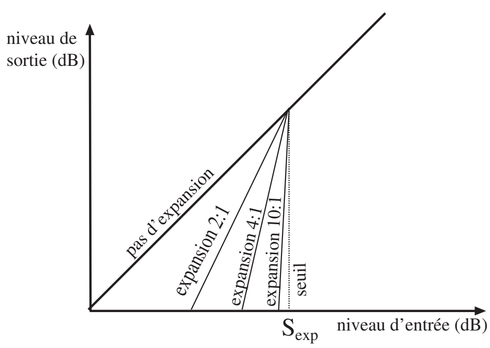

Fig. 2.2: *Relation entrée-sortie de l'expandeur*

On peut représenter la relation entrée-sortie d'un expandeur à l'aide d'un graphique tel que fig. 2.2. Le niveau d'entrée est représenté en abcisse, le niveau de sortie en ordonnée.

Le point de la courbe S_{exp} où la pente change est appelé **seuil** du limiteur et est ajustable par l'utilisateur. C'est la valeur du niveau d'entrée à partir de laquelle le contrôle de gain est effectif. Le rapport d'expansion R_{exp} indiqué (par exemple 4 : 1) s'applique en Décibel (dB), puisque c'est l'unité de mesure de niveau sonore. Le *noise gate* correspond à un rapport d'atténuation de 10 : 1.

$$N_{sortie} = \begin{cases} N_{entree} & \text{si } N_{sortie} \geq Seuil \\ R_{exp}N_{entree} + (1 - R_{exp})Seuil & \text{si } N_{sortie} \leq Seuil \end{cases} \tag{2.4}$$

La fonction d'évaluation du niveau sonore s'effectuant sur une moyenne du signal entrant, il existe un temps de retard entre un changement de niveau dans le signal d'entrée et le déclenchement d'un changement de gain. On appelera **temps d'attaque** (noté $\tau_{a,exp}$) le temps nécessaire à l'expandeur pour restaurer à 1 le gain d'un signal lorsque le niveau d'entrée dépasse le seuil. De la même manière, le temps nécessaire à l'expandeur pour réduire le gain après que le signal est passé en dessous du seuil s'appelle **temps de relaxation** et se note $\tau_{r,exp}$.

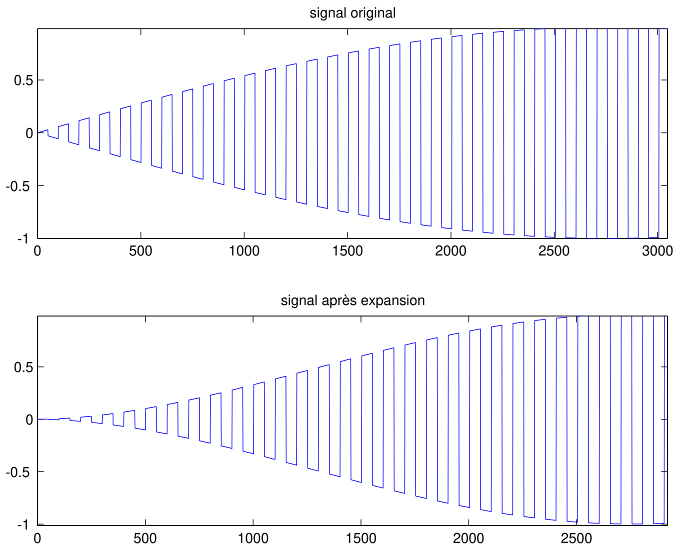

Fig. 2.3: *Signal carré modulé en amplitude avant et après expansion*

Intérêt de cet effet

Cet effet est utilisé dans les appareillages électroniques. Par exemple, il sert à accroître la dynamique d'enregistrement sur cassettes ou disques vinyls dont la dynamique est limitée. En utilisant le limiteur lors de la restitution, la dynamique sera alors plus étendue.

La plus grande application de l'expandeur est sans doute la réduction de bruit de fond. Cela permet de réduire les risques d'effets de larsen (*feedback*). Le *noise gate* est utilisée pour éliminer les bruits ou masquer des bruits risquant d'être amplifiés ou entendus lorsque un instrument ne joue pas. Le seuil doit alors être suffisamment élevé pour supprimer les bruits indésirables, mais pas trop non plus pour éviter de couper intempestivement des notes instrumentales ou prématurément les parties décroissantes de notes.

Les expandeurs peuvent aussi être utilisés conjointement aux compresseurs pour réduire les effets de bruit de fond d'une chaîne de transmission de signal, audio ou autre. Tout canal de transmission ayant une bande passante dynamique réduite, on compresse le signal à transmettre afin d'élever le niveau moyen du signal par rapport au niveau du bruit. Un expandeur est utilisé à la réception afin de rendre au signal sa dynamique originale. On appelle *companding* ce processus de compression-expansion. Un exemple de ce processus est la technique de réduction de bruit Dolby A utilisé pour les cassettes.

Compléments d'information

Le placement de cet effet dans une chaîne d'effet est d'une grande importance. La plupart des effets produisant du bruit, on préfère placer l'expandeur en fin de chaîne, de façon à ce que le bruit généré et éventuellement amplifié soit inaudible lorsqu'aucun signal n'est envoyé en entrée (le musicien utilisant la chaîne d'effets de joue pas). Cependant, il est préférable de placer les effets d'écho et de réverbération en dernier, pour éviter de dupliquer le bruit, et aussi afin d'éviter une rupture brutale de derniers échos ou des dernières répliques de la réverbération.

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser cet effet sont :

- le seuil S_{exp};
- le rapport d'expansion R_{exp} ;
- le temps d'attaque $\tau_{a,exp}$;
- le temps de relaxation $\tau_{r,exp}$.

2.2.4 Compresseur, limiteur

Définition

Le compresseur est un traitement non linéaire qui réduit la plage dynamique d'un signal. Il est très utilisé en enregistrements audio, en production, pour la réduction de bruit de fond, et pour les performances, mais doit être manipulé avec soin.

Fonctionnement

Le diagramme de fonctionnement du compresseur est le même que celui de l'expandeur (cf. fig. 2.1), seule la fonction de contrôle de gain diffère. Le gain est réduit lorsque le niveau du signal est élevé, ce qui rend les passages forts un peu plus doux, en réduisant la plage dynamique.

La relation entrée-sortie du compresseur (cf. fig. 2.4) diffère de celle de l'expandeur. Elle s'explique de la même manière que la relation entrée-sortie de l'expandeur : au dessus du seuil S_{comp}, un rapport de compression R_{comp} de 4 : 1 signifie qu'il faudra que le niveau d'entrée augmente de $4dB$ pour que celui de sortie augmente de $1dB$. Pour un taux de compression extrême ($R_{comp} = 10$), on parle alors de **limiteur**.

$$N_{sortie} = \begin{cases} N_{entree} & \text{si } N_{sortie} \leq Seuil \\ \frac{1}{R_{comp}}(N_{entree} - Seuil) + Seuil & \text{si } N_{sortie} \geq Seuil \end{cases} \tag{2.5}$$

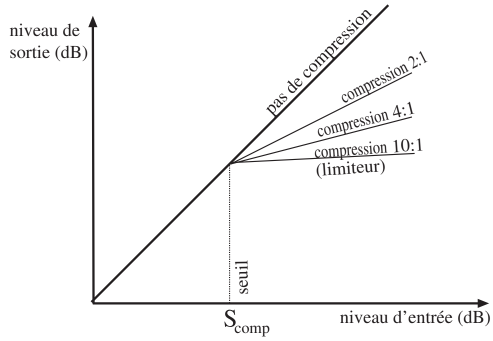

Fig. 2.4: *Relation entrée-sortie du compresseur*

Le niveau d'énergie est mesuré par moyenne sur une portion courte du signal (ex. par RMS), mais peut aussi l'être par pic instantané ou valeur échantillonnée : dans ces deux cas, le compresseur devient un limiteur.

On définit encore une fois un temps d'attaque ($\tau_{a,comp}$), temps nécessaire au compresseur pour répondre à un accroissement de niveau franchissant le seuil, ainsi qu'un temps de relaxation ($\tau_{r,comp}$), correspondant au temps nécessaire au compresseur pour répondre à une diminution de niveau passant en dessous du seuil, temps généralement plus grand que le temps d'attaque.

Selon l'application, il peut être préférable d'avoir un faible temps d'attaque ou un faible temps de relaxation, ce qui requiert une rapide variation de gain : cela sera perçu comme un effet de "respiration" (breathing) ou de "pompage" (pumping). Lorsque le niveau sonore est inférieur au seuil, le gain augmente et passe à 1. Le signal est alors plus proche du niveau de bruit du système, ce qui rend le bruit audible. Un compresseur plus sophistiqué étudiera plus en détail le signal d'entrée et ajustera le gain lorsque le niveau d'entrée atteint momentanément 0, afin de limiter cet effet indésirable de "respiration".

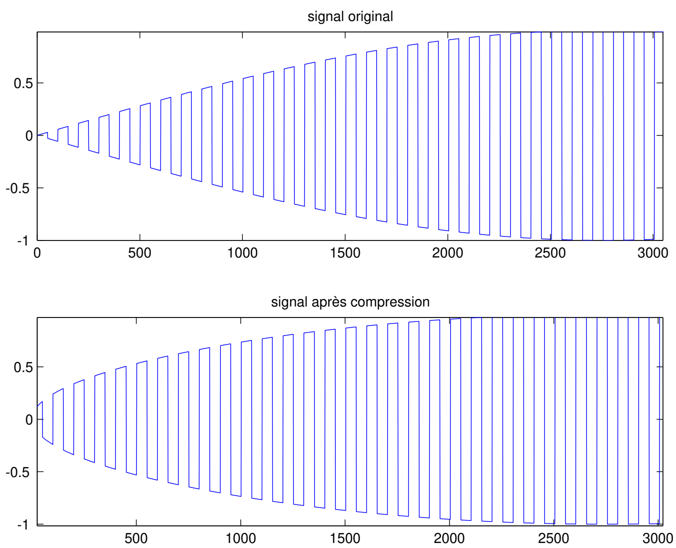

Fig. 2.5: *Signal carré modulé en amplitude avant et après compression et re-normalisation.*

Intérêt de cet effet

Limiter le niveau de certains signaux permet d'éviter qu'une piste magnétique ne "bave" sur les voisines (magnétisation trop forte d'une piste qui magnétise partiellement les particules des pistes voisines) lors d'enregistrement studios. Cela permet aussi d'éviter les phénomènes de distorsions indésirés, du fait que les appareils électroniques analogiques ou numériques recevant les signaux ont un niveau d'entrée maximal (protégeant en même temps le matériel).

En studio, la compression est un outil pratique pour découper des pistes et ajuster des mixages (en diminuant les différences excessives de niveau d'une piste, par exemple). En utilisant un temps d'attaque approprié, on conservera le naturel de l'attaque d'un son instrumental après compression. Dans certains cas, la compression permet de réduire le besoin d'égalisation.

Une autre utilisation du compresseur est l'augmentation de tenue (sustain) d'un instrument. Le compresseur peut maintenir un niveau quasi constant pour le signal de sortie en amplifiant convenablement le signal d'entrée. Une faible compression empêchera le niveau de l'instrument de changer trop radicalement, ce qui est ressenti comme une augmentation de tenue, ou un "lissage" du son. Un plus grand temps de relaxation que le temps de décroissance du son instrumental permet de conserver les propriétés du son de l'instrument après compression.

Compléments d'information

Le limiteur croisé (*ducking/cross limiting*) correspond à un limiteur appliqué à un signal (en l'occurence la musique) et piloté par un autre signal (la voix). Il permet par exemple à

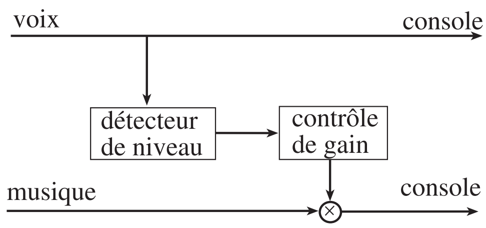

Fig. 2.6: *Diagramme du limiteur croisé*

un animateur ou un Disk Jokey de gérer automatiquement le gain du fond sonore lors des prises de parole. De plus, il permet de donner de l'importance à certains éléments (tels les percussions) lors de l'enregistrement studio, augmentant leur présence en baissant le niveau des autres pistes.

En appliquant le limiteur à certaines bandes fréquentielles, on obtient un *de-esser* (cet effet supprime les "s").

Lorsqu'on utilise le compresseur avec d'autres effets, il est préférable de la placer en premier dans la chaîne, tout d'abord pour des raisons de bruit. Un compresseur en fonctionnement (réduisant la gamme dynamique de la sortie) suivi d'un gain additionnel sur la sortie amplifie le bruit en même temps que le son de l'instrument. D'autres effets peuvent introduire du bruit de fond, que l'on préfère ne pas amplifier.

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser cet effet sont :

- le seuil S_{comp};
- le rapport d'expansion R_{comp} ;
- le temps d'attaque $\tau_{a,comp}$;
- le temps de relaxation $\tau_{r,comp}$.

2.2.5 Trémolo

Définition

Le trémolo correspond à une modulation d'amplitude en dessous de 20 Hz, audible dans le domaine temporel. Il est utilisé en tant que tel comme mode de jeu, et est aussi présent dans le vibrato (volontairement ou non de la part du musicien).

Fonctionnement

Le trémolo s'obtient à l'aide d'un simple modulateur d'amplitude. Pour sonner comme tel, la fréquence du trémolo f_{tr} doit se trouver entre 7 et 14 Hz. Soit $x(n)$ le signal à traiter. Une fois le trémolo appliqué, on obtient le signal modulé :

$$y(n) = \left[1 - d_{tr}\left(1 - \frac{\sin(2\pi f_{tr} n / F_e)}{2}\right)\right].x(n) \tag{2.6}$$

avec F_e la fréquence d'échantillonnage et $d_{tr} \in [0; 1]$ la profondeur.

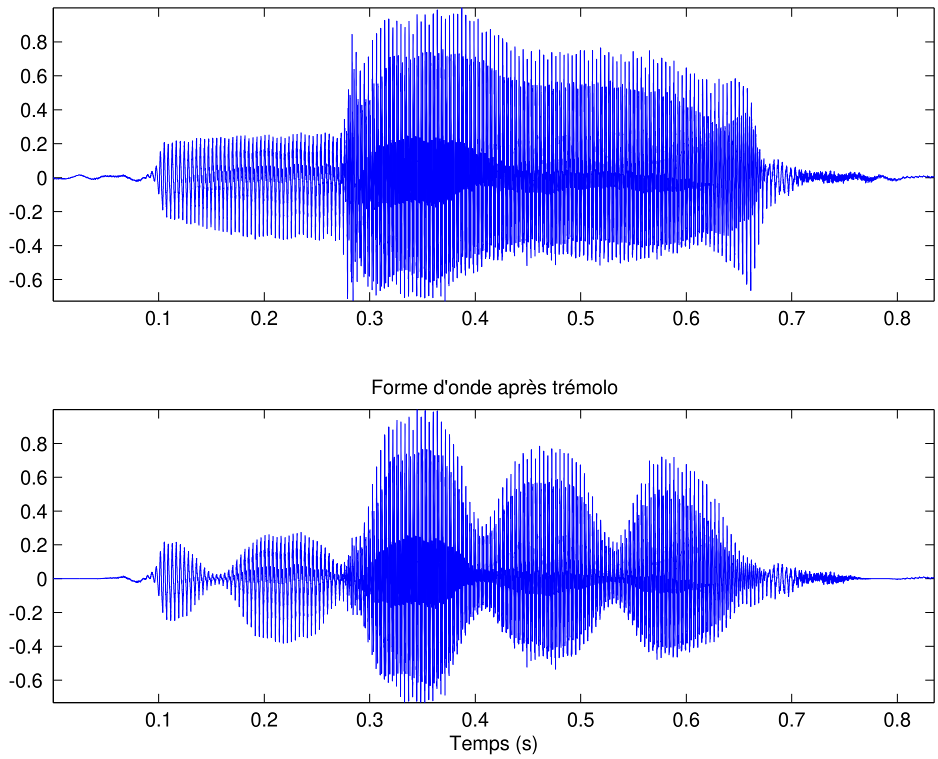

Fig. 2.7: *Enveloppes temporelles d'un son de voix et du même son après application d'un trémolo de fréquence* 6 Hz *et de profondeur 0.7.*

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser cet effet sont :

- la fréquence du trémolo f_{tr} en Hertz ;
- sa profondeur $d_{tr} \in [0; 1]$.

2.3 Traitements modifiant la durée

Le temps et la fréquence sont intrinsèquement liés. Ainsi, prenons une sinusoïde à 440 Hz échantillonnée à 44100 Hz. Sur une période de 1 seconde, on a 44100 valeurs de cette sinusoïde discrétisée, avec 440 périodes durant cette seconde. Si l'on procède à un changement de fréquence d'échantillonnage (par exemple 22500 Hz), on n'aura plus que 220 périodes en une seconde (donc un son à 220 Hz, une octave en dessous), et il faudra deux fois plus de temps pour lire le même signal.

Ce problème de changer l'échelle temporelle indépendamment du contenu fréquentielle est classique en traitement du son. Il est lié à la perception. La somme de deux sinusoïdes peut être vue come le produit d'une porteuse et d'une modulante. Une fois l'étirement temporel effectué, doit-on avoir la somme des deux mêmes sinusoïdes ou la même porteuse multipliée par une modulante différente ? En dessous de 10 Hz, une modulation est entendue comme un trémolo ou un vibrato ; au-dessus, cela donne des fréquences audibles.

Il faut donc mettre en œuvre des méthodes d'analyse-synthèse afin de pouvoir modifier le temps sans modifier les fréquences. L'idée de base est de répéter une portion de son : si je prends un grain de son fenêtré, et que je l'ajoute en sortie avec un décalage temporel, pour peu que les phases soient en continuité, j'aurai un son plus long, de même hauteur. Ce principe peut être appliqué dans le domaine temporel (TD-PSOLA), dans le domaine temps-fréquence (avec le vocodeur de phase) ou dans le domaine fréquentiel (modèle spectral).

2.3.1 Changement d'échelle temporelle linéaire sans conservation de la hauteur

Le changement de durée sans conservation de la hauteur (ni des formants) consiste à modifier l'échantillonnage d'un signal, de façon à modifier la hauteur perçue. Ce traitement ne conserve ni les formants, ni la durée du signal.

Il est expliqué en 2.7, du fait qu'à la fois la durée et la hauteur soient modifiés.

2.3.2 Changement d'échelle temporelle linéaire avec conservation de la hauteur

Définition

Le changement d'échelle temporelle ou *time-scaling* (ou encore ralenti-accéléré) linéaire avec conservation de la hauteur consiste à ralentir ou accélérer un son tout en préservant la hauteur d'origine.

Fonctionnement

La manière de réaliser cet effet consiste à conserver l'information fréquentielle lorsque l'étirement temporel est réalisé. Pour ce faire, plusieurs méthodes existent.

Les traitements segment-temporels (type SOLA, PSOLA) répètent ou suppriment des segments d'échantillons, de la taille d'une période fondamentale, afin de modifier la durée. Ceci fonctionne particulièrement bien pour des signaux harmoniques et surtout la partie stable.

Les traitements spectraux utilisent l'information de module et de fréquence des partiels analysés pour synthétiser un son de la durée requise, en interpolant les valeurs des modules et fréquences de synthèse à partir des valeurs d'analyse. Le bruit quant à lui est resynthétisé par filtrage, l'enveloppe spectrale du filtre étant déduite par interpolation des enveloppes spectrales d'analyse.

Les traitements temps-fréquence (vocodeur de phase) synthétisent le son en utilisant eux aussi un pas d'analyse différent du pas de synthèse, et en recalculant les phases à chaque itération de synthèse, de façon à ne pas provoquer de rupture de phase.

En effet, changer les valeurs d'une transformée de Fourier de manière arbitraire n'est pas une opération valide (la Tranformée obtenue n'est pas forcément l'image d'un signal réel) : le son de synthèse peut alors être différent de ce que à quoi on s'attendrait. Un bon algorithme de transformation consiste à trouver une stratégie qui préserve l'aspect de changement d'échelle temporelle sans introduire trop d'artefacts.

Pour le vocodeur de phase par bloc FFT/IFFT, l'idée est donc de conserver la magnitude et de modifier la phase de façon à préserver la fréquence instantanée f_i. La différence de phase du panier de fréquence k est dont $\Delta\psi(k) = \frac{R_s}{R_a}\Delta\phi(k)$, avec R_a le pas d'analyse, R_s le pas de synthèse, d'où $\frac{R_s}{R_a}$ le facteur de changement d'échelle temporelle.

La différence de phase à l'analyse se calcule comme suit :

$$\Delta\phi(k) = \frac{2\pi R_a(k-1)}{N} + \text{princarg}\left(\tilde{\phi}\left((s+1)R_s, k\right) - \tilde{\phi}\left(s\ R_s, k\right) - \frac{2\pi R_a(k-1)}{N}\right) \qquad (2.7)$$

d'où l'on obtient la nouvelle phase de synthèse :

$$\psi\left((s+1)R_s, k\right) = \text{princarg}\left(\psi\left(sR_s, k\right) + \Delta\phi(k)\frac{R_s}{R_a}\right) \qquad (2.8)$$

Intérêt de cet effet

Cet effet permet de modifier la durée d'un son, en donnant un expressivité toute autre que celle de départ. Il permet aussi de synchroniser deux signaux de durées initiales différentes ; il est aussi utilisé en post-production cinéma-vidéo, du fait que les formats sont différents (24 images/s pour l'un, 25 images/s pour l'autre).

Compléments d'information

Plusieurs remarques quant à la mise en œuvre du changement d'échelle temporelle par vocodeur de phase sont à prendre en compte. Tout d'abord, le changement de phases correspond à l'application d'un filtre passe-tout. Il faut appliquer une fenêtre de resynthèse, sinon un effet de convolution circulaire (dûe au filtrage) implique des discontinuités aux extrémités des fenêtres. Ensuite, même avec l'application d'une fenêtre de synthèse, il reste un aspect de convolution circulaire (c'est la version repliée d'une FFT infinie) : il est nécessaire de procéder à un bourrage de 0 à l'analyse comme à la synthèse.

Pour s'assurer d'une reconstruction parfait lorsque les pas et les fenêtres d'analyse et de synthèse sont identiques, la somme des carrés des fenêtres (soit leur puissance) espacées d'un pas doit toujours être de 1, pour chaque échantillon (on doit s'assurer de ne pas ajouter de modulation d'amplitude avec l'utilisation de fenêtres).

Par exemple, pour une fenêtre de Hanning sans bourrage de 0, le pas de synthèse R_s doit être un diviseur de $\frac{N}{4}$.

Pour les fenêtres de Hamming et Blackman, les lobes secondaires sont plus bas que pour la fenêtre de Hanning, mais leur valeur aux extrémités est non nulle, ce qui implique des problèmes aux extrémités si on les utilise seules. Il faut un pas de synthèse R_s diviseur de $\frac{N}{8}$.

Pour une fenêtre de Gauss tronquée, la somme des fenêtres décalées d'un pas produit toujours des oscillations (modulations d'ampitude), mais qui peuvent être rendues en dessous du seuil de perception avec un pas bien choisi.

Remarquons enfin que lorsque le pas de synthèse est multiple entier du pas d'analyse : $\frac{R_s}{R_a} = k \in \mathbb{N}$, il n'est plus nécessaire de calculer le déroulement de phase, puisque la relation modulo 2π est conservée. Par contre, lorsque ce ratio est réel, il résulte un problème de dispersion des phases. En effet, toutes les phases sont forcées de tourner à la vitesse à laquelle elles tourneraient si une sinusoïde de fréquence égale à la fréquence centrale du panier de fréquence existait dans le signal. En réalité, seules les sinusoïdes présentes dans le signal devraient être modélisées par des phases tournant entre deux FFT consécutives[1]. C'est ce que réalisent Dolson et Laroche, en ne déroulant les phases que des paniers contenant des pics ainsi que leurs voisins (un pic correspondant à un maximum local par rapport à ses 2 ou 3 voisins). Le déroulement de phase se fait alors à la même fréquence pour tous les paniers du pic (ie. ses voisins aussi).

Rappel des paramètres de l'effet

Le paramètre permettant de caractériser cet effet est le ratio de contraction/dilatation γ_0 (facteur de changement d'échelle).

2.3.3 Changement d'échelle temporelle non linéaire (adaptatif) avec conservation de la hauteur

Définition

L'étirement temporel non linéaire est un changement d'échelle dont le facteur d'étirement varie au cours du temps. Le changement d'échelle temporelle **adaptatif** (ou sélectif) est un étirement temporel dont le facteur d'étirement dépend d'un descripteur extrait du son à traiter. Ainsi, le facteur d'étirement est une fonction du ou des descripteurs $\mathcal{D}_i(n)$ utilisés :

$$\gamma(n) = f\left(\mathcal{D}_1(n), \ldots, \mathcal{D}_I(n)\right) \tag{2.9}$$

Fonctionnement

Un étape d'extraction de paramètre doit être ajoutée à l'algorithme de changement d'échelle temporelle présenté en 2.3.2. Ce paramètre est transformé de façon à obtenir une valeur variant entre les bornes définies par l'utilisateur.

[1] les autres paniers de fréquence devraient avoir un déroulement de phase aléatoire, puisqu'ils correspondent à du bruit

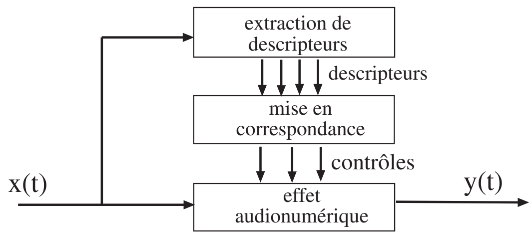

Fig. 2.8: *Diagramme de l'effet audionumérique adaptatif*

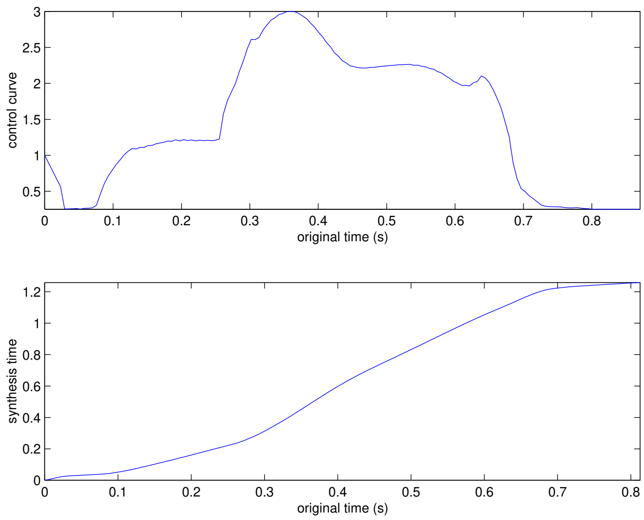

Fig. 2.9: *Courbe de contrôle et Diagramme du temps du grain de synthèse en fonction du temps du grain d'analyse : une contraction temporelle correspond à une pente supérieure à 1 et un étirement (dilatation) à une pente inférieure à 1.*

Intérêt de cet effet

Par exemple, en utilisant l'indice de voisement, on peut créer le ralenti sélectif, qui ne ralenti que les voyelles, et laisse les consonnes d'une voix parlée ou chantée intelligible. Un autre exemple consiste à prendre un facteur d'étirement allant d'une valeur inférieure à 1 à une autre supérieure à 1. Cela permet alors de ralenti ou accélérer la lecture du son de départ. D'une manière générale, cela permet de ré-interpréter une phrase musicale.

Compléments d'information

Les paramètres permettant de caractériser cet effet sont :

- les descripteurs $\mathcal{D}_i, i = 1, \ldots, I$ utilisés ;
- la fonction de transformation des descripteurs en valeur de contrôle f ;
- les bornes données à cette fonction de transformation, à savoir le facteur de contraction maximal γ_- et le facteur de dilatation maximal γ_+.

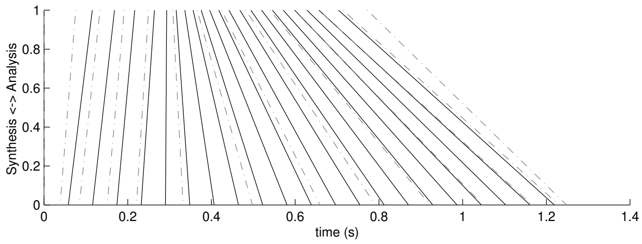

Fig. 2.10: *Représentation des marques d'analyse (en haut) et de synthèse (en bas) ainsi que leurs correspondances, à pas d'analyse constant (traits pointillés) et à pas de synthèse constant (traits pleins).*

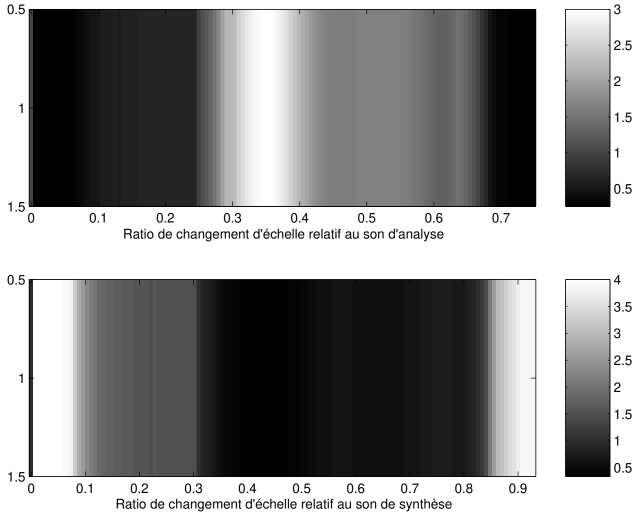

Fig. 2.11: *Ratios d'étirement avec pour référence le son original (fig supérieure) et le son de synthèse (fig. inférieure)*

2.3.4 Ralenti-accéléré non linéaire avec conservation de la hauteur et de la durée globale

Définition

Un changement d'échelle temporelle non linéaire avec conservation de la hauteur et de la durée consiste à appliquer un changement local de durée tout en assurant que le son de synthèse, dans sa globalité, sera de même durée que le son d'analyse. C'est un raffinement du changement d'échelle temporelle non linéaire avec conservation de la hauteur, présentée précédemment en 2.3.3.

Fonctionnement

Une fois l'indice de contraction/dilatation $\gamma(l)$, $l = 1, \ldots, NT$ calculé, on lui applique une transformation de façon à ce que la durée du son obtenue soit la même que le son de départ. Deux schémas peuvent être utilisés, l'un avec pas d'analyse constant, l'autre avec pas de synthèse constant :

Pas de synthèse Δt_S constant Le temps d'analyse est donné par :

$$T_A^1(k) = T_A^1(k-1) + \tilde{\gamma}\left(T_S^1(k-1)\right)\Delta t_S \tag{2.10}$$

avec $\tilde{\gamma}$ l'interpolation linéaire de γ, et le temps de synthèse par :

$$T_S^1(k) = T_S^1(k-1) + \Delta t_S \tag{2.11}$$

Notons $t_0 = T_A^1(0)$, on obtient la formulation récurrente :

$$T_A^1(k) = t_0 + \Delta t_S \sum_{l=1}^{k} \tilde{\gamma}\left(t_0 + l\Delta t_S\right) \tag{2.12}$$

$$T_S^1(k) = t_0 + k\,\Delta t_S \tag{2.13}$$

Pas d'analyse Δt_A constant Le temps d'analyse est donné par :

$$T_A^2(k) = T_A^2(k-1) + \Delta t_A \tag{2.14}$$

et le temps de synthèse par :

$$T_S^2(k) = T_S^2(k-1) + \tilde{\gamma}\left(T_A^2(k-1)\right)\Delta t_A \tag{2.15}$$

En remarquant que $t_0 = T_S^2(0)$, on obtient la formulation récurrente :

$$T_A^2(k) = t_0 + k\,\Delta t_A \tag{2.16}$$

$$T_S^2(k) = t_0 + \Delta t_A \sum_{l=1}^{k} \tilde{\gamma}\left(t_0 + l\Delta t_A\right) \tag{2.17}$$

On considére que les deux schémas sont à peu près équivalents, en tout cas suffisamment proches l'un de l'autre pour les considérer équivalents.

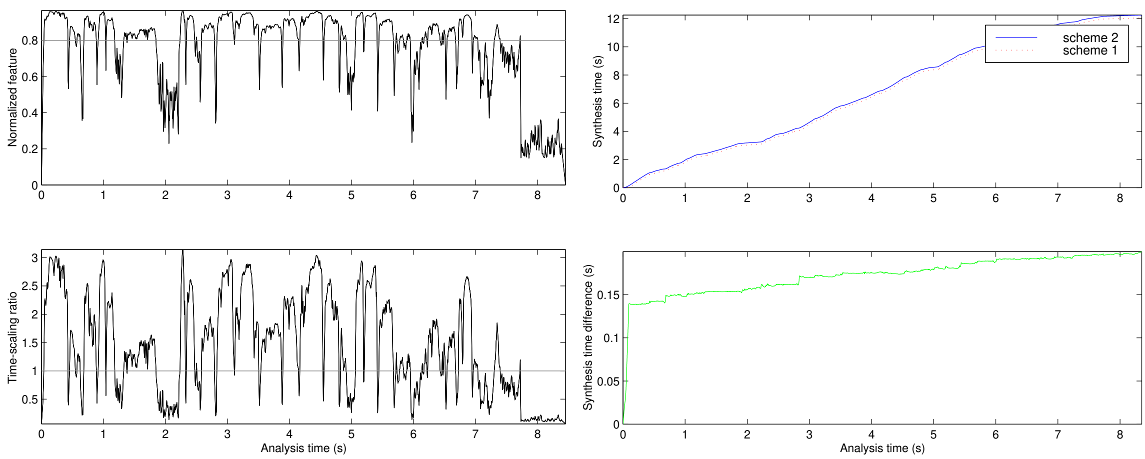

Fig. 2.12: *Gauche-haut : descripteur normalisé ; gauche-bas : ratio de contraction/dilatation utilisé ; droite-haut : les deux schémas pour un son de 8.5 s ; droite-bas : différence de temps de synthèse entre les deux schémas.*

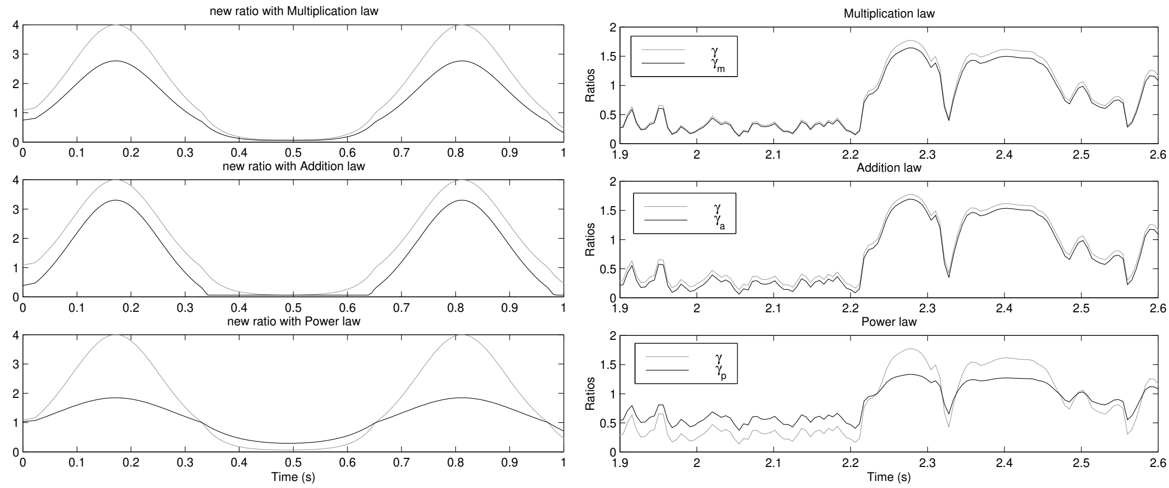

Fig. 2.13: *Modification du ratio de contraction/dilatation par addition, multiplication, puissance.*

Remarquons que la condition de durée égale du son d'analyse et du son de synthèse pour le schéma i est $t_s = T_A^i(s) = T_S^i(s)$ avec t_s le temps de synchronisation et s l'indice auquel ce temps est atteint. Pour le schéma 1, cette condition peut encore s'exprimer :

$$\sum_{l=1}^{s} \tilde{\gamma}(t_0 + l\ \Delta t_S) = s \tag{2.18}$$

Pour le schéma 2, elle s'écrit :

$$\sum_{l=1}^{s} \tilde{\gamma}(t_0 + l\ \Delta t_A) = s \tag{2.19}$$

En prenant soin d'utiliser le même pas pour l'analyse dans le schéma 2 et la synthèse dans le schéma 1 ($\Delta t_U = \Delta t_A = \Delta t_S$), on obtient une condition unique pour les deux schémas :

$$\sum_{l=1}^{s} \tilde{\gamma}(t_0 + l\ \Delta t_U) = s \tag{2.20}$$

Nous proposons trois façons de modifier la courbe γ : par addition, par multiplication et par puissance.

Synchronisation par translation des valeurs de γ Etant donné l'intervalle de variation de γ (imposé par l'utilisateur)$I_\gamma = [\gamma_{min}; \gamma_{max}] = [\frac{1}{M}; N]$, on désire trouver γ_a un facteur additif tel que la courbe $\Gamma = \gamma + \gamma_a$ vérifie la condition (2.20). Si l'on ne tient pas compte de I_γ, l'intervalle de variation de γ, le calcul est direct et donne :

$$\gamma_a = 1 - \frac{\sum_{l=1}^{s} \gamma(l)}{s} \tag{2.21}$$

Cependant, on prendrait ainsi le risque d'avoir soit un facteur de dilatation trop grand (ce qui peut être accepté), soit un facteur de contraction négatif (ce qui, par contre, ne peut être accepté, sinon cela signifierait de lire le son à l'envers).

Pour éviter ces désagrements, on utilise une méthode itérative de recherche de γ_a, en imposant $\Gamma(l) \in I_\gamma,\ l = 1, \ldots, s$.

Synchronisation par multiplication des valeurs de γ On désire trouver γ_m un facteur multiplicatif (ou additif en échelle logarithmique) tel que la courbe $\Gamma = \gamma \times \gamma_m$ vérifie la condition (2.20). Si l'on ne tient pas compte de I_γ, l'intervalle de variation de γ, le calcul est direct et donne :

$$\gamma_m = \frac{s}{\sum_{l=1}^{s} \gamma(l)} \tag{2.22}$$

Cependant, les risques sont identiques à ceux existant dans le cas de la translation. Pour éviter ces désagrements, on utilise une méthode itérative de recherche de γ_m, en imposant $\Gamma(l) \in I_\gamma,\ l = 1, \ldots, s$.

Synchronisation par puissance des valeurs de γ On désire trouver γ_p un facteur exponentiel (ou multiplicatif en logarithmique) tel que la courbe $\Gamma = \gamma^{\gamma_p}$ vérifie la condition (2.20). Si l'on ne tient pas compte de I_γ, le calcul n'est pas direct et nécessite une méthode itérative. Les risques sont identiques à ceux existant dans le cas de l'addition et de la multiplication. Pour les éviter, on impose encore une fois $\Gamma(l) \in I_\gamma,\ l = 1, \ldots, s$.

Intérêt de cet effet

En utilisant cette notion de synchronisation ou de conservation de durée, on obtient des sons de même durée, mais d'expressivité différentes. Cela permet aussi de réaliser des changements de prosodie (cf. 2.7.4), d'expressivité (cf. 2.7.5) poussés.

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser cet effet sont :

- les descripteurs $\mathcal{D}_i, i = 1, \ldots, I$ utilisés ;
- la fonction de transformation des descripteurs en valeur de contrôle f ;
- les bornes données à cette fonction de transformation, à savoir le facteur de contraction maximal γ_- et le facteur de dilatation maximal γ_+ ;
- la fonction de synchronisation (par addition, multiplication ou puissance) et ses paramètres (facteur additif γ_a, multiplicatif γ_m ou puissance γ_p).

2.4 Traitements modifiant la hauteur

Les différents traitements existant sur la hauteur sont basés sur la transposition (avec ou sans conservation des formants).

2.4.1 Transposition sans conservation de la durée ni des formants

La transposition sans conservation de la durée ni des formants consiste à modifier l'échantillonnage d'un signal, de façon à modifier la hauteur perçue. Ce traitement ne conserve ni les formants, ni la durée du signal. C'est la transformation duale du changement d'échelle temporelle sans conservation de la hauteur. Elle est expliquée en 2.7.

2.4.2 Transposition sans conservation des formants

Définition

La transposition sans conservation des formants consiste à modifier la hauteur de manière brute, sans extraire l'enveloppe spectrale. Il s'agit avant tout d'un artefact de méthodes de changement de hauteur, baséess ur le ré-échantillonnage. Cet effet conduit facilement à des effets de type “Donald Duck”.

Fonctionnement

La hauteur d'un signal étant donnée par sa répartition en fréquences (un signal harmonique composé d'harmoniques, partiels de fréquences multiples de la fréquence fondamentale), en changer la hauteur se fait en multipliant toutes les fréquences par un même ratio γ.

L'utilisation d'un modèle spectral (SMS, SAS) permet aisément cette manipulation. Le vocodeur de phase par bloc quant à lui nécessite une mise en œuvre plus complexe. Le changement de hauteur est en lien direct avec l'algorithme de changement de durée avec le vocodeur de phase : si l'on applique un changement de durée d'un facteur $\gamma_{duree} = \frac{R_s}{R_a}$ puis un ré-échantillonnage avec le ratio inverse $\gamma_{hauteur} = \frac{1}{\gamma_{duree}}$, on obtient un son de même durée que l'original, mais dont la hauteur a changé du fait du ré-échantillonnage.

Soit $L_i = \frac{R_s}{R_a} L_a$ la durée du signal rallongé, la durée du signal de synthèse est, après ré-échantillonnage, $L_s = \frac{R_s}{R_a}\, L_i = \frac{R_a}{R_s}\frac{R_s}{R_a}\, L_a = L_a$.

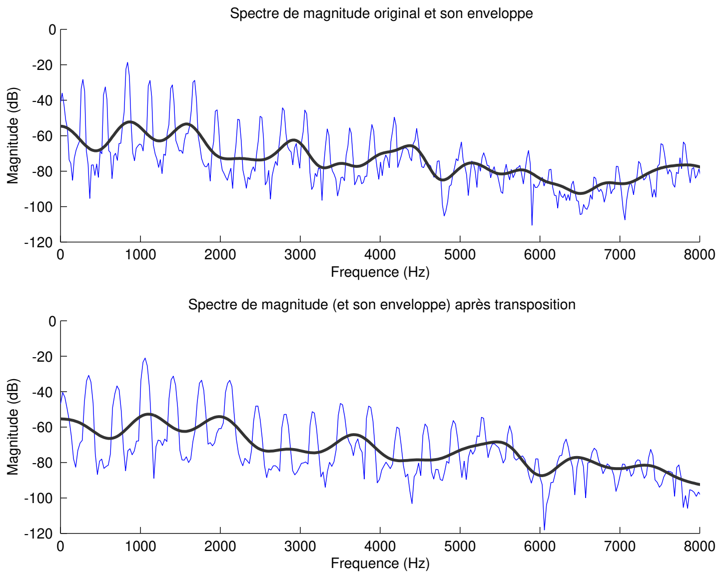

Fig. 2.14: *Sonagramme d'un son voisé avant et après transposition sans conservation des formants de hauteur (l'enveloppe spectrale est dilatée).*

Intérêt de cet effet

Cet effet permet la création de nouveaux sons, puisque les formants ne sont pas respectés (ainsi, l'exemple de l'effet “Donald Duck”).

Compléments d'information

Cet effet peut s'obtenir par des traitements temps-fréquence : par banc de filtres, bloc-par-bloc. De même on peut l'appliquer à l'aide de traitements segment-temporels : par contraction-étirement temporel et ré-échantillonnage (comme expliqué précédemment), par lecture à vitesse variée (avec 2 ou 4 têtes de lectures tournantes), par modulation de longueur d'une ligne à retard.

Rappel des paramètres de l'effet

Le paramètre permettant de caractériser cet effet est le ratio de transposition γ.

2.4.3 Transposition avec conservation des formants

Définition

La transposition avec conservation des formants consiste à appliquer un changement de hauteur tout en gardant l'information formantique du signal. Cette transposition est en générale plus proche de la transposition effectuée par un instrument de musique que le traitement ne préservant pas les formants.

Fonctionnement

De même qu'on ne peut directement modifier la hauteur d'un signal sans en modifier la durée, on ne peut modifier la hauteur d'un signal sans affecter les formants qu'à condition de mettre en œuvre des techniques particulières. Aussi, il fait d'abord calculer l'enveloppe spectrale du son, effectuer la transposition, puis corriger les amplitudes des pics fréquentiels en fonction de l'enveloppe spectrale d'origine.

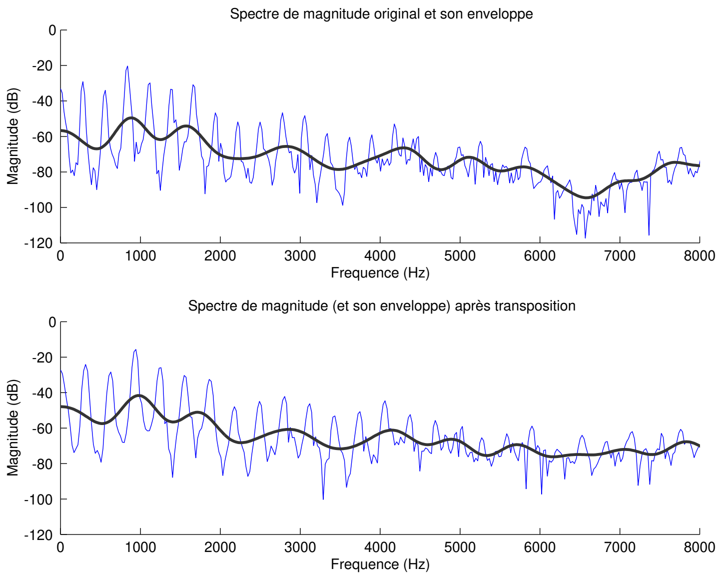

Fig. 2.15: *Sonagramme d'un son voisé avant et après transposition avec conservation des formants de hauteur (l'enveloppe spectrale est conservée).*

Intérêt de cet effet

Cet effet permet par exemple de transposer une voix humaine sans effet "Donald Duck", ou de transposer un instrument en conservant mieux le timbre.

Compléments d'information

Cet effet peut s'obtenir par les trois classes de méthodes suivantes : les traitements source-filtre (lpc, cepstre, etc), les traitements segment-temporels (par PSOLA) et les traitements spectraux.

Rappel des paramètres de l'effet

Le paramètre permettant de caractériser cet effet est le ratio de transposition γ.

2.4.4 Discrétisation sur une échelle tempérée

Définition

La discrétisation sur une échelle tempérée consiste à modifier la hauteur d'un signal de telle sorte que la fréquence fondamentale appartienne toujours à la gamme tempérée. Ceci revient en quelque sortes à accorder le signal sur une échelle prédéterminée.

Fonctionnement

Il s'agit donc dans un premier temps de calculer la hauteur (à défaut la fréquence fondamentale) du signal, puis de la modifier. Dans une certaine mesure, on pourra considérer que des écarts de hauteur entre la hauteur mesurée et la plus proche hauteur de l'échelle tempérée supérieurs à une tolérance impliqueront la modification, tandis qu'un écart inférieur à la tolérance ne sera pas modifié.

Intérêt de cet effet

Cet effet permet d'accorder le signal sur une échelle pré-établie, et par exemple de mixer ensemble des pistes qui n'étaient auparavant pas accordées, sans désagréments audibles.

2.4.5 Harmoniseur

Définition

Un harmoniseur est un effet qui ajoute à un signal sa version transposée. Selon le nombre de transpositions ajoutées et leur facteur de transposition, on obtient des accords.

Fonctionnement

Le principe est simple : une fois la règle d'harmonisation définie (par exemple, pour obtenir un accord parfait majeur, il faut transposer d'une tierce majeure et d'une quinte), les versions transposées $T_{\gamma_i}(x)(n)$ du signal sont calculées (avec γ_i le ratio de transposition), puis ajoutées au signal de départ $x(n)$.

Intérêt de cet effet

Cet effet permet par exemple de remplacer une voix monophonique par un choeur harmonisé, et donc d'enjoliver des mélodies.

Compléments d'information

Le changement de hauteur utilisé est un changement de hauteur avec conservation de la durée et des formants. Il peut utiliser une méthode segment-temporelle, temps-fréquence ou spectrale.

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser cet effet sont :

- le nombre de voix N_{harmo} ;
- les ratios de transposition $\gamma_i, i = 1, \ldots, N_{harmo}$.

2.4.6 Vibrato

Définition

Le vibrato est une lente modulation de fréquence, généralement comprise entre 4 et 8 Hz. C'est une spécificité de la voix chantée, un de ses attributs perceptifs. En effet, un vibrato est perçu comme un mode de jeu et non comme une modulation de fréquence : le système auditif le traduit comme attribut perceptif. La plupart des instruments utilisent le vibrato dans leurs modes de jeu (vents, cuivres, cordes frottées ou pincées comme la guitare). Il s'accompagne généralement d'un trémolo (modulation d'amplitude) de même fréquence et de profondeur de l'ordre de 3 dB.

Fonctionnement

Quelle que soit la méthode de changement de hauteur utilisée (PSOLA, vocodeur de phase, modèle spectral, avec ou sans conservation des formants), le facteur de changement de hauteur est donné par :

$$\gamma_{hauteur}(n) = 1 + d_{vib}.\sin(2\pi\, f_{vib} n) \tag{2.23}$$

avec d_{vib} la profondeur du vibrato, de l'ordre de 10 cent (soit 10/200 ton), et f_{vib} la fréquence de la modulation.

Compléments d'information

Il a été montré que le vibrato, en produisant un balayage de l'enveloppe spectrale, permet au système auditif de reconnaître plus facilement les formants d'un son. Aussi, il est préférable d'utiliser une méthode de changement de hauteur qui conserve les formants, de manière à reproduire fidèlement cette mise en relief, ce parcours de l'enveloppe spectrale.

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser cet effet sont :

- d_{vib} la profondeur du vibrato ;
- f_{vib} la fréquence de la modulation.

2.5 Traitements modifiant la localisation, la spatialisation

Nous abordons maintenant les effets portant sur les notions de localisation, de spatialisation, de ditance entre l'auditeur et la source, les effets produits par la salle, par les mouvements de l'auditeur ou de la source.

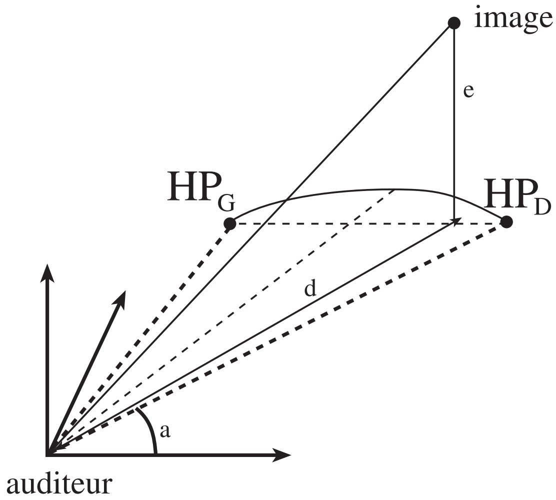

Fig. 2.16: *Localisation par un rendu 3D d'un son : azimuth (a), distance (d) et élévation (e).*

2.5.1 Écho

Définition

L'écho (ou délai) est l'un des effets les plus simples à réaliser. Il s'agit d'une ou plusieurs répliques du son d'entrée décalées dans le temps.

Fonctionnement

Le délai simple prend un signal en entrée et le décale dans le temps d'une durée τ_{delay}. Ce temps de retard peut varier entre quelques millisecondes et quelques secondes. La figure 2.17 présente le diagramme de cet écho simple.

Un délai simple étant un effet limité, des systèmes de délai plus complexes sont réalisés avec une boucle de retour ou de réinjection (*feedback*/régénération), cf. fig.2.18. La sortie du premier écho est mixée avec l'entrée, et pourvu que le gain soit inférieur à 1 (pour des raisons de stabilité de l'appareil : dans le cas contraire, l'entrée sature très vite), on peut répéter durablement un son, dont le niveau baisse progressivement au fur et à mesure de son passage dans la boucle d'écho. Théoriquement, le son devrait se répéter indéfiniment avec la

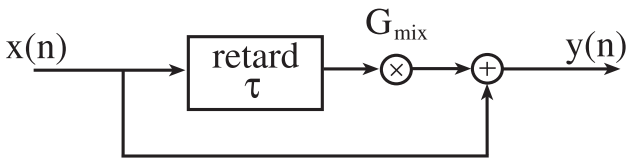

Fig. 2.17: *Diagramme du délai simple.*

réinjection, mais il passe vite en dessous du niveau de bruit ambiant, et devient inaudible de ce fait.

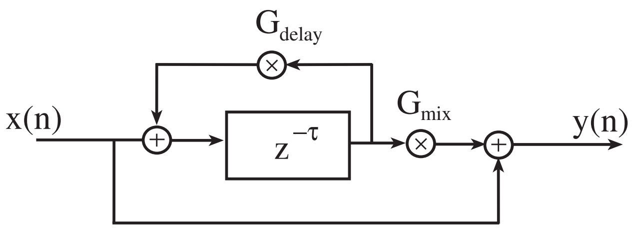

Fig. 2.18: *Diagramme du délai avec réinjection (feedback).*

Intérêt de cet effet

Les délais sont très utilisés pour enrichir un son instrumental. En utilisant de faibles temps de retard (50 à 100 millisecondes), il se crée un effet de doublage, comme si deux instruments jouaient à l'unisson.

L'utilisation de plusieurs délais avec réinjection peut être utilisé pour créer un effet de réverbération, bien qu'une unité de réverbération classique crée un son plus complexe.

Pour des temps de retard supérieurs à 100 millisecondes, l'effet est moins subtile, la répétition s'entend comme telle. Aussi, une possibilité intéressante consiste à faire correspondre ce temps au tempo, de façon à ce que les répétitions tombent exactement sur l'un des battements suivants.

Les systèmes de bouclage et d'échantillonnage découlent de ceci, en prenant une portions de mélodie ou de progression harmonique ou mélodique et en la répétant indéfiniment (techno). Cela permet aussi d'improviser sur ses propres grilles, par exemple (jazz). Plusieurs autres styles de musiques populaires sont construites sur ce modèle.

Les délais sont aussi d'un grand intérêt pour le mixage d'enregistrements stéréophoniques. Cela grandit le son du mixage, et permet une meilleure sensation de placement des instruments (en modifiant à la fois le gain de chaque piste et le délai, qui correspond au temps d'arrivée du son à chaque oreille de l'auditeur) dans un espace à deux dimensions (cf. 2.5.4). Un délai de 20 millisecondes peut donner une grande différence au mixage final.

Compléments d'information

L'implémentation numérique d'un délai se fait à l'aide d'une mémoire tampon circulaire (circular buffer), dont la longueur est déterminée par le temps de retard τ_{delay}. Ainsi, pour un retard d'une seconde de son échantillonné à $44100Hz$, il faut un buffer de 44100 valeurs. Dans le cas où le temps de retard ne correspond pas à un nombre d'échantillons entier, on utilise des lignes à retard fractionnaire (cf. 1.4.3).

D'autres types de délais, plus perfectionnés mais basés sur ces principes, existent :

Slapback : c'est le nom du délai simple à temps de retard très court (40 à 120 ms).

Multi-Tap delay : il s'agit de plusieurs délais avec réinjection (cf. fig.2.19 pour un exemple à deux répétitions), combinés de façon à pouvoir créer des motif rythmiques complexes.

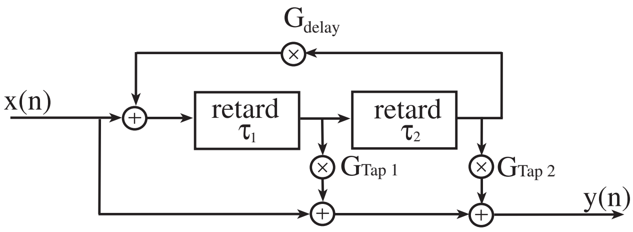

Fig. 2.19: *Diagramme du multi-délai.*

Ping-pong delay : cet effet concerne un signal stéréophonique, dont les deux canaux peuvent être identiques. Les échos de chaque canal sont envoyés dans l'autre canal, et ceci avec une réinjection. Cela crée un effet de rebond.

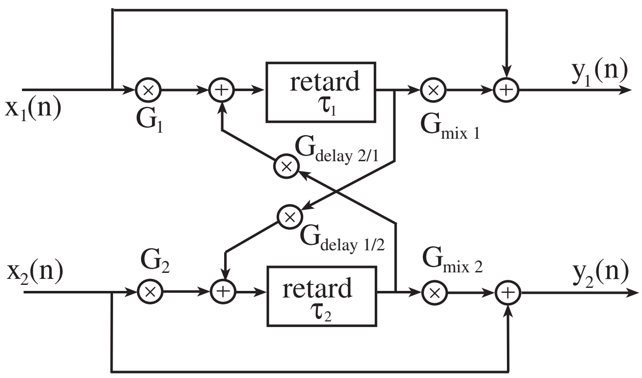

Fig. 2.20: *Diagramme du délai ping-pong.*

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser le délai simple sont :

- τ_{delay} le temps de retard ;
- G_{mix} le gain de mixage, entre 0 et 1;
- G_{delay} le gain de la réinjection, entre 0 et 1.

2.5.2 Panoramisation

Définition

La panoramisation consiste à modéliser avec un système à N haut-parleurs l'azimuth d'un source sonore. Dans le cas de deux haut-parleurs par exemple, il s'agit de pouvoir produire un signal semblant provenir de n'importe quel point entre les deux haut-parleurs. Dans un système à N haut-parleurs situés autour des auditeurs, le positionnement du son en azimuth se fait à partir de deux haut-parleurs seulement. L'auditeur est situé placé au centre du cercle de haut parleurs, ou du moins à égale distance des deux hauts-parleurs (en stéréophonie). La méthode de traitement utilisée est une modulation d'amplitude.

Fonctionnement

Soit deux haut-parleurs fomant un angle θ_{max} avec l'auditeur. La **panoramisation linéaire** est la plus simple ; les amplitudes correctrices du signal dans chaque haut parleur sont données par :

$$\alpha_G = \frac{\theta}{\theta_{max}} \tag{2.24}$$

$$\alpha_D = 1 - (\theta - \theta_{max}) \tag{2.25}$$

$$\tag{2.26}$$

Le grand défaut de cette panoramisation est qu'elle crée un "trou" au milieu, entre les deux haut-parleurs. Le système auditif code un signal plus fort aux extrémités (ie. aux haut-parleurs) qu'au centre. La loi d'intensité sonore établit que l'intensité perçue (ou sonie) d'un son est proportionnelle à son intensité (approximation). Au milieu de la panoramisation l'intensité vaut $I = \sqrt{\alpha_G^2 + \alpha_D^2} = \sqrt{0.5^2 + 0.5^2} = 0.707$. Cette chute d'intensité correspond à une chute de 3 dB, ce qui correspond environ à un changement de nuance en musique (de mf à f par exemple, sur l'échelle *ppp*, *pp*, *p*, *mp*, mf, f, ff, fff). La sensibilité de l'oreille est davantage liée à l'intensité : le son semble être plus faible au centre, comme s'il était éloigné de l'auditeur.

La **panoramisation à puissance constante** utilise des courbes sinusoïdales pour contrôler l'amplitude du signal émis par chaque enceinte. Ceci crée l'impression d'une panoramisation ayant une sonie plus stable :

$$\alpha_G = \frac{\sqrt{2}}{2}(\cos\theta + \sin\theta) \tag{2.27}$$

$$\alpha_D = \frac{\sqrt{2}}{2}(\cos\theta - \sin\theta) \tag{2.28}$$

Elle est adaptée à une situation où les enceintes sont placées avec un angle de $\pi/2$ radians, $\pi/4$ radians de part et d'autre de l'auditeur. L'intensité est donnée par :

$$\begin{aligned}
I &= \sqrt{\alpha_G^2 + \alpha_D^2} && (2.29)\\
&= \sqrt{\frac{1}{2}\left((\cos\theta + \sin\theta)^2 + (\cos\theta - \sin\theta)^2\right)} && (2.30)\\
&= \sqrt{\frac{1}{2}\left(2\cos^2\theta + 2\sin^2\right)} && (2.31)\\
&= 1 && (2.32)\\
& && (2.33)
\end{aligned}$$

L'intensité est donc bien la même pour tout angle $\theta \in [-\frac{\pi}{4}; \frac{\pi}{4}]$.

Intérêt de cet effet

La panoramisation permet de modifier l'azimuth d'un son dans les systèmes multi-haut-parleurs.

2.5.3 Réverbération

Définition

La réverbération d'un son est la somme des réflexions du son original sur les parois d'un espace clos. Il s'agit d'un filtrage, d'une coloration donnée par un lieu sur les sons. C'est un indice perceptif fort, qui nous permet de reconnaître le type de lieu (autant la taille que les types de surfaces, réfléchissantes, diffusantes ou absorbantes) qui recouvrent les parois.

La réverbération se décompose en trois constituants : le son direct, les réflexions primaires, correspondant à des échos très courts, en nombre fini, et les réflexions secondaires, correspondant aux réflexions plus tardives, et de ce fait plus filtrées par la salle.

Fonctionnement

La manière de simuler une réverbération consiste à modéliser les trois constituants, soit distinctement, soit à l'aide d'un dispositif qui reproduit de facto ces constituants. On dénote quatre approches : par ligne à retard, par ensemble de filtres passe-tout / filtres en peigne, par réseau de lignes à retard avec réinjection, et enfin par convolution avec une réponse impulsionnelle de salle. Chacune de ces techniques demandant à elle seule plusieurs pages de développement, nous n'entrerons pas dans les détails, mais donnerons des pistes pour en comprendre la méthodologie.

Réverbération par systèmes de filtres et lignes à retard Les premiers systèmes de réverbération numérique, réalisés par Schroeder aux laboratoires Bell dans les années soixante, utilisaient un filtre passe-tout basé sur une ligne à retard récursive (ou à réinjection). Ce filtre permet une réponse impulsionnelle dense et une réponse en fréquence plate. Jusqu'au début des années 80, cette structure simple, efficace et peu onéreuse en temps de calculs a été très largement utilisée. On considère souvent qu'un filtre passe-tout ne donne aucune coloration au son : ceci n'est vrai (d'un point de vue perceptif) que lorsque la ligne à retard est de longueur inférieure au temps d'intégration de l'oreille, à savoir environ 50 ms. Cette limitation explique la recherche d'autre modèles par la suite.

Dans les années soixante-dix, Gerzon a généralisé le filtre passe-tout à une entrée et une sortie, pour obtenir une structure de type réseau à N entrées et N sorties. L'idée était d'augmenter la complexité de la réponse impulsionnelle sans introduire de coloration notable.

Fin des années soixante-dix, Moorer développa les études de Schroeder en mettant en lien les structures de calcul (lignes à retard, filtre en peigne et filtre passe-tout) avec les propriétés physiques des salles. Cela lui permit de proposer des modèles plus réalistes.

Au milieu des années 80, Smith proposé les réseaux de guides d'ondes numériques comme modèle, entre autres, des réverbérations. Un guide d'onde est modélisé par une ligne à retard bi-directionnelle, avec filtres. Enfin, en 1982, Stautner et Puckette ont jeté les bases de ce qui, formalisé par Jot en 1992, forme les réseaux de délais à réinjection.

Réverbération par convolution avec réponse impulsionnelle Dès lorsque l'on a enregistré la réponse impulsionnelle d'une salle, on sait que le son réverbéré est obtenu par convolution de cette réponse impulsionnelle $h(n)$ avec le signal $x(n)$, d'où le signal obtenu :

$$y(n) = (x * h)(n) \tag{2.34}$$

Parmi les défauts de cette méthode, on relève la nécessité de longues réponses impulsionnelles afin d'obtenir des réverbérations naturelles, et l'impossibilité de manipuler simplement (par interpolation) les réponses impulsionnelles, notamment pour passer de l'un à l'autre simplement. Les structures ave filtres et lignes à retard se prêtent beaucoup mieux à ce genre de manipulations.

Intérêt de cet effet

Cet effet permet d'ajouter un effet de salle à un son brut, par exemple enregistré en chambre anéchoïque ou en champ très proche et dans une salle à parois absorbantes.

Compléments d'information

Les attributs perceptifs d'une réverbération permettant de caractériser le qualité acoustique d'une salle se répartissent en 3 catégories :

- la perception de la source (en lien avec le spectre et l'énergie relative du son direct et des premières réflexions) :
  - la présence (rapport entre énergie du son direct et celle des réflexions primaires) ;
  - la brillance (variation hautes fréquences des réflexions primaires) ;
  - la chaleur (variation basses fréquences des réflexions primaires) ;
- l'intéraction source – salle (en lien avec les énergies relatives du son direct, des réflexions primaires et secondaires, et avec les temps de décroissance primaires) :
  - l'enveloppement (énergie des réflexions primaires par rapport au son direct) ;
  - la présence de la salle (énergie du son réverbéré tardif) ;
  - réverbération traînante (temps de décroissance primaire de la réponse impulsionnelle de la salle) ;
- la perception de la salle (en lien avec les temps de décroissance secondaires et leurs variations en fréquence) :
  - la réverbération tardive (temps de décroissance secondaire de la réponse impulsionnelle de la salle) ;
  - la lourdeur ou *heaviness* (variation basses fréquences du temps de décroissance) ;
  - la vivacité ou *liveness* (variations hautes fréquences du temps de décroissance).

2.5.4 Distance

Définition

La distance d'une source sonore est évalué par l'auditeur à l'aide de plusieurs mécanismes : tout d'abord le niveau du signal émis, ensuite le fitrage passe-bas du son (effectué par l'air), puis l'écho et la réverbération (le rapport entre la puissance du signal clair sur la puissance de la réverbération du signal), et enfin des indices cognitifs que nous n'aborderons pas ici (c'est notamment grâce à eux que l'on reconnaît une voix chuchotée d'une voix parlée ou criée, enregistrés en chambre source donc sans réverbération, et égalisées en sonie).

Fonctionnement

Les mécanismes que nous pouvons modéliser sont la réverbération, expliquée en 2.5.3, le filtrage passe-bas et la puissance à l'aide d'un simple potentiomètre.

Le filtrage passe-bas et le changement du gain correspondent à ce qui se passe en espace ouvert (à l'extérieur), lorsque le son traverse de grande étendues d'air. Dans ce cas, les hautes fréquences sont absorbées par l'agitation thermique des molécules présentes dans l'air.

Les indications d'écho et de réverbération modélisent ce qui se passe dans un espace clos, tel qu'une salle de concert. Pour simuler un distance spécifique à l'intérieur d'une pièce, le plus simple consiste à garder le même niveau de réverbération $P_{reverb} = P^0_{reverb}$ et d'échelonner le signal direct de façon à ce qu'il soit inversement proportionnel à l'inverse de la distance. En effet, la puissance du signal direct décroît en proportion inverse de la distance $P_{direct} = \frac{P_0}{d}$, tandis que la puissance du signal réverbéré décroît en proportion de l'inverse de la racine carrée de la distance $P_{reverb} = \frac{P_0}{\sqrt{d}}$. Cette seconde modélisation est bien plus fidèle à la réalité.

Compléments d'information

Pour modéliser un espace fermé avec une ouverture sur un second espace (plus grand), on utilise une réverbération globale relativement faible das l'ensemble des N haut-parleurs, et une réverbération locale plus forte dans deux haut-parleurs adjacents, simulant la direction de l'ouverture.

Cette distinction entre réverbération locale et globale possède un autre intérêt : en utilisant la relation $R_{locale} > \left(1 - \frac{1}{d}\right) R_{globale}$ qui permet d'augmenter la réverbération locale en fonction de la distance, on évite l'effet de masquage qui apparaît à des distances où les amplitudes des signaux directs et réverbérés globaux sont égales. En effet, lorsque la distance tend vers 0 (la source est très proche de l'auditeur), la puissance de la réverbération locale tend vers celle de la réverbération globale, elle est distribuée uniformément, dans tous les canaux. Au contraire, lorsque la source s'éloigne de l'auditeur, le signal réverbérant se concentre dans la direction de la source.

2.5.5 Doppler

Définition

L'effet Dopper correspond aux modifications de niveau et de hauteur perçues lors des déplacements d'une source sonore qui s'approche et/ou s'éloigne de l'auditeur. Cet effet permet à l'auditeur d'analyser le mouvement et d'en connaître les changements de direction et de distance.

Fonctionnement

Dès lors qu'une relation de mouvement existe entre la source sonore et l'auditeur, l'onde sonore s'allonge ou se comprime. Un signal sonore qui s'approche est perçu de façon plus aigüe qu'il n'est émis à la source, et plus grave lorsqu'il s'éloigne. La vitesse de déplacement de la source s'ajoute à celle de propagation du son. Lorsque les deux vont dans le même sens, les vitesses s'ajoutent et la fréquence perçue augmente ; lorsqu'elles vont en sens contraire, la fréquence diminue. De plus lorsque le changement de sens de la source est brusque, un effet complémentaire de rapprochement-éloignement s'ajoute.

Notons c la célérité du son dans l'air et c_s la vitesse de déplacement de la source, c_a la vitesse de déplacement de l'auditeur. Considérons que l'auditeur et la source se déplacent selon un seul et même axe. Soit f_s la fréquence d'une sinusoïde émise par la source, la fréquence perçue par l'auditeur sera :

$$f_d = f_s \frac{c}{c_s - c_a} \tag{2.35}$$

Le changement de hauteur est réalisé par l'une des méthodes présentée en 2.4, avec un rapport contrôlé par la vitesse relative $c_s - c_a$ entre source et auditeur. Le ratio de changement de hauteur est réalisé par une courbe du type (cf. fig. 2.21) :

$$\gamma_{hauteur} = -\sin\left(\pi \frac{log(|t|+1)}{|t|+1}\right) \tag{2.36}$$

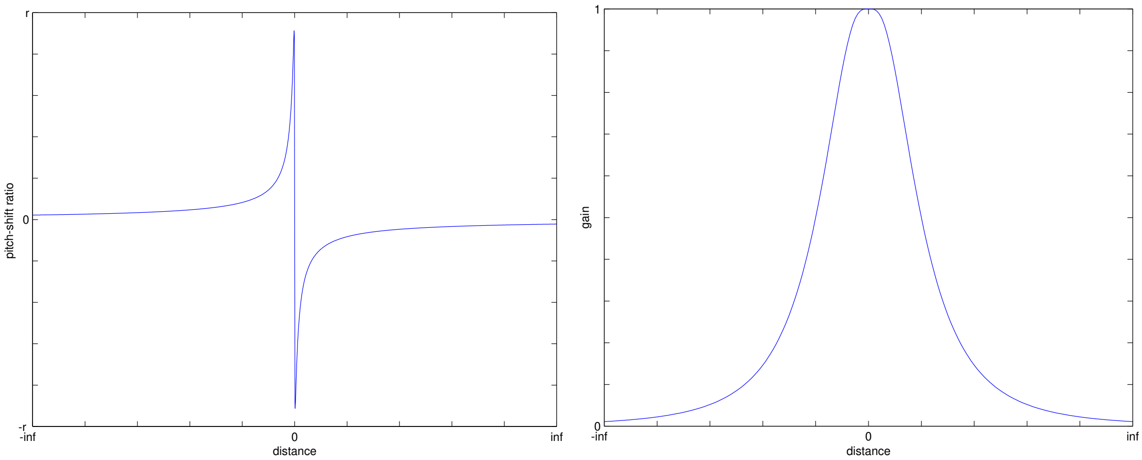

Fig. 2.21: *Ratio de changement de hauteur (fig. gauche) et gain (fig. droite) à appliquer pour réaliser l'effet Doppler.*

Intérêt de cet effet

Cet effet permet de modéliser les déplacements rapides de sources sonores (utilisation principalement pour des sons bruités, notamment par John Chowning, ou pour la réalisation de bandes sons de films).

Compléments d'information

Remarquons qu'un changement de niveau est nécessaire lui aussi (le niveau sonore augmente lorsque la source s'approche, et diminue lorsqu'elle s'éloigne). Un effet connu en psychoacoustique est que lorsque la sonie (intensité sonore perçue) augmente, la hauteur perçue est elle aussi légèrement modifiée, ce qui s'ajoute au changement de hauteur imposé par les différences de vitesses auditeur-source.

Rappel des paramètres de l'effet

Les paramètres permettant de caractériser cet effet sont :

- c_a la vitesse de déplacement de l'auditeur ;
- c_s la vitesse de déplacement de la source.

2.5.6 Effet Rotary sur les enceintes (Leslie)

Définition

La radiation d'un son émis par un haut-parleur rotatif crée un effet de spatialisation saisissant. L'effet Rotary consiste à mettre en rotation une enceinte. Cet effet rend vivants même les sons sourds et stables en les animant de qualités variant dans le temps.

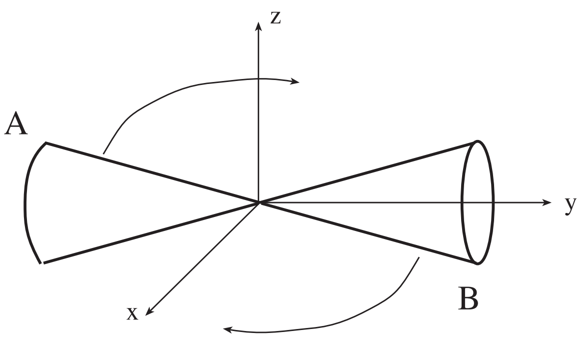

Fig. 2.22: *Système Rotary (à enceintes rotatives) permettant l'effet Leslie.*

Fonctionnement

Le mécanisme originel d'enceinte rotative fut la Cabine Leslie, qui envoyait un signal entrant dans deux mécanismes rotatifs séparés : un pavillon rotatif pour les hautes fréquences et un écran rotatif (bloquant et débloquant un boomer stationnaire) pour les basses fréquences. Une télécommande pour la vitesse du moteur permettait au musicien d'ajuster la vitesse de rotation. Le pavillon résonant de la Cabine Leslie la rend immédiatement identifiable. Ce système fut à l'origine construit pour enrichir le son d'orgues électriques, tels le Hammond B3, avec lequel il était souvent couplé.

Les effets induits par la rotation sont nombreux, mettant en jeu le vibrato de l'effet Doppler, le filtrage variant dans le temps, les déphasages, les distorsions causées par la turbulence de l'air, les réflexions d'échos des surfaces adjacentes, ainsi que les caractéristiques de transfert des amplificateurs et des enceintes utilisés.

Le pavillon est double, le haut-parleur A ayant ses modifications de hauteur et d'intensité dans la direction opposée à celle du haut-parleur B.

Tous ces effets simultanés sont difficiles à simuler de façon convaincante en utilisant le traitement numérique du signal. Une combinaison de modulateurs (pour l'intensité) et de lignes à retard (pour moduler la hauteur) permet de simuler assez bien cet effet. La simulation de l'effet Doppler des deux pavillons opposés est réalisée par deux lignes à retard modulées avec des configurations de vibrato en déphasages de π radians. La caractétistique de directionnalité des haut-parleurs tournant peut être rendue par une modulation en amplitude des signaux en sortie des lignes à retard. Cette modulation est synchrone à la modulation de la ligne à retard, de manière à ce que le mouvement en arrière du pavillon produise une baisse de la hauteur et de la puissance du son. Au point de demi-tour, l'intensité est minimal et la hauteur non modifiée par rapport au son original. Le mouvement inverse, de l'arrière vers l'auditeur, produit les effets inverses : hausse de la hauteur et de l'intensité. Un effet stéréophonique de haut-parleurs tournant est obtenu par un mixage inégal des deux lignes à retard dans les sorties correspondant aux canaux gauche et droit.