



DILATATION ET TRANSPOSITION SOUS CONTRAINTES PERCEPTIVES DES SIGNAUX AUDIO : APPLICATION AU TRANSFERT CINEMA-VIDEO

Grégory Pallone

► To cite this version:

Grégory Pallone. DILATATION ET TRANSPOSITION SOUS CONTRAINTES PERCEPTIVES DES SIGNAUX AUDIO : APPLICATION AU TRANSFERT CINEMA-VIDEO. Modélisation et simulation. Université de la Méditerranée - Aix-Marseille II, 2003. Français. NNT : . tel-00003363v1

HAL Id: tel-00003363

<https://theses.hal.science/tel-00003363v1>

Submitted on 11 Sep 2003 (v1), last revised 20 Feb 2007 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE DE LA MEDITERRANEE - AIX-MARSEILLE II
FACULTE DES SCIENCES DE LUMINY

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE d'AIX-MARSEILLE II

ECOLE DOCTORALE DE MECANIQUE, PHYSIQUE ET MODELISATION

**Spécialité : ACOUSTIQUE, TRAITEMENT DU SIGNAL ET INFORMATIQUE
APPLIQUES A LA MUSIQUE**

présentée et soutenue publiquement le 20/06/2003 par

Grégory PALLONE

**DILATATION ET TRANSPOSITION SOUS CONTRAINTES
PERCEPTIVES DES SIGNAUX AUDIO :**

APPLICATION AU TRANSFERT CINEMA-VIDEO.

Devant la commission d'examen formée de :

M. BOUSSARD Patrick	(Examineur)
Dr. CANEVET Georges	(Examineur)
Pr. DEPALLE Philippe	(Rapporteur)
Dr. KRONLAND-MARTINET Richard	(Directeur de thèse)
Pr. SERRA Xavier	(Rapporteur)
Pr. TORRESANI Bruno	(Président du jury)

Ce dont on ne peut parler, il faut le taire.

L. Wittgenstein, *Tractatus logico-philosophicus*.

Remerciements

Cette thèse est le fruit d'un travail d'équipe où l'implication de chacun des partenaires a permis de mener à bien cette recherche. J'ai ainsi eu le privilège de partager mon temps entre la société GENESIS et l'équipe S2M du LMA-CNRS.

Je souhaite remercier en premier lieu Patrick Boussard, directeur de GENESIS, de m'avoir fait confiance depuis le début de ce projet, et sans qui ce travail n'aurait pas vu le jour. Merci aussi à toute l'équipe de GENESIS pour l'agréable ambiance de travail qu'ils ont su créer, et tout particulièrement à Florent Jaillet, Manu Deschamps et Benoit Jacquier pour leur aide précieuse.

Je remercie également Richard Kronland-Martinet, qui a dirigé ma thèse, pour m'avoir accueilli au sein de l'équipe scientifique et de m'avoir prodigué tant de bons conseils. Merci aussi à tous les membres de l'équipe S2M pour leur participation, et plus particulièrement à Philippe Guillemain pour le développement des méthodes temps-fréquence et Thierry Voinier pour sa disponibilité.

Je remercie Bruno Torresani qui me fait l'honneur de présider ce jury. Je remercie aussi Philippe Depalle et Xavier Serra d'avoir bien voulu accepter la charge de rapporteur. Je remercie également Georges Canévet d'avoir bien voulu participer au jugement de ce travail.

J'exprime ma gratitude à Dominique Toussaint de Cinestéréo pour avoir encouragé le développement de l'HARMO, à Michel Baptiste de la CST, à l'ANRT pour leur financement à travers la Convention Industrielle de Formation par la Recherche (CIFRE), au Ministère de l'Economie des Finances et de l'Industrie et au Centre National de la Cinématographie (CNC) pour leur financement à travers leur Programme pour la Recherche et l'Innovation dans l'Audiovisuel et le Multimédia (PRIAMM).

J'adresse mes remerciements à Pascal Gobin, Jordi Bonada, et toute l'équipe "Audio 3D" de Creative et en particulier Jean-Marc Jot pour son accueil lors d'un stage sur la spatialisation.

Enfin, pour finir sur une note plus personnelle, je remercie ma famille et mes amis du "Nord" de la France qui m'ont soutenu, et un merci tout spécial à Margot.

Avertissement aux lecteurs

Les résultats de ce travail sont valorisés au travers de la commercialisation d'un produit baptisé HARMO. La thèse a été initiée et financée par la société GENESIS avec l'aide de l'ANRT (bourse CIFRE). L'algorithme a été développé dans le cadre d'un partenariat entre cette société et le LMA-CNRS (équipe Modélisation, synthèse et contrôle des signaux sonores et musicaux). Il en résulte la confidentialité de la dernière annexe de ce document, présente uniquement dans les exemplaires à caractère confidentiel. Pour toute information, veuillez contacter l'une des trois personnes suivantes:

- Patrick Boussard (patrick.boussard@genesis.ac)
- Richard Kronland-Martinet (kronland@lma.cnrs-mrs.fr)
- Grégory Pallone (gregory@pallone.fr)

Le présent document est accompagné d'un Compact Disc contenant 97 pages audio lisibles par la plupart des lecteurs de CD audio, ainsi qu'une session CD-ROM, accessible par la plupart des ordinateurs, sur laquelle se trouvent une version de la thèse (au format PDF), les autres publications de l'auteur, des documents concernant l'HARMO, ainsi qu'un certain nombre de sons (au format WAV) originaux et traités par le système, issus du CD de démonstration de l'HARMO.

Dans la thèse, les références aux pistes audio sont notées entre crochets, à l'aide du numéro de page. Les références bibliographiques sont également notées entre crochets, avec les premières lettres du nom de l'auteur (ou les initiales pour plusieurs auteurs), précédant l'année de publication.

Table des matières

Introduction	17
Différence des formats cinéma et vidéo	17
Conséquences liées au son	18
Un besoin induit : l'harmoniseur	18
Plan du document	19
 1 Problématique	 21
1.1 Généralités sur la dualité temps/fréquence	21
1.1.1 Dilatation d'un signal quelconque	21
1.1.2 "Dilatation temporelle mathématique"	21
1.1.3 "Transposition fréquentielle mathématique"	21
1.1.4 Inadéquation des solutions proposées	22
1.1.5 "Lecture à vitesse variable" et "Rééchantillonnage"	24
1.1.6 Conclusion sur la dualité temps-fréquence	25
1.2 Concept de "transformation-p"	26
1.2.1 Concept de "dilatation-p"	26
1.2.2 Concept de "transposition-p"	27
1.2.3 Des problèmes "semi-duaux"	27
1.2.4 Prise en compte des spécificités de l'oreille	29
1.2.5 Fonctions de dilatation et transposition	30
1.3 Des contraintes particulières pour le transfert cinéma/vidéo	31
1.3.1 Contraintes technologiques	31
1.3.2 Contraintes de qualité sonore	33
1.4 D'autres applications des transformations-p	36
1.4.1 Historique des besoins en transformation-p	36
1.4.2 Applications à une source vocale	36
1.4.3 Applications à une source sonore complexe	38
1.5 Technologie actuellement disponible	40
 2 Classification des méthodes	 41
2.1 Préambule à la classification	41
2.1.1 Transformation-p d'une représentation	41
2.1.2 Propriétés des transformation-p	45
2.1.3 Discussion sur la modification des formants	45
2.1.4 Bilan des représentations	46
2.2 Méthodes temporelles	47
2.2.1 Principe général des méthodes temporelles	47
2.2.2 Méthodes "aveugles"	55
2.2.3 Méthodes adaptatives	64

2.2.4	Méthodes recourant à des décompositions préalables	76
2.2.5	Bilan des "méthodes temporelles"	80
2.3	Méthodes fréquentielles	82
2.3.1	Principe général des méthodes fréquentielles	82
2.3.2	Méthodes "aveugles" de dilatation-p	97
2.3.3	Méthodes "aveugles" de transposition-p	101
2.3.4	Méthodes adaptatives	103
2.3.5	Bilan des "méthodes fréquentielles"	105
2.4	Méthodes temps-fréquence	106
2.4.1	Principe général des méthodes temps-fréquence	106
2.4.2	Représentation temps-échelle	107
2.4.3	Représentation basée sur l'échelle des Barks	109
2.4.4	Représentation adaptative	109
2.4.5	Représentation temps-fréquence multi-résolution	110
2.4.6	Méthodes recourant à des décompositions préalables	112
2.5	Conclusions sur la classification	113
3	Innovations algorithmiques et évaluations	115
3.1	Etude de l'anisochronie émanant des méthodes temporelles	115
3.1.1	Tempo "lent"	116
3.1.2	Tempo "rapide"	119
3.1.3	Tempo "modéré"	120
3.1.4	Bilan sur l'anisochronie	120
3.2	Méthodes temps-fréquence adaptées à l'audition	123
3.2.1	Transposition-p par une méthode temps-fréquence	126
3.2.2	Dilatation-p par une méthode temps-fréquence	129
3.2.3	Discussion sur les méthodes adaptées à l'audition	131
3.3	Dilatation-p par méthodes couplées	132
3.3.1	Décomposition en sous bandes	132
3.3.2	Décomposition hybride	134
3.3.3	Discussion sur les méthodes couplées	136
3.4	Algorithme de dilatation-p HARMO	137
3.4.1	Principe de la méthode	137
3.4.2	Développement de l'outil HARMOLAB	139
3.4.3	Critère de sélection de K	141
3.4.4	Critère de sélection de I	148
3.4.5	Critère de sélection de FE	154
3.4.6	Mesure de la distorsion	162
3.4.7	Traitement multicanal	164
3.5	Evaluations des méthodes	169
3.6	Conclusions sur les innovations algorithmiques	173
4	Conceptions matérielle et logicielle	175
4.1	Développement matériel	175
4.1.1	Enoncé du besoin	175
4.1.2	Choix de l'architecture et des composants	175
4.2	Développement logiciel	184
4.2.1	Processeur P1	185
4.2.2	Processeur P0	188

<i>Table des matières</i>	11
4.2.3 Relations P0/P1	191
4.3 Problématique temps-réel	194
4.3.1 Difficultés spécifiques liées à l'algorithme	194
4.3.2 Défauts et points forts de l'algorithme	195
5 Conclusion	197
A Dilatation et transposition avec conservation de l'énergie	201
B Les principaux systèmes et logiciels de transformation-p	203
C Méthode PSOLA	209
D Magnétophone à têtes tournantes	215
E Films et sons traités par l'HARMO	223
Références bibliographiques	227
Références aux pistes audio	243

Notations

Variables

L	Marque de lecture
E	Marque d'écriture
H	Support ou durée de la fenêtre h
P	Période fondamentale du signal
Γ	Tolérance temporelle sur les marques de lecture/écriture
K_A, K_B	Segments à mixer
K_M	Segment mixé
K	Durée du segment inséré
k_{min}	Durée minimale du segment inséré
k_{max}	Durée maximale du segment inséré
k_{limite}	Durée-seuil entre indiquant la limite entre "court" et "long" segment
K_1	Durée optimale du segment inséré pour un segment "court"
K_2	Durée optimale du segment inséré pour un segment "long"
R	Durée du segment résiduel à insérer pour terminer l'itération
N_c	Durée sur laquelle est estimée la similarité
N	Taille de la FFT (en échantillons)
t	Variable temporelle continue associée au signal
$n = t_n$	Variable temporelle discrète associée au signal
τ	Variable temporelle continue associée à la transformée temps-fréquence
$p = \tau_p$	Variable temporelle discrète associée à la transformée temps-fréquence
Ω	Variable fréquentielle continue associée à la transformée temps-fréquence
Ω_k	Variable fréquentielle discrète associée à la transformée temps-fréquence
ω	Variable fréquentielle continue associée au signal
ω_k	Variable fréquentielle discrète associée au signal
	Par abus de langage, nous appelons ω "fréquence" alors qu'il s'agit littéralement d'une "pulsation" : $\omega = 2\pi f$.

Notations générales

Transformation-p	Transformation sous contraintes perceptives
Dilatation-p	Dilatation temporelle sous contraintes perceptives
Transposition-p	Transposition fréquentielle sous contraintes perceptives
α	Taux de dilatation-p, transposition-p ou de rééchantillonnage
F_e	Fréquence d'échantillonnage
j	$\sqrt{-1}$
\bar{x}	Complexe conjugué de x
$\langle x, y \rangle$	Produit scalaire de x et y
\bar{w}	Complexe conjugué de w
C^{te}	Constante

Opérateurs et fonctions

Dp_α	Opérateur de dilatation-p de taux α
Tp_α	Opérateur de transposition-p de taux α
$R_\alpha \equiv D_\alpha = \alpha T_{\frac{1}{\alpha}}$	Opérateur de rééchantillonnage de taux α
$*$	Opérateur de convolution
F	Opérateur de Fourier
$\hat{s}(f)$	Transformée de Fourier du signal s
$S(\tau, \Omega)$	Transformée temps-fréquence
$S(p, \Omega_k)$	Transformée temps-fréquence discrète
$M(\tau, \Omega)$	Module de la transformée temps-fréquence
$\varphi(\tau, \Omega)$	Phase de la transformée temps-fréquence
$D(t)$	Fonction de dilatation
$T(t)$	Fonction de transposition
$s(t)$	Signal temporel
$h(t)$	Fenêtre temporelle (d'analyse)
$v(t)$	Fenêtre temporelle (de synthèse)
$g(t)$	Grain temporel ou signal fenêtré
$r(t)$	Signal de référence pour un traitement multicanal
$C(k)$	Mesure de similarité dépendant du retard k
$C_{xy}(k)$	Fonction de corrélation
$CN_{xy}(k)$	Fonction de corrélation normalisée

Abbreviations

IOI	"Inter Onset Interval" : intervalle entre 2 pulsations
BPM	Battement Par Minute
JND	"Just Noticeable Difference" : la plus petite différence audible
TFD	Transformée de Fourier Discrète
TFDI	Transformée de Fourier Discrète Inverse
TFCT	Transformée de Fourier à Court Terme
RFCT	Représentation de Fourier à Court Terme (obtenue par TFCT)
TO	Transformée en ondelettes
FFT	"Fast Fourier Transform" : algorithme rapide de TFD
IFFT	"Inverse Fast Fourier Transform" : algorithme rapide de TFDI
OLA	"OverLap-Add" (Recouvrement-Addition)
WOLA	"Weighted OverLap-Add" (Recouvrement-Addition pondéré)
SOLA	"Synchronous OverLap-Add" (Recouvrement-Addition synchronisé)
PSOLA	"Pitch-Synchronous OverLap-Add" (Recouvrement-Addition synchronisé à la période fondamentale)

Introduction

Cette thèse a pour cadre une collaboration entre l'entreprise GENESIS S.A. située à Aix-en-Provence et le CNRS-LMA de Marseille. Elle a également été financée en partie par l'Association Nationale de la Recherche Technique (ANRT).

La recherche algorithmique a été valorisée par la réalisation d'un produit audionumérique multicanal baptisé "**HARMO**", utilisé actuellement dans plusieurs studios de post-production cinématographique en Europe. Ce projet a reçu le soutien financier du Ministère de l'Economie des Finances et de l'Industrie et du Centre National de la Cinématographie (CNC) à travers leur Programme pour la Recherche et l'Innovation dans l'Audiovisuel et le Multimédia (PRIAMM).

Différence des formats cinéma et vidéo

La problématique que nous exposons ici trouve sa source dans l'existence de formats différents entre le cinéma et la vidéo en ce qui concerne la vitesse de projection des images. Cependant, comme nous allons le voir, ce problème touche également le son.

Le cinéma, tout comme la vidéo et la télévision, consiste à présenter à un spectateur une succession rapide d'images fixes créant ainsi l'illusion d'un mouvement continu, grâce à la persistance rétinienne. Bien que cette illusion puisse s'exercer à des cadences aussi basses que 12 images par seconde (i/s) [Coo90], la cadence utilisée pour le cinéma moderne est de 24 i/s, alors que celle retenue en Europe pour la vidéo est de 25 i/s [BP00, Whi02].

On désire parfois convertir d'un format vers l'autre. On parle alors de transfert : passage d'un film tourné à 24 i/s vers le format vidéo à 25 i/s (pour une utilisation télédiffusion, VHS ou DVD), ou bien passage d'une production vidéo tournée à 25 i/s (à l'aide d'une caméra vidéo ou numérique) vers le format cinéma à 24 i/s en vue d'une projection en salle. Ce dernier cas, assez marginal jusqu'à peu, devient de plus en plus fréquent du fait de l'apparition des caméras numériques haute définition ("Digital Video" ou DV) [Col01].

Cette conversion pourrait s'effectuer en modifiant le nombre d'images de manière à conserver la durée globale du film. Il faudrait alors utiliser une technique de "compensation", basée sur la répétition ou la suppression de certaines images, altérant le rendu des mouvements continus ainsi que la synchronisation précise du son. C'est pourquoi cette méthode n'est généralement retenue que dans l'étape intermédiaire de montage avec le télécinéma [Vil00]. Il serait également possible d'effectuer une interpolation d'images si cette technique était suffisamment développée.

En Europe, la conversion finale est généralement réalisée en conservant le nombre d'images total. Nous nous intéressons ici uniquement à ce type de conversion, qui est l'alternative aux techniques précédentes. Dans ce type de conversion, la durée globale du film se trouve altérée (accélération du film lors du passage de 24 i/s vers 25 i/s donc diminution de sa durée, et inversement, ralentissement du film lors du passage de 25 i/s vers 24 i/s donc augmentation de sa durée) puisqu'un même nombre d'images ne sera pas visualisé dans un même intervalle de

temps selon la vitesse de lecture.

La modification de la durée du film implique inévitablement la même modification sur la durée de la bande-son, sans quoi la synchronisation image et son ne serait plus assurée. Le changement de la durée de la bande-son est réalisé en fixant les vitesses (pour l'analogique) ou les fréquences d'échantillonnage (pour le numérique) des appareils de lecture et d'enregistrement dans le même rapport que celui de la dilatation temporelle désirée.

Et c'est là que les ennuis commencent pour la bande-son...

Conséquences liées au son

En effet, une accélération d'un son induit un "rehaussement" des fréquences. Il suffit pour s'en convaincre d'écouter un disque vinyle 33 tours/minute lu à 45 tours/minute. Inversement, un ralentissement induit un "abaissement" des fréquences. Ainsi, une dilatation temporelle entraîne inéluctablement une transposition fréquentielle.

Le contenu fréquentiel de la bande-son est donc transposé de 24/25 (-4%) lors de la projection à 24 i/s d'un film initialement tourné à 25 i/s (les sons paraissent plus graves), et transposé de 25/24 (+4,2%)¹ lors de la projection à 25 i/s d'un film initialement tourné à 24 i/s (les sons paraissent plus aigus).

Il est préférable, lors de la projection, de retrouver le contenu fréquentiel de la bande-son originale, malgré le changement de sa durée. Exprimé d'une autre manière, et dans le cas cinéma vers vidéo, nous devons compenser les modifications fréquentielles dues à l'accélération du film pour retrouver lors de la diffusion, des sons certes plus rapides, mais paraissant "naturels".

Un besoin induit : l'harmoniseur

Cependant, il est possible d'insérer dans la chaîne de conversion de format, une machine de transformation du son permettant de compenser préalablement la déformation fréquentielle sonore due à l'accélération ou au ralentissement. Cet appareil doit donc transposer toutes les fréquences de la bande-son : ce système est appelé par le milieu professionnel "harmoniseur" [Vil00], en référence à la marque de produit "Harmonizer" (un des premiers appareils commercialisés effectuant ce type de traitement) commercialisé par la société EVENTIDE [UQA96].

Actuellement, ce traitement tend à se généraliser mais il n'est pas encore appliqué à toutes les productions cinématographiques et audiovisuelles. Les raisons en sont d'une part un prix plus important, et d'autre part une qualité de traitement pas totalement satisfaisante, ainsi que l'inexistence jusqu'en 2000 d'une machine temps-réel (indispensable pour la productivité des studios de post-production) adaptée aux nouveaux formats sonores multicanaux de type 5.1 [Neu98].

La **problématique générale** à laquelle nous devons faire face est la dilatation temporelle globale d'une bande-son avec conservation des qualités spectrales² subjectives : en effet, le son présenté au final à l'auditeur est un son ralenti ou accéléré par rapport à l'original.

Cependant, la **problématique particulière** à laquelle nous sommes confrontés est la transposition fréquentielle d'une bande-son sans modification de la durée, sachant que l'objectif est une dilatation (ce but à atteindre est très important car il nous guidera lors de certains choix que nous aurons à faire).

1. Dans un souci de simplification, le rapport 25/24, qui vaut environ +4,16667%, est arrondi à +4,2%.

2. Hauteurs tonales et timbres identiques.

La figure 1 schématise les opérations effectuées en studio de post-production audiovisuelle. On y montre un traitement sans harmonisation, pour lequel le signal sonore, lu à une vitesse variable, est globalement dilaté en temps, mais aussi en fréquence. On y montre également l'harmonisation de la bande-son, à travers les 2 possibilités pour effectuer le traitement. Dans ce cas, les fréquences finales sont restituées fidèlement aux originales. Les deux problématiques, générale et particulière, y sont représentées, et l'on indique la durée et la fréquence du signal en chacun des points de la chaîne de traitement.

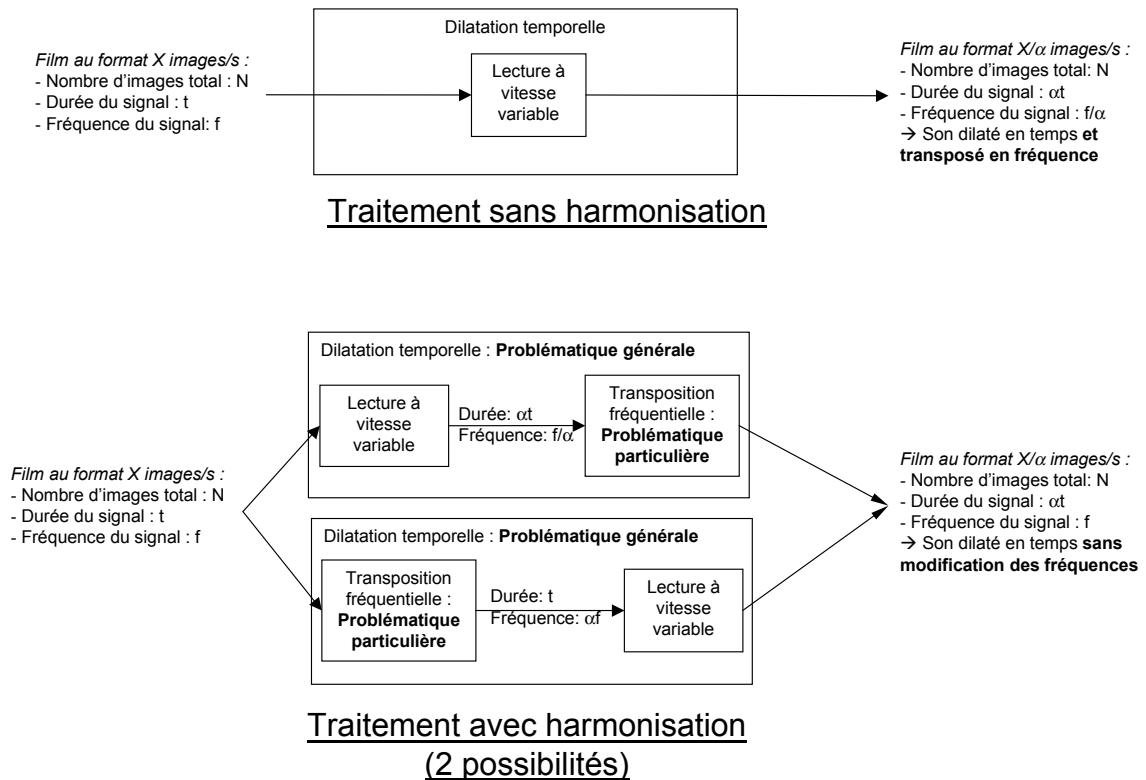


Figure 1 – Illustration des problématiques générale et particulière

Plan du document

Dans le premier chapitre, nous posons le problème de manière formelle, à savoir la question de la dilatation temporelle sous contraintes perceptives ou "dilatation-p", c'est-à-dire la dilatation temporelle pour laquelle aucune modification du timbre ou des fréquences n'est perçue. Le problème de la transposition fréquentielle sans modification de la durée est également évoqué. Nous montrons que la solution immédiate consistant à dilater la forme d'onde ne répond pas à ces problèmes. Nous sommes donc amenés à introduire de nouveaux concepts.

Le second chapitre est dédié au recensement et à la classification des différentes méthodes qui permettent d'effectuer en pratique ces transformations. Il s'agit, au delà d'un simple état de l'art quasi-exhaustif, de proposer un formalisme commun pour comprendre les relations souvent très fortes qu'entretiennent les différentes méthodes.



Figure 2 – Photo du produit HARMO, un système audionumérique multicanal de transposition fréquentielle en temps réel, dans lequel se trouve l’algorithme développé dans cette thèse.

Nous explorons dans le troisième chapitre quelques pistes afin d’améliorer un certain nombre de ces différentes méthodes. Cette étude, réalisée dans un contexte de collaboration recherche/industrie, a eu pour objectif de mettre sur le marché un produit répondant à des critères technologiques mais aussi de qualité sonore. C’est pourquoi il est nécessaire, à un moment donné de l’étude, de sélectionner la méthode qui donne les meilleurs résultats perceptifs, même si d’autres méthodes semblent prometteuses et n’ont pas bénéficié de recherches assez poussées.

Le quatrième chapitre se concentre sur les aspects électronique et programmation de la machine. L’algorithme retenu est optimisé, à la fois en terme de qualité mais aussi de puissance de calcul, puis valorisé à travers une implantation temps-réel pour le produit professionnel HARMO, satisfaisant actuellement la plupart de ses utilisateurs.

Chapitre 1

Problématique

1.1 Généralités sur la dualité temps/fréquence

1.1.1 Dilatation d'un signal quelconque

On appelle dilatation de taux α d'une fonction $s \in L^2(\mathbb{R})$ l'homothétie de rapport α . L'opérateur D_α associé à cette transformation est donc défini par :

$$D_\alpha[s](x) = s\left(\frac{x}{\alpha}\right) \quad \alpha \in \mathbb{R}^{+*} \quad (1.1)$$

Cet opérateur de changement de variable homothétique modifie le support de la fonction sur laquelle il est appliqué. En effet,

$$\text{Supp } s = [0, X] \Rightarrow \text{Supp } D_\alpha[s] = [0, \alpha X]$$

Par conséquent, effectuer un changement de variable homothétique sur une fonction, en posant $X = \frac{x}{\alpha}$, revient à multiplier son support par l'inverse de la constante d'homothétie.

1.1.2 "Dilatation temporelle mathématique"

On appelle **dilatation temporelle mathématique** la transformation de dilatation lorsqu'elle est appliquée à un signal temporel. Elle permet d'augmenter ou diminuer la durée du signal sans changer sa forme d'onde globale. En d'autres termes, on obtient un signal ralenti ou accéléré par rapport à l'original.

D_α agit sur le support temporel d'un signal s dans le sens d'une élongation pour $\alpha > 1$, et dans le sens d'une contraction pour $\alpha < 1$. Nous considérerons ici uniquement les rapports $\alpha > 0$.

1.1.3 "Transposition fréquentielle mathématique"

Intéressons-nous maintenant aux conséquences de l'action de l'opérateur D_α dans le domaine fréquentiel.

Pour cela, on notera F l'opérateur de Fourier :

$$\begin{aligned} F : \mathcal{L}^2(\mathbb{R}) &\rightarrow \mathcal{L}^2(\mathbb{R}) \\ s &\rightarrow F[s] = \hat{s} \end{aligned}$$

$$\text{tel que } \hat{s}(\omega) = F[s](\omega) = \int_{-\infty}^{+\infty} s(t)e^{-j\omega t} dt$$

La transformée de Fourier d'un signal dilaté est alors donnée par :

$$F[D_\alpha[s]](\omega) = \alpha \hat{s}(\alpha\omega) = \alpha D_{\frac{1}{\alpha}}[\hat{s}](\omega)$$

Par conséquent, l'action de l'opérateur D_α dans le domaine fréquentiel est une dilatation de l'axe des fréquences de rapport $1/\alpha$. On obtient alors un signal plus aigu ou plus grave. Nous appelons cette transformation **transposition fréquentielle mathématique**, notée T_α , lorsque l'opérateur de dilatation agit sur un signal de type fréquentiel. On a donc :

$$T_\alpha = \alpha D_{\frac{1}{\alpha}} \quad (1.2)$$

T_α agit sur le support fréquentiel d'un signal s dans le sens d'une transposition vers les hautes fréquences pour $\alpha > 1$, et dans le sens d'une transposition vers les basses fréquences pour $\alpha < 1$. Nous considérerons ici uniquement les rapports $\alpha > 0$.

1.1.4 Inadéquation des solutions proposées

Les transformations précédemment définies ne sont pas les solutions recherchées aux problèmes de la production d'un signal dilaté temporellement *et conservant le support fréquentiel* (conservant le spectre) ou de la production d'un signal transposé en fréquence *et conservant le support temporel* (conservant la durée).

En effet, il découle de la relation de dualité de l'équation 1.2 qu'une dilatation de l'axe temporel entraîne une dilatation inverse de l'axe fréquentiel :

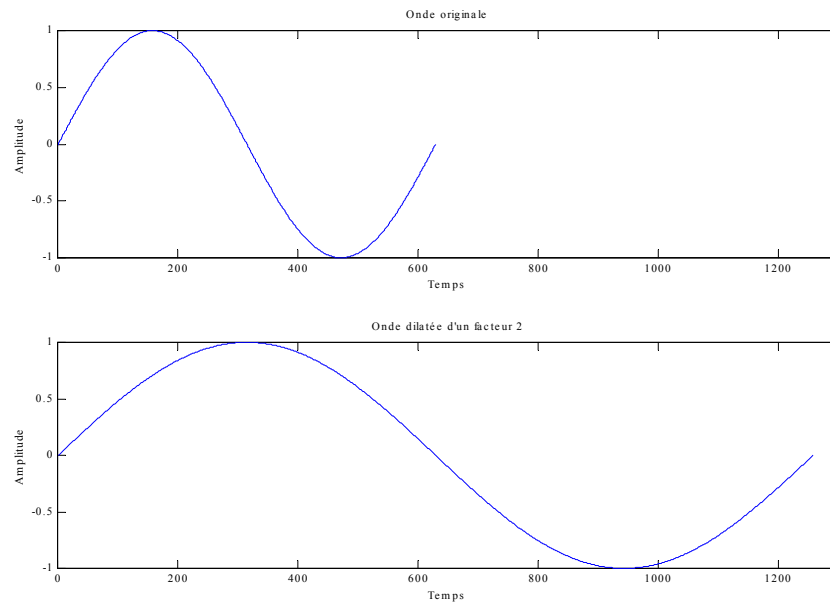
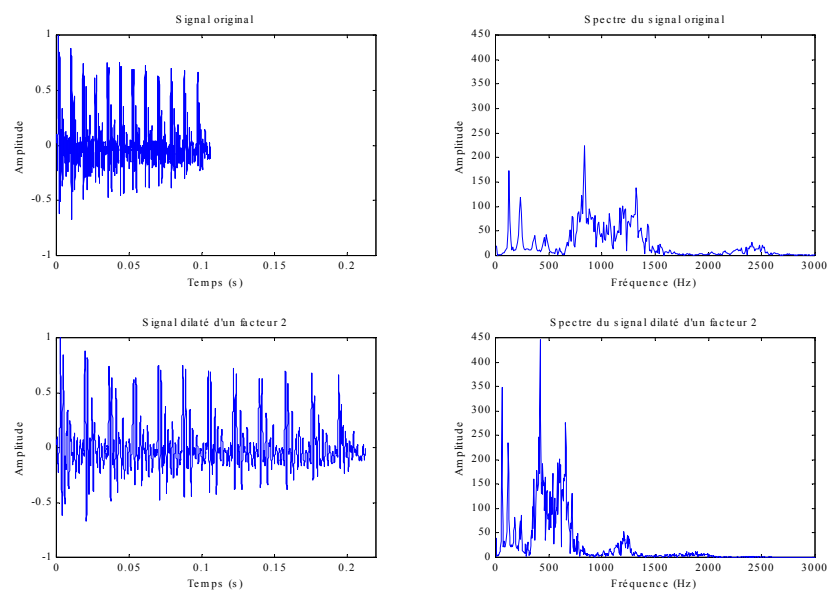
$$s'(t) = s(\alpha t) \Leftrightarrow \hat{s}'(\omega) = \frac{1}{\alpha} \hat{s}\left(\frac{\omega}{\alpha}\right) \quad (1.3)$$

Ainsi, une accélération (resp. ralentissement) d'un signal temporel entraîne inéluctablement une transposition de ses fréquences vers les aigus (resp. vers les graves). C'est ce qui se passe lorsque l'on réalise l'expérience de la lecture d'un disque vinyle 33 tours/min en 45 tours/minute (que l'on appelle "effet 33T/45T") : on transpose toutes les fréquences d'un facteur 1,36 soit 36% mais la durée se trouve altérée d'un facteur de $1/1,36=0,74$ soit -26%.

La figure 1.1 illustre ce phénomène en montrant une période d'une onde sonore, avant et après une dilatation mathématique de facteur 2. On comprend alors que l'allongement de l'onde entraîne une diminution de la fréquence (qui est par définition l'inverse de sa période). Le son [3] et la figure 1.2 illustrent également ce phénomène sur un signal vocal.

Si l'on avait considéré les transformations avec conservation de l'énergie, un terme de normalisation $\alpha^{-1/2}$ serait apparu dans ces définitions. Il en aurait résulté la disparition du terme α dans l'équation 1.2 (voir annexe A).

Cependant, la conservation de l'énergie n'est pas respectée en pratique lorsqu'on lit un disque plus rapidement (l'amplitude reste alors constante) ou lorsque l'on crée des échantillons par l'opération de rééchantillonnage numérique, c'est pourquoi nous préférons conserver la définition donnée par l'équation 1.2, qui reflète mieux ce qui se passe dans la réalité. Ceci explique notamment pourquoi le signal temporel dilaté de la figure 1.2 est globalement plus énergétique que celui de l'original, et pourquoi le spectre du signal dilaté possède des valeurs plus élevées que celui de l'original : une même puissance instantanée (énergie par unité de temps) est émise mais pendant une durée plus longue.

Figure 1.1 – *Dilatation mathématique de facteur 2 d'une onde sinusoïdale*Figure 1.2 – *Dilatation mathématique de facteur 2 d'un signal réel*

1.1.5 "Lecture à vitesse variable" et "Rééchantillonnage"

Nous appelons "**lecture à vitesse variable**" (communément appelé par son terme anglais "varispeed", de "variable-speed" vitesse variable) l'opération analogique consistant à lire une bande magnétique à une vitesse différente de celle utilisée lors de l'enregistrement. Il en résulte une accélération ou un ralentissement du son, qui s'accompagne inéluctablement d'un réhaussement ou d'un abaissement des fréquences ("effet 33T/45T").

Le terme "lecture à vitesse variable" peut parfois désigner le changement dynamique de vitesse, cependant, nous le définissons ici comme étant un rapport fixe non trivial ($\neq 1$) entre vitesse de lecture et d'enregistrement (le terme "lecture à vitesse différente" serait alors plus approprié).

Ainsi, on parle par exemple d'une "lecture à vitesse variable" de +4% lorsque la vitesse de lecture est supérieure de 4% à la vitesse d'enregistrement. Cet exemple correspond à la diffusion sur support vidéo (25 images/seconde) d'un film tourné à 24 images/seconde.

Nous appelons "**rééchantillonnage**" l'opération numérique qui consiste à transformer un signal numérique d'entrée de X échantillons en un signal numérique de sortie de Y échantillons. La contrainte régissant cette transformation est que les signaux continus correspondants doivent être liés entre eux par une homothétie. On passe donc d'un signal numérique de X échantillons, échantillonné à la fréquence F_e , à un signal numérique de αX échantillons par une opération de rééchantillonnage d'un facteur α .

D'une part, si ce signal rééchantillonné est lu à une fréquence d'échantillonnage de αF_e , on obtient les mêmes caractéristiques temporelles et fréquentielles que le signal original, aux conditions de Cauchy-Nyquist-Shannon ([Cau41]-[Nyq28]-[Sha48])¹ sur la bande passante près (les caractéristiques sont identiques dans la bande fréquentielle $[0, F_{e_{min}}/2]$, avec $F_{e_{min}}$ la fréquence d'échantillonnage la plus faible entre F_e et αF_e).

D'autre part, si ce signal rééchantillonné est lu à la fréquence d'échantillonnage initiale F_e , nous obtenons une dilatation temporelle accompagnée d'une transposition fréquentielle : il s'agit alors de l'équivalent numérique de la "lecture à vitesse variable".

Dans le reste de cet exposé, chaque fois que sera utilisé le terme "rééchantillonnage", il s'agira de la transformation du domaine numérique modifiant simultanément les supports temporels et fréquentiels. Ce terme est donc équivalent au terme anglais "resampling" et la méthode de transformation employée se nomme "conversion de fréquence d'échantillonnage" ("Sample Rate Conversion" en anglais, souvent abrégé en "SRC"). Les transformations D_α et $T_{1/\alpha}$ définies précédemment sont identiques à l'opération de rééchantillonnage (à un facteur près, cf. équation 1.2) et sont donc désormais remplacées par R_α :

$$R_\alpha = D_\alpha = \alpha T_{\frac{1}{\alpha}} \quad (1.4)$$

Un rééchantillonnage de facteur $\alpha > 1$ avec relecture à la fréquence d'échantillonnage initiale correspond donc à une élongation du signal (ralentissement) avec transposition des fréquences vers les graves.

On utilisera également l'opérateur R_α pour désigner l'opération de "lecture à vitesse variable", où α correspond au rapport des durées finales et initiales, ainsi qu'à l'inverse du rapport des vitesses (ou fréquences d'échantillonnage F_e) finales et initiales.

$$\alpha = \frac{Duree_{finale}}{Duree_{initiale}} = \frac{Vitesse_{initiale}}{Vitesse_{finale}} = \frac{F_{e_{initiale}}}{F_{e_{finale}}} \quad (1.5)$$

1. Pour un historique du théorème de l'échantillonnage, on pourra consulter [Poh00].

Les différentes techniques de rééchantillonnage sont exposées dans [Lar95], et une méthode efficace en terme de calculs peut être trouvée dans [SG84].

1.1.6 Conclusion sur la dualité temps-fréquence

La dualité temps-fréquence exposée jusqu'ici est intrinsèquement liée à la définition donnée à la fréquence f , qui est l'inverse d'une durée t :

$$f = \frac{1}{t}$$

Cette définition mène à une relation forcément étroite entre ces deux concepts. Ainsi, nous avons vu que la dilatation d'un signal temporel implique une dilatation de ses fréquences ("effet 33T/45T"). Nous appelons cette transformation dans le domaine analogique "lecture à vitesse variable" et dans le domaine numérique "rééchantillonnage". Bien que cette opération nous est utile par la suite, elle ne convient pas pour notre application dont le but final est de modifier la durée du signal sans modifier ses fréquences.

1.2 Concept de "transformation-p"

Jusqu'ici, nous n'avons pas trouvé de solution satisfaisant les problèmes de dilatation temporelle sans modification des fréquences, et de transposition fréquentielle sans modification de la durée. En effet, la théorie du signal se heurte à la dualité temps-fréquence qui stipule que la dilatation mathématique d'un signal induit simultanément et inévitablement une dilatation temporelle et une transposition fréquentielle. Cette transformation implique une modification des 2 dimensions (temporelle et fréquentielle), alors que l'on souhaite une modification agissant uniquement sur une seule de ces deux dimensions.

Cependant, d'un point de vue perceptif, il semble exister des solutions intuitives à ces problèmes! On imagine tout à fait ce que serait la transformation de la voix d'une personne parlant plus lentement mais à la même hauteur, ou encore la transformation du son d'un instrument jouant à l'octave mais au même tempo. Cependant, ces solutions ne sont pas uniques et dépendent de ce que l'on désire réellement obtenir. Pour ce qui est de la dilatation temporelle, cherche-t-on à ralentir ou accélérer les transitoires? Pour ce qui est de la transposition fréquentielle, cherche-t-on à modifier les fréquences des transitoires, la position des formants (indices caractéristiques des timbres des locuteurs)? On s'aperçoit que seule la perception peut nous guider vers une modification adéquate.

Dans l'optique d'une démarche liée à la perception, il est nécessaire d'introduire de nouveaux concepts de transformations construites pour répondre à des contraintes perceptives. Nous les nommerons pour cela "**transformation sous contraintes perceptives**" noté "**transformations-p**".

Nous attirons l'attention du lecteur sur le fait que le concept de transformation-p en lui-même n'est pas nouveau puisque de nombreux auteurs ont déjà étudié ce sujet. Cependant la littérature, française en particulier, fait rarement une distinction terminologique entre le terme de dilatation mathématique et les termes de dilatation temporelle ou transposition fréquentielle du point de vue perceptif qui représentent pourtant des transformations très différentes. C'est pourquoi nous avons ressenti le besoin d'introduire les concepts et les notations qui vont suivre afin de lever toute ambiguïté quant à la terminologie employée.

1.2.1 Concept de "dilatation-p"

La "**dilatation temporelle sous contraintes perceptives**" notée "**dilatation-p**" (en anglais, les termes communément employés sont "time-scaling" ou "time-stretching") est définie comme un concept de dilatation temporelle du point de vue perceptif, la plupart des autres critères demeurant fixes par ailleurs, comme par exemple, la position des formants, la fréquence fondamentale et le timbre. La dilatation-p est donc une transformation qui donne la sensation d'un son ralenti ou accéléré sans modification des fréquences et sans défaut audible.

Le but de cette transformation est de modifier l'évolution temporelle des événements perçus, en se basant sur une fonction de dilatation, décrite en section 1.2.5. En d'autres termes, nous désirons que la version dilatée du signal acoustique soit perçue comme étant constitué de la même séquence d'événements acoustiques du signal original, mais reproduite sur une échelle temporelle dilatée [SVW94, Ver00].

Cependant, l'établissement de ce concept n'est pas sans soulever d'autres problèmes; il reste à trancher arbitrairement en ce qui concerne le comportement de certains caractères du son : les transitoires doivent-ils être dilatés ou conservés? Faut-il ralentir ou non les modulations d'amplitude et de fréquence? Il semble inévitable de se référer au mode de production d'un son

pour connaître les critères que l'on doit faire varier et ceux qui doivent rester constant. Ainsi, dans le cas de la voix, il s'agit a priori d'accélérer ou de ralentir les parties quasi-stationnaires, tout en conservant intacte la position des formants, la fréquence fondamentale ainsi que les transitoires.

Dans la suite de ce document, nous notons α le taux de dilatation, et nous le définissons de sorte que $\alpha > 1$ corresponde à une élongation temporelle (i.e. un ralentissement) et $\alpha < 1$ corresponde à une contraction temporelle (i.e. une accélération).

Nous notons l'opérateur associé à cette transformation Dp_α pour "dilatation-p de rapport α ".

1.2.2 Concept de "transposition-p"

La "**transposition fréquentielle sous contraintes perceptives**" notée "**transposition-p**" (en anglais, les termes communément employés sont "pitch-scaling" et "pitch-shifting") est un concept de transposition du point de vue perceptif, qui conserve les caractères temporels du signal. Cette transformation doit conserver l'aspect harmonique du signal, ce qui n'est pas le cas lorsque l'on réalise par exemple une modulation hétérodyne². La transposition-p est donc une transformation qui donne la sensation d'un son plus aigu ou plus grave sans modification de la durée et sans défaut audible.

Comme pour la dilatation-p, il faut trancher arbitrairement en ce qui concerne le comportement de certains caractères du son : les transitoires doivent-ils être transposés ou conservés? Faut-il transposer ou non les zones de résonance telles que les formants? Pour les sons de type impulsion-résonance, il semble naturel de conserver les transitoires ainsi que les formants, sans quoi la source sonore (voix ou instrument) est difficilement identifiable. Dans certains cas, on peut vouloir transposer les transitoires également.

De même, dans le reste de ce document, nous noterons α le taux de transposition, et nous le définissons de sorte que $\alpha > 1$ corresponde à une transposition vers les aigus et $\alpha < 1$ corresponde à une transposition vers les graves.

Nous noterons l'opérateur associé à cette transformation Tp_α pour "transposition-p de rapport α ".

1.2.3 Des problèmes "semi-duaux"

Il existe cependant des relations entre les contraintes régissant la dilatation-p et la transposition-p :

Pour un type de dilatation-p donné, il est possible de lui associer un type de transposition-p grâce à la transformation mathématique biunivoque R_α définie par l'équation 1.4. Cependant, ces deux types de transformation associées ne correspondent pas forcément à ce qu'on aurait pu s'attendre du point de vue perceptif.

Ainsi, dans le cas d'instruments de musique où l'on cherche à dilater les parties quasi-stationnaires et conserver les transitoires (les attaques restent généralement les mêmes quel que soit le tempo) pour simuler un instrumentiste jouant à un tempo différent (dilatation-p à transitoires conservés notée Dp_α^{tc}), la transposition-p associée à la dilatation-p après "rééchantillonnage" modifie le timbre des transitoires (transposition-p à transitoires transposés notée Tp_α^{tt}) puisque ceux-ci sont dilatés à la fois en temps et en fréquence. En effet, contracter l'axe temporel revient à étirer l'axe fréquentiel comme on peut le voir sur la figure 1.3, et l'entendre avec le son [4] correspondant à cette figure.

2. Ce décalage des fréquences, appelé en anglais "frequency shifting", est très simple à réaliser - multiplication par une exponentielle complexe - mais transforme un son harmonique en un son inharmonique. Une explication

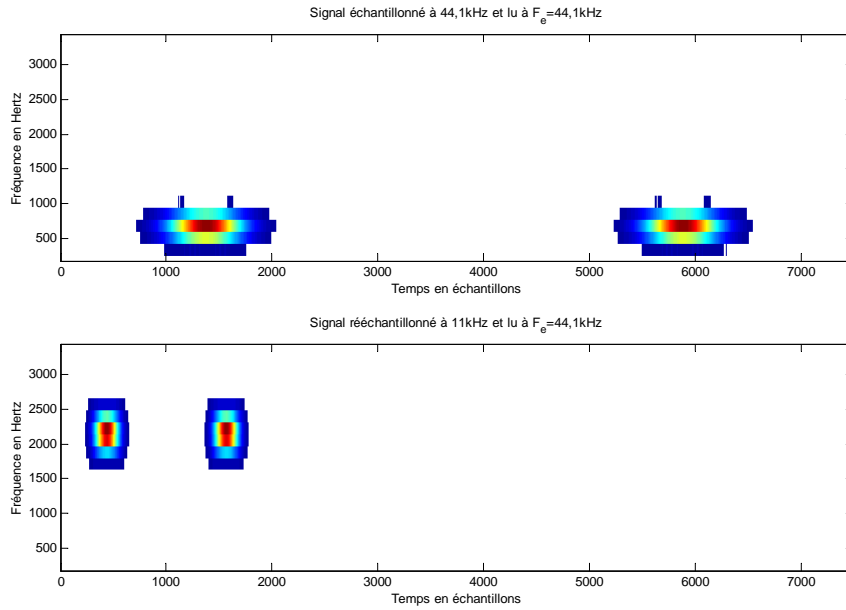


Figure 1.3 – Images temps-fréquence représentant un son composé de 2 transitoires, échantillonnés à 44,1 kHz et 11 kHz mais tous deux lus à 44,1 kHz

On a donc équivalence, à une opération de rééchantillonnage près, entre la "dilatation-p à transitoires conservés" et la "transposition-p à transitoires transposés" :

$$Dp_{\alpha}^{tc} \equiv R_{\alpha} [Tp_{\alpha}^{tt}] \quad (1.6)$$

On peut remarquer que pour obtenir un ralentissement ($\alpha > 1$), on doit transposer vers les aigus et rééchantillonner par un facteur α . La figure 1.4 montre ce type d'équivalence pour un taux de dilatation-p de 2.

D'autre part, la "transposition-p à transitoires conservés" (Tp_{α}^{tc}) correspond à une dilatation-p où ces derniers sont dilatés (Dp_{α}^{td}) :

$$Tp_{\alpha}^{tc} \equiv R_{1/\alpha} [Dp_{\alpha}^{td}] \quad (1.7)$$

On remarque que pour obtenir une transposition-p vers les aigus, ($\alpha > 1$), on doit ralentir le signal et le rééchantillonner par un facteur $1/\alpha$.

En conséquence de quoi, lorsque l'on a mis au point une méthode de dilatation-p, un simple rééchantillonnage permet d'obtenir la transposition-p duale. C'est pourquoi on ne cherchera pas à classer les méthodes selon le résultat final désiré puisqu'il y a deux manières de l'obtenir : indirectement ou directement (par l'utilisation du rééchantillonnage ou non).

En pratique, le rééchantillonnage est une technique qui peut n'introduire aucun défaut audible si elle est réalisée convenablement. Par conséquent, les artefacts engendrés par une méthode indirecte (avec rééchantillonnage) sont dus aux opérations de dilatation-p ou de transposition-p.

plus détaillée ainsi que des sons sont disponibles sur [Spr02].

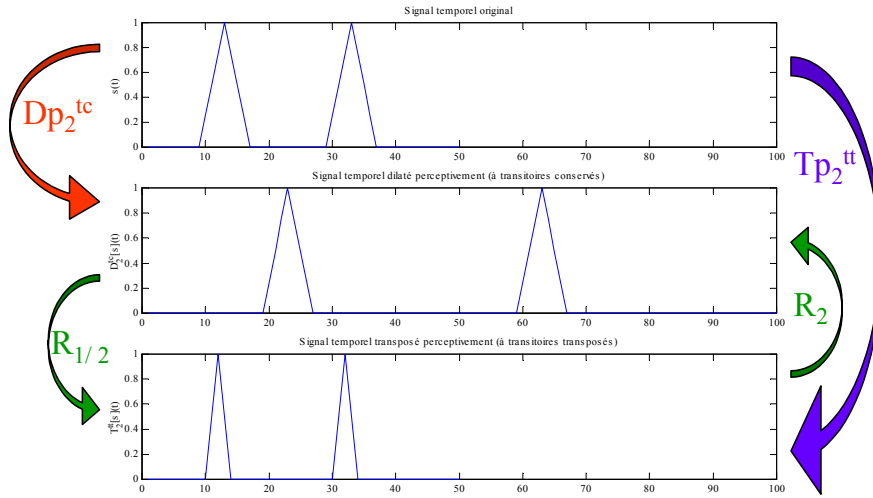


Figure 1.4 – "Semi-équivalence" entre dilatation-p et transposition-p

Il existe différents concepts de dilatation-p et de transposition-p selon le type de transformation-p désiré, à savoir les contraintes perceptives que l'on se fixe. Ces contraintes peuvent malheureusement être différentes selon la source sonore à traiter, bien que les caractéristiques des signaux soient similaires : par exemple dans le cas d'une dilatation-p, il est dans certains cas souhaitable de conserver le rythme original d'une modulation lente d'amplitude (comme le vibrato d'une voix [AD98] dont la fréquence est de l'ordre du Hertz) alors que dans la majorité des cas, le rythme doit être dilaté (comme les notes répétées à une fréquence du même ordre de grandeur que dans le cas précédent). Il semble donc impossible d'utiliser une transformation unique, valable pour tous les types de sons, car les contraintes peuvent varier d'une source à l'autre.

Le but de ce travail de thèse est donc de construire les opérateurs de dilatation-p idéale " Dp_{α}^{ideal} " et de transposition-p idéale " Tp_{α}^{ideal} ". Ce but semble difficile à atteindre par une méthode totalement automatique, mais nous nous emploierons à nous en approcher au mieux, en profitant de nos connaissances sur les caractéristiques de l'audition pour tenter de satisfaire au mieux la perception auditive (par l'introduction d'artefacts éventuels dans des zones inaudibles).

1.2.4 Prise en compte des spécificités de l'oreille

Dilatation-p et transposition-p ont comme unique but dans notre étude la transformation d'un signal audio, message qui s'adresse donc à l'oreille humaine. Or, l'oreille n'est pas un capteur idéal et possède ses spécificités qui font l'objet de recherches poussées dans le domaine de la psychoacoustique. La psychoacoustique est une branche de la psychophysique (discipline qui étudie les relations entre les stimulus physiques et les sensations engendrées), où le stimulus est une vibration mécanique de l'air, et la sensation est auditive [Can00].

Du fait de l'absence de solution mathématique triviale et de la multiplicité des solutions, les méthodes de transformation introduisent des artefacts. Cependant, on peut tirer parti des caractéristiques de l'oreille afin de se rapprocher au mieux d'une transformation perceptivement parfaite. Tous les algorithmes de dilatation-p et de transposition-p conduisent donc à faire des compromis et à utiliser des "astuces". L'oreille, dernier maillon de la chaîne sonore (la bande-son est en fin de compte destinée aux spectateurs), est l'ultime juge de la qualité des compromis effectués.

1.2.5 Fonctions de dilatation et transposition

Fonction de dilatation

La dilatation-p peut être spécifiée par une fonction de dilatation $D(t)$ [ML95, Ver00]. Cette fonction est une sorte de "loi de correspondance" entre l'échelle temporelle originale et l'échelle temporelle dilatée. Cette fonction indique que l'événement sonore, produit à l'instant t dans le signal original, devra être entendu à l'instant t' dans le signal dilaté. Les algorithmes mis en pratique ne répondent toutefois qu'à des approximations de cette loi idéale.

Pour un taux de dilatation constant $\alpha(t) = \alpha_0$, la fonction de dilatation est linéaire :

$$t \rightarrow t' = D(t) = \alpha_0 t$$

Elle correspond à un changement global de tempo, comme c'est le cas dans l'application de transfert cinéma/vidéo. Pour $\alpha_0 > 1$, la dilatation-p correspond à un ralentissement temporel du signal original et pour $0 < \alpha_0 < 1$, la dilatation-p correspond à une accélération temporelle du signal original.

Pour un taux de dilatation variant dans le temps $\alpha(t)$, la fonction de dilatation devient non-linéaire. Elle peut être utile, par exemple, pour spécifier le synchronisme entre la voix doublée et les mouvements des lèvres d'un acteur.

Dans ce cas, la fonction de dilatation est dérivée de $\alpha(t)$ par la formule suivante :

$$t \rightarrow t' = D(t) = \int_0^t \alpha(u) du$$

Dans tout cet exposé, nous supposons que le taux de dilatation, noté α , reste constant. Cette hypothèse nous est dictée à la fois par un souci de simplification d'écriture et par l'utilisation pratique qui est faite dans notre application.

Fonction de transposition

La transposition-p peut également être spécifiée en définissant une fonction de transposition. Il s'agit d'une correspondance entre l'échelle fréquentielle originale et l'échelle fréquentielle dilatée.

Soit $P(t)$ la période d'une composante sinusoïdale présente à l'instant t dans le signal original, la fonction de transposition est définie par

$$t \rightarrow T(t) = \frac{P(t)}{\alpha(t)}$$

Cette fonction indique que la composante sinusoïdale devra être entendue à l'instant t à une fréquence $\frac{1}{T(t)}$ dans le signal transposé.

Dans tout cet exposé, nous supposons que le taux de transposition, noté α , reste constant. Nous prenons la même notation que pour la dilatation car nous n'utiliserons pas simultanément les deux transformations.

Pour $\alpha > 1$, la transposition-p correspond à une augmentation des fréquences du signal original (le son devient plus aigu) et pour $0 < \alpha < 1$, la transposition-p correspond à une diminution des fréquences du signal original (le son devient plus grave).

1.3 Des contraintes particulières pour le transfert cinéma/vidéo

1.3.1 Contraintes technologiques

L'harmoniseur, destiné à restituer les hauteurs des bandes-son des films lors du passage d'un standard vers un autre (cinéma vers vidéo ou inversement), doit respecter un certain nombre de contraintes technologiques pour pouvoir être utilisé par les professionnels du son auxquels il est destiné (techniciens des studios de post-production cinématographique, ingénieurs du son, mixeurs...).

Format numérique des entrées/sorties

L'ensemble de l'équipement des studios de post-production tend à se généraliser vers une solution "tout numérique". Sans rentrer dans le débat son numérique - son analogique, il semble évident qu'un appareil aux entrées-sorties analogiques au sein d'une chaîne numérique ne peut qu'apporter des dégradations au son, en raison des conversions analogique/numérique et numérique/analogique (au nombre de 2 chacune si le traitement à l'intérieur de la machine est réalisé en numérique). C'est pourquoi il est aujourd'hui souhaitable d'utiliser des appareils possédant des **entrées/sorties numériques**.

Adaptation au son multicanal

Avec la généralisation du son multicanal dans les cinémas et chez les particuliers ("home-theatre"), il est nécessaire que l'appareil de traitement soit également adapté à ce mode de restitution sonore. Les techniques de spatialisation reposent sur les principes de différences interaurales de niveau et de phase [Bla97]. Généralement, les mixeurs de films utilisent uniquement des différences de niveau entre canaux pour mettre les sons en espace, et ils n'introduisent donc aucune différence de phase. Il est alors important de ne pas modifier cette synchronisation des phases, sous peine de ruiner les efforts mis en œuvre pour effectuer la spatialisation de scènes sonores.

Les canaux en question peuvent être au nombre de 4 pour un mixage "LtRt" ("Left total - Right total"³) destiné aux formats amateurs "Dolby Surround ProLogic" et professionnels "Dolby Stereo", 6 pour un mixage 5.1 (5 canaux discrets plus un caisson de basses-fréquences) destiné aux formats amateurs "Dolby Digital" et professionnels "Dolby SR-D" ("Spectral Recording - Digital") ou "DTS" ("Digital Theater Systems"), 7 pour un mixage 6.1 (6 canaux discrets plus un caisson de basses-fréquences) destiné au format "Dolby Digital Surround EX", 8 pour un mixage 7.1 (7 canaux discrets plus un caisson de basses-fréquences) destiné au format "SDDS". Pour plus de détails on pourra consulter [Bes98], [Dol02a], [SDD02], [DTS02]. De plus, certains formats nécessitent un matriçage de plusieurs canaux, qui ne souffrent aucun décalage temporel relatif: c'est le cas du mixage "LtRt".

L'harmoniseur doit donc pouvoir fonctionner simultanément sur les différents canaux d'une bande-son de film **sans en modifier les relations de phase**.

Fonctionnement en temps-réel

Nous appelons "traitement temps réel" un traitement dont le débit des échantillons en sortie correspond au débit des échantillons en entrée. Il faut évidemment un certain temps, aussi petit

3. "LtRt" est un encodage utilisé dans les formats Dolby Surround et permettant de regrouper 4 canaux en seulement 2.

soit-il, à un échantillon d'entrée pour ressortir traité. Cette durée, que l'on appelle "temps de latence" (retard dû au traitement), est généralement mesurée en millisecondes.

Un temps de latence faible, dont l'ordre de grandeur est inférieur à la milliseconde, est compatible avec une utilisation interactive entre un musicien et son instrument ayant subi un traitement sonore. Des temps de latence élevés, supérieurs à la dizaine de millisecondes, ne permettent plus une utilisation musicale en direct, mais imposent quand même des contraintes à la structure des algorithmes, sans oublier la puissance de calcul qu'il faut aussi répartir convenablement.

Le fonctionnement en temps-réel permet une certaine souplesse aux utilisateurs des studios de post-production cinématographique, ainsi qu'un gain de temps non négligeable (la bande son est totalement traitée dès que le lecteur a terminé d'émettre). En effet, un tel traitement permet d'écouter immédiatement (moyennant un temps de latence faible) et en continu le résultat sonore, contrairement aux systèmes temps-différé où le traitement (dont la durée est souvent supérieure à la durée du son) doit être effectué d'un bloc sur toute la bande-son avant de pouvoir être contrôlé. Ainsi, les utilisateurs peuvent vérifier au fur et à mesure la qualité de la bande-son, éventuellement s'arrêter en cas de problème, reprendre à un autre endroit...

Ils évitent également la fastidieuse opération de transfert du support original (disque Magnéto-Optique, cassette numérique multipiste type "DA88"...) vers le support de traitement (disque dur du système de traitement).

Bien qu'un temps de latence de l'ordre de la seconde ne soit pas un inconvénient majeur pour les utilisateurs concernés, **l'harmoniseur doit donc fonctionner en temps-réel**, et ce, malgré les contraintes techniques mais aussi algorithmiques qui en découlent.

Adaptation du taux de transposition (-4% et +4,2%)

Les studios de post-production doivent corriger la transposition en fréquence due aux rapports de vitesse 25/24 et 24/25. Par conséquent, leurs besoins en terme de taux de transposition se situent à -4% et +4,2%. On remarque au passage qu'une augmentation de fréquence d'un demi-ton correspond à un taux de transposition de $2^{(1/12)}$, soit environ 6%, ce qui signifie que les taux de transposition utilisés sont légèrement inférieurs au demi-ton.

Il est important de connaître les valeurs des taux habituellement utilisés, car certaines méthodes algorithmiques sont mieux adaptées à certains taux. De plus, l'algorithme peut être optimisé pour ces valeurs, même si parfois il n'existe pas de problème théorique pour une utilisation avec d'autres rapports de transposition (on se heurte alors plutôt à des problèmes de qualité sonore). La qualité de traitement de l'harmoniseur doit donc être **optimisée pour les taux de transposition de -4% et +4,2%**.

Cependant, nous ne nous limitons pas à l'étude de ces deux rapports de transposition et nous nous intéressons à tous les rapports compris entre -20% et +20%. Nous devons donc déterminer un pas de réglage que nous fixons à 0,1%. Ce choix nous est dicté par le fait qu'il correspond à une différence fréquentielle de moins de 2 cents⁴; or des études ont montré [WJG77] que la plus petite différence de fréquence notable était en moyenne d'environ 3 cents pour un son pur autour de 500 Hz présenté à 80 dB SPL, valeurs auxquelles l'oreille est la plus sensible à ce type de variation.

4. Un cent correspond à un centième de demi-ton (il y a donc 1200 cents dans un octave), soit un rapport de fréquence de $\sqrt[1200]{2} \simeq 1,000578$.

De plus, l'erreur introduite lorsque l'on approxime $25/24$ par un rapport de 4,2%, est d'environ 0,034%, soit inférieure à 1 cent et donc théoriquement inaudible.

Nous n'étudions ici uniquement des transformations-p (dilatation-p ou transposition-p) dont les taux sont fixes dans le temps, bien que toutes ces méthodes soient théoriquement compatibles avec des taux variables.

Interface utilisateur simple

L'interface homme-machine doit être simple dans sa compréhension et son utilisation. Au-delà de l'aspect ergonomique, cette contrainte de simplicité affecte le mode de fonctionnement de l'algorithme, qui ne doit posséder aucun paramètre lié à la connaissance a priori du signal : **un seul et unique pré-réglage** doit être capable de traiter n'importe quel son.

En revanche, pour ce qui est du format multicanal, connu de l'utilisateur avant le lancement du traitement, il est autorisé d'introduire certains paramètres afin d'optimiser le fonctionnement en fonction du matériau à traiter.

1.3.2 Contraintes de qualité sonore

Les contraintes de qualité sonore pour une transformation-p telle que la dilatation-p sont très fortes, puisqu'il s'agit de traiter dans son intégralité la bande-son d'un film, avec toutes les préoccupations artistiques et techniques que cela implique.

Respect de la totalité des sons (et principalement la voix)

La difficulté principale pour cette application réside dans le traitement simultané de toutes les sources sonores constituant la bande-son, que ce soit la parole, la musique, les bruitages, les ambiances... Il faut donc respecter sur chacun des sons toutes les contraintes énumérées dans la suite. Or, les sources sonores indépendantes n'étant pas accessibles, on doit effectuer le traitement sur la totalité de la bande-son : les contraintes deviennent alors souvent incompatibles. C'est pourquoi on devra réaliser des compromis.

Cela implique également qu'il faut éviter les paramètres de contrôle que l'utilisateur pourrait choisir en fonction du matériel à traiter, puisqu'une fois le traitement lancé, n'importe quel type de son est susceptible d'apparaître.

Respect des formants

Les formants sont des zones fréquentielles de résonance, caractéristiques de certains instruments et surtout de la voix (voir la définition de "formant" donnée dans [UQA96]). Ceux-ci subissent un déplacement lors de la dilatation temporelle dû au changement de vitesse (voir figure 1.5). Cette modification formantique rend plus difficile la reconnaissance d'un instrument ou d'une voix particulière. Ainsi, il est nécessaire, lors de l'étape de transposition-p réalisée par l'harmoniseur, de replacer ces formants à leurs emplacements d'origine.

C'est sans aucun doute ce type de défaut qui pousse les mixeurs de films à utiliser un harmoniseur, car un déplacement de 4% des formants peut suffire à rendre une voix méconnaissable, comme on peut s'en rendre compte avec les exemples sonores [5, 6] et [8, 9].

Respect du timbre

Le timbre est parfois défini comme étant l'attribut de la sensation auditive qui permet à un auditeur de distinguer deux sons qui ont la même hauteur tonale et la même sonie [ANS60].

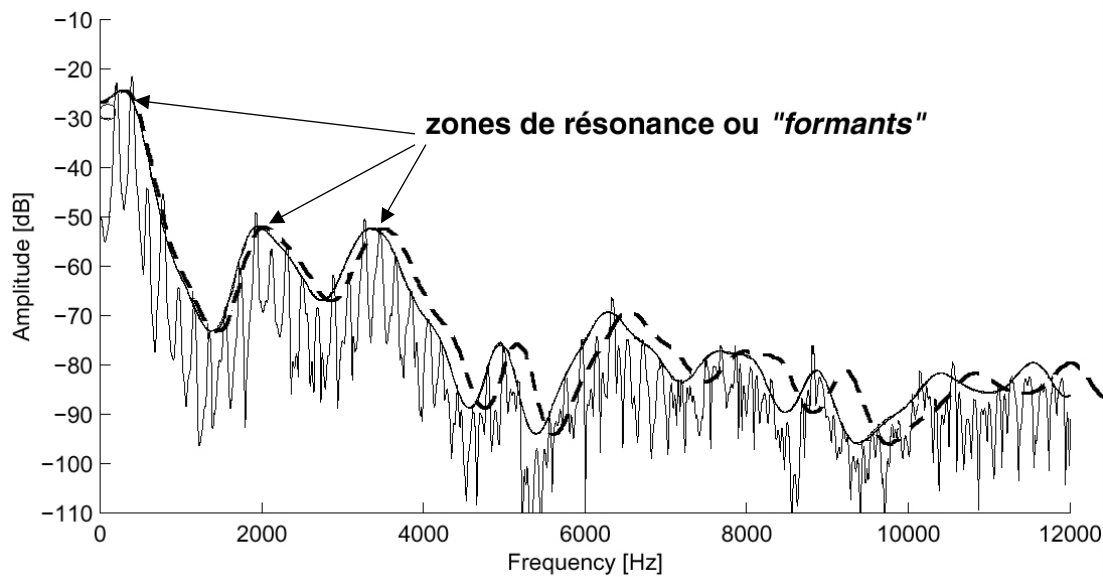


Figure 1.5 – *Formants d'un son de voix original (traits pleins) et formants déplacés (pointillés) dû à une accélération de 4%.*

Le timbre dépend principalement du spectre du stimulus, mais il dépend aussi de l'enveloppe spectrale, de l'enveloppe temporelle, et des variations de chacune d'elles [PMH00].

Cette notion, dont les formants font partie, englobe également la signature des transitoires et des attaques. En effet, le contenu spectral de ces types de sons varie avec l'accélération ou le ralentissement du film. Il est donc important de restituer ces caractéristiques fidèlement après un traitement d'harmonisation.

Respect du rythme

Il s'agit non pas de conserver le tempo puisque l'on cherche à dilater le son dans le temps, mais de respecter la structure rythmique de l'original au sens où les rapports de durées des notes et des silences sont préservés. Ainsi, des claquements réguliers de castagnettes doivent rester réguliers.

De plus, il semble important que la synchronisation entre l'image et le son soit respectée, ce qui renforce l'idée que la dilatation-p après harmonisation doit être répartie régulièrement le long du film. C'est pourquoi les méthodes de dilatation-p non linéaires (dont le taux varie au cours du temps) ne seront pas étudiées spécifiquement ici. Il a été montré qu'elles donnent cependant de bons résultats sur la voix [CWS98, HG00].

Respect de la hauteur tonale

L'utilisation d'un harmoniseur lors du transfert cinéma/vidéo permet de retrouver les hauteurs des sons d'origine, malgré des durées différentes. Sans harmoniseur, chacune des notes est transposée d'un même rapport, ainsi les intervalles de fréquence entre notes sont conservés. Quel que soit le tempérament utilisé⁵, la musique conserve toujours sa "couleur" après la transposition fréquentielle, ce qui n'est pas le cas d'une transposition de la partition : par exemple, la "Tocatta et Fugue en Ré mineur" de J.S. Bach jouée en Ré bémol mineur ou en Ré dièse mineur

5. Pour une discussion sur le tempérament, consulter [Ass85, Bai95].

possèdera une couleur différente (sauf si le tempérament de tous les instruments est strictement égal [AB95]).

Ainsi, le désagrément de ne pas retrouver la note d'origine n'affecte généralement que des personnes possédant (consciemment ou non) "l'oreille absolue", c'est-à-dire des auditeurs ayant une mémoire très précise de la tonalité. Pour les autres, l'utilisation d'un harmoniseur dans le but de retrouver la hauteur tonale reste anecdotique, surtout pour des rapports de transposition aussi faibles que 4%.

1.4 D'autres applications des transformations-p

1.4.1 Historique des besoins en transformation-p

Le besoin en dilatation-p s'est fait ressentir très tôt dans une application cinématographique. Un brevet de 1935 [Fre35] décrit une machine permettant de synchroniser l'enregistrement sonore optique avec les images d'un film tourné avec une caméra haute-vitesse. Les supports optiques des medias sonores et visuels étant de taille différentes, il est nécessaire d'adapter la longueur de la pellicule son à la longueur de la pellicule image, pour pouvoir ne produire qu'un seul et unique film, et ce, sans modifier les fréquences sonores originales. Selon l'auteur, cet appareil fonctionne aussi bien pour la voix que pour d'autres types de matériaux sonores comme la musique.

A cette époque, la vidéo n'existait pas donc le problème spécifique "24/25" non plus, mais cette invention permettait d'effectuer des dilatations-p pour d'autres applications que celui des caméras haute-vitesse comme l'étude des détails d'un film [Fre35].

Un brevet encore plus ancien [FZ28], datant de 1928, décrit une méthode permettant de réduire la largeur de bande requise pour transmettre des signaux électriques, dont le principe est à la base de beaucoup de méthodes de dilatation-p. Ainsi, les Tempophon [Man85], Phonogènes [Bat98] et autres appareils à têtes tournantes comme ceux de Fairbanks, ne semblent être que des adaptations (et seulement parfois des améliorations) du principe établi en 1928. Toutes ces méthodes sont décrites en détail au chapitre 2.

Nous venons de voir que le besoin en dilatation-p et en transposition-p n'était pas nouveau puisque des solutions à ce problème ont été proposées dès 1928. Depuis, le nombre d'articles et de brevets concernant ce sujet a augmenté considérablement, sans doute en réponse aux nouvelles applications qui ont pu voir le jour. Nous avons tenté ici d'établir une liste documentée, mais certainement pas exhaustive, de ces applications.

1.4.2 Applications à une source vocale

Apprentissage d'une langue étrangère

L'apprentissage d'une langue étrangère peut être facilitée par l'écoute d'un locuteur dont le débit s'adapte aux progrès de compréhension de l'étudiant [Mal79, Lar98]. Les élèves peuvent plus facilement imiter les gestes articulatoires de la production de parole lorsque la voix est ralentie. De plus un enregistrement de leur propre voix leur permet de corriger leurs erreurs d'articulation [Sco67]. D'autre part, des études ont montré qu'écouter deux fois à une vitesse double un matériau sonore d'apprentissage était plus efficace que l'écouter une seule fois à vitesse normale [Sti69].

Lecture pour les aveugles

Des appareils de compression temporelle de la voix ont déjà été utilisés dans des programmes d'éducation [Fou64]. D'autre part, la compréhension d'une phrase prononcée plus rapidement qu'on ne peut le réaliser physiquement est possible; ainsi comme des taux d'accélération jusqu'à 2 fournissent toujours une bonne intelligibilité et une bonne compréhension, les textes peuvent être lus beaucoup plus rapidement que ne le fait la lecture en braille [FEJ59, Sco67, BP59, Lee72, Mal79, Aro92, Aro97, Lar98]. Les bibliothèques sonores peuvent donc utiliser ce type d'outil.

Adaptation de la voix pour les malentendants

Certaines personnes ont des pertes auditives dans les hautes fréquences (au-dessus de 1500 Hz). L'utilisation d'une transposition-p vers les basses fréquences leur permet, avec de l'entraînement, de comprendre le message original sans que le tempo soit altéré et malgré la réduction de bande passante [TB61, Sco67].

Apprentissage à la lecture rapide

L'apprentissage de la lecture rapide peut être amélioré si le sujet lit un texte en même temps qu'il écoute la voix accélérée [OFW56, Sco67].

Reconnaissance de la parole

Des systèmes de reconnaissance de la parole possèdent une phase de pré-traitement qui consiste à normaliser temporellement les mots. Cette étape est réalisée grâce à la dilatation-p [Eng77, Mal79].

Modification de la prosodie

Certaines méthodes de dilatation-p permettent la modification de la prosodie d'une voix naturelle [MC90]. Ces méthodes sont utilisées pour améliorer la qualité des systèmes de synthèse de la parole à partir du texte ("Text-to-speech").

Répondeurs téléphoniques, dictaphones et serveurs vocaux

L'accélération ou le ralentissement de la voix est utile dans les systèmes de répondeurs téléphoniques [Max80, VR93, Hej90] (accélération pour la recherche rapide des messages [RW85], ralentissement pour améliorer l'intelligibilité [TL00]), les serveurs d'informations vocales [Ver00] et également les dictaphones [VR93] (synchronisation du débit vocal à la vitesse de frappe).

Outil d'études psychoacoustiques

La dilatation-p fournit un moyen d'altérer la base temporelle des signaux sans modifier sensiblement leurs spectres. L'investigation d'un certain nombre de problèmes temporels liés à l'audition peut être effectué grâce à cette technique [Sco67]. Par exemple, des travaux ont été effectués sur l'intelligibilité de la parole lorsque l'on ôte des segments plus ou moins longs de signal [ML50, Gar53] (il semble que 60% du signal peut être ôté sans que l'intelligibilité ne chute en dessous de 80% [Gar53]). On parle alors de *redondance temporelle* [FEJ54].

Réduction de largeur de bande et compression de données

Grâce à la transposition-p de rapport $\alpha < 1$, il est possible de réduire la largeur de bande du signal vocal sans affecter la durée du signal [Mal79]. Cette technique permet par exemple d'augmenter la capacité des lignes de transmission conventionnelles [FEJ54, Sch66b, FG66]. Historiquement, cette réduction de largeur de bande permettait aussi de limiter l'atténuation (proportionnelle à la fréquence) due aux transmissions filaires, et d'utiliser une sélectivité plus grande au niveau du récepteur entraînant une réduction conséquente de la quantité d'interférence statique reçue [FZ28]. Pour des applications de codage à bas débit, cette technique ne permet que de faibles taux de compression mais elle peut être appliquée en combinaison avec d'autres types de techniques de codage plus classiques [CCJ83, MCC93, MEJ86, Lar98].

De la même manière, il est possible d'utiliser la dilatation-p pour réaliser du codage à bas débit. Le signal est alors accéléré au niveau de l'émetteur avant d'être transmis, puis ralenti au niveau du récepteur [RW85, WW88].

Cette technique est également utilisée pour transmettre le signal vocal par un réseau IP ("Internet Protocol") réalisant ainsi un bon compromis entre la latence et la perte de paquets

[LFG02].

Amélioration du rapport signal à bruit

Selon Wayman et Wilson [WW88], la contraction temporelle suivie de l'expansion temporelle peut être utilisée comme un filtre de corrélation pour améliorer le rapport signal à bruit dans le cadre de signaux vocaux bruités.

Amélioration de la communication en plongée hyperbare

La vitesse de propagation du son dans l'hélium présent dans les voies respiratoires donne aux plongeurs une voix à caractère nasal qui diminue l'intelligibilité. Ce phénomène est communément appelé "effet Donald Duck" en rapport au personnage des dessins animés. La parole est déformée encore davantage par la densité du mélange respiratoire, qui acquiert certaines des propriétés d'un liquide. Pour compenser cette modification du son, due au déplacement des formants (mais pas des fréquences) des corrections effectuées sur la fonction de transfert du conduit vocal, soit dans le domaine fréquentiel ou temps-fréquence [Meu90, AKM92], soit directement sur la fonction d'auto-corrélation [BHA98].

Cette application n'est pas réellement issue des méthodes de transposition-p puisque ni la durée, ni les fréquences ne sont modifiées; seuls les formants sont altérés. Cependant, on la cite ici du fait des rapports étroits qu'il existe entre fréquences et formants.

1.4.3 Applications à une source sonore complexe

Edition sonore

Deux composantes brèves faiblement espacées dans le temps issues de sons complexes courts sont difficiles à détecter par l'oreille (comme par exemple la fermeture d'une agrafeuse). La dilatation-p permet alors de discriminer ces composantes séquentielles [QDH95].

De même, la localisation des limites phonétiques est simplifiée lorsque le signal vocal a été dilaté [Sco67].

D'autre part, la dilatation-p peut être utilisée par les musiciens qui souhaitent transcrire de la musique enregistrée. Un ralentissement efficace leur permet de distinguer par exemple les notes rapides d'un solo de guitare. C'est également un moyen d'avoir le plaisir de jouer avec son artiste préféré, mais à un rythme moins élevé que l'original.

Production audio et vidéo

Une fonction de recherche rapide est parfois nécessaire dans les appareils de lecture audio [SVW94] et vidéo [IK99, APB⁺00]. Elle permet de repérer le passage désiré à haute vitesse sans modification des fréquences. Cette application requiert des taux de dilatation élevés, mais se satisfait généralement d'une qualité médiocre.

La dilatation-p permet également de réaliser de petits ajustements de tempo pour corriger des imperfections d'interprétation [SVW94]. Cette application utilise cette fois des taux de dilatation faibles, mais requiert des modifications de très haute qualité.

La transposition-p peut, quant à elle, être utilisée pour ajuster la hauteur tonale d'un enregistrement avant de l'intégrer dans un mixage, si les instruments ne sont pas tous accordés entre eux. Le taux de transposition peut être fixe ou variable dans le temps et il est généralement inférieur à 6% (environ un demi-ton)[Lar93]. Certaines machines d'enregistrement multipiste sont d'ailleurs dotées de cette caractéristique [Lar98].

Création musicale

Les outils de dilation-p et de transposition-p sont très usités dans la création musicale [Tru94, Ris99] et se retrouvent sous forme de processeurs d'effets musicaux ou d'éditeurs de sons logiciels [LD99a, Fav01]. Les compositeurs sont intéressés par un contrôle indépendant du temps et des fréquences [Lar98]. Ils s'approprient aussi souvent des effets étranges résultant d'artefacts incontrôlés (comme par exemple l'effet "choral" [Moo78]).

Selon [Roa96], l'appareil mécanique "Tempophon" de la compagnie allemande Springer avait été utilisé pour traiter les sons parlés dans la pièce de musique électronique de 1963 de H. Eimert "Epitaph für Aikichi Kuboyama". Selon [Bod84], le premier harmoniseur numérique commercial fut le H910 de Eventide, commercialisé en 1975.

Les "Disc-Jockey" peuvent faire varier le tempo d'une musique durant une performance sans introduire de distorsion audible grâce à des lecteurs CD professionnels [TL00], ou bien enchaîner continûment deux musiques dont les tempi sont différents [DDPZ02], ou encore de faire parler un "rappeur" bien plus vite qu'il ne pourrait le faire en réalité.

De nombreux logiciels de création sonore "par boucles" (les segments sonores sont répétés automatiquement) utilisent aussi la dilatation-p afin de faire tenir n'importe quel segment sonore dans les limites temporelles imposées par le tempo.

Synthèse par échantillonnage

Certains types de synthétiseurs, basés sur la technique de l'échantillonnage, utilisent des segments de sons enregistrés. Le principe de cette méthode est de reconstituer l'étendue des sons d'un instrument à partir d'un nombre restreint d'échantillons de notes stockées. Ainsi, il peut être fait appel à la dilatation-p pour allonger les parties entretenues des sons (bouclage dans la partie entretenue), et à la transposition-p pour synthétiser des notes différentes de la note originale [Lar98, Mas98, Bou02].

Publicité

A l'aide de la dilatation-p, il est possible de fournir plus d'information publicitaire en une durée donnée, et un matériau publicitaire donné peut être ajusté à une période de temps allouée [FEJ54, FEJ59].

Tatouage numérique

Le tatouage numérique audio ("audio watermarking" en anglais) est une technique de marquage consistant à insérer une signature invisible à l'intérieur d'un fichier son permettant ainsi de lutter contre la fraude et le piratage. Certains algorithmes utilisent des méthodes de dilatation-p afin d'introduire les données de marquage [MT01].

Multimédia

Une grande quantité d'information numérique est présente sur Internet. Les entreprises y publient des allocutions et des formations, les universités mettent en ligne leurs cours sous forme de vidéo, les médias y proposent du contenu audiovisuel. Toutes ces informations peuvent être distribuées à des vitesses adaptées à chaque personne, selon sa capacité ou son temps disponible, grâce à des algorithmes de dilatation-p [HG00, APB⁺00]. Cette technique permet de gagner du temps.

Synchronisation audio/vidéo

La dilatation-p peut être utilisée pour ajuster la durée d'un enregistrement de manière à ce qu'il corresponde exactement à la durée de la séquence sans modifier son contenu spectral [Lar93, VR93, Lar98]. Cela évite par exemple d'avoir à ré-enregistrer la voix d'un acteur pour faire correspondre l'image des mouvements de lèvres au son [Bon00b], ou encore ré-enregistrer

un orchestre quand on cherche à synchroniser la musique aux mouvements des personnages du film [TL00].

Cette application est valable pour de faibles taux de dilatation dans la post-production audiovisuelle et cinématographique, mais également pour des taux plus importants dans les outils de recherche rapide pour les systèmes vidéo [Wat94, KIS⁺99].

1.5 Technologie actuellement disponible

Le problème de transfert de la bande-son n'est ni récent, ni rare puisque les professionnels du cinéma et de l'audiovisuel y sont confrontés par exemple à chaque fois qu'ils doivent convertir un film pour la télévision ou le DVD ("Digital Versatile Disc"). Ces derniers ont un choix à faire : soit ils ne traitent pas la bande-son, qui est alors altérée d'une transposition de l'ordre de 4% lors de la restitution et donc modifie les formants (c'est encore actuellement le cas pour nombre de productions françaises), soit ils utilisent une machine commerciale leur permettant de compenser cette transposition.

De nombreuses machines effectuant une transposition-p sont présentes sur le marché. Cependant, la plupart d'entre elles ne répondent pas aux critères technologiques pour cette application. D'autre part, la majorité des appareils introduisent des artefacts audibles largement supérieurs aux limites acceptables pour ce type d'application. Les algorithmes utilisés dans les stations de travail n'ont pas véritablement retenu l'attention des professionnels de l'audiovisuel. De plus, la solution "station de travail" est loin d'être la plus souple et la plus efficace pour ces derniers. Les algorithmes des logiciels d'application vocale ne sont pas adaptés à traiter des bandes-sons complètes de films.

L'annexe B dresse un très large panorama des machines et logiciels qui effectuent de la dilatation-p et/ou de la transposition-p en 2003.

Aucun des algorithmes existants ne semble être totalement adapté à nos contraintes : certains ne peuvent conserver les relations de phase pour un traitement multicanal, d'autres sont loin d'être temps-réel, et de surcroît, la totalité d'entre eux ne donnent pas une entière satisfaction du point de vue de la qualité sonore (il existe toujours des sons qui posent des problèmes).

Jusqu'à présent, une seule machine professionnelle semblait donner satisfaction pour des signaux stéréo : il s'agit de la Lexicon 2400, conçue en 1987 par Dattorro [Dat87]. Celle-ci respecte les contraintes de temps-réel et de taux de transposition, mais ne peut satisfaire aux nouvelles contraintes technologiques de son multicanal et d'entrées/sorties numériques. De plus, cette dernière montre ses limites sur des sons de type impulsif [11, 12], des sons complexes polyphoniques [15, 16], des sons inharmoniques comme une simple note de piano [18, 19] ou encore certains sons de voix parlée [21, 22].

C'est pourquoi une demande mercantile s'est créée⁶, alimentant ainsi une mise en œuvre des nouvelles technologies (entrées/sorties au format numérique AES/EBU [RW99]), mais aussi des recherches scientifiques portant sur l'algorithme, qui doivent non seulement déboucher sur une qualité sonore accrue, mais aussi s'adapter au son multicanal.

6. Au début de cette étude, les machines de T.C. Electronic (dont la qualité sonore de l'algorithme de transposition-p se révéla être décevante) et de Dolby n'étaient pas encore sur le marché.

Chapitre 2

Classification des méthodes

La classification proposée ici a pour but de rassembler sous un même formalisme toutes les méthodes de dilatation-p et de transposition-p que l'on a pu rencontrer jusqu'à présent durant notre étude.

Il s'agit en quelque sorte d'un état de l'art présenté sous un éclairage particulier et personnel, mais nous espérons que ce travail sera interprété comme allant au-delà d'un simple travail bibliographique.

Nous souhaitons en effet, par cette typologie, faciliter la compréhension des différentes techniques, et surtout mettre en relief les relations parfois très étroites qui existent entre les différentes méthodes, relations qui sont loin d'être explicites dans les divers articles.

Nous avons réalisé la classification la plus exhaustive possible, en espérant que le lecteur saura y retrouver les méthodes qu'il connaît, et que les méthodes non encore découvertes y trouveront leur place tout naturellement.

Cependant, un autre but de ce chapitre est aussi de simplifier les approches pour les futures recherches en dilatation-p / transposition-p.

En effet, cette classification facilite l'amélioration d'une méthode existante en s'inspirant des idées exploitées dans d'autres méthodes comme c'est le cas pour les méthodes proposées au chapitre 3.

2.1 Préambule à la classification

2.1.1 Transformation-p d'une représentation

Les transformations-p sont des transformations qui agissent sur la représentation temporelle du signal audio.

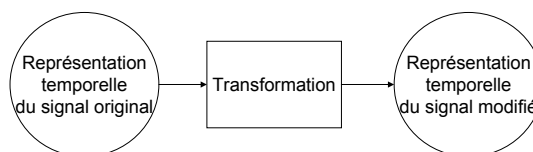


Figure 2.1 – *Illustration de la transformation de la représentation temporelle*

Leur mise en œuvre s'effectue soit directement sur la **représentation temporelle** du signal

(développé au chapitre 2.2, voir figure 2.1), soit au travers de **représentations intermédiaires** (développées aux chapitres 2.3 et 2.4, voir figure 2.2), mieux adaptées à la modification de certains types de signaux. La représentation intermédiaire est obtenue à l'issue d'une étape d'**analyse**, et l'étape de **synthèse** fournit la représentation temporelle modifiée après avoir effectué la **transformation**.

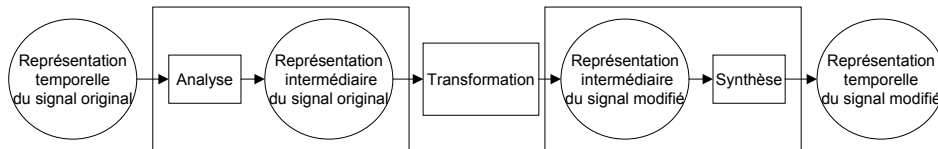


Figure 2.2 – Illustration de la transformation de la représentation intermédiaire

Les représentations intermédiaires se classent habituellement selon 2 types : représentations *paramétriques* et *non-paramétriques*. Cette distinction est faite selon que la représentation repose sur un modèle ou non. De manière générale et dans le cas d'un signal arbitraire, les représentations paramétriques perdent de l'information, alors que ce n'est pas le cas des représentations non-paramétriques. Nous différencions dans la suite ces deux types de représentation en nous attachant à justifier le choix que l'on fait de nous intéresser uniquement aux représentations non-paramétriques.

La figure 2.3 schématise la notion que l'on a du concept d'analyse/synthèse en s'appuyant sur la distinction réalisée entre les différentes représentations.

Représentation paramétrique

La représentation paramétrique repose sur une utilisation explicite d'un modèle physique ou de signal, d'où sont tirés des **paramètres**. Il s'agit d'estimer au mieux l'ensemble de ces paramètres potentiellement variables au cours du temps, symbolisés par le vecteur $\vec{p}(t)$, afin que le signal synthétisé à partir des paramètres d'analyse non modifiés soit le plus "proche" possible du signal original (au sens de la perception sonore). Si le modèle est adapté au signal et que l'estimation des paramètres d'analyse $\vec{p}(t)$ est convenable, il est alors possible de resynthétiser correctement le signal original.

L'intérêt de ce type de représentation réside dans la simplicité, l'efficacité, et surtout la qualité des transformations *lorsque le modèle est bien adapté au signal*. Ces représentations sont réputées pour diminuer la quantité d'information nécessaire à décrire fidèlement un signal donné (codage de la parole par exemple).

Les modèles physiques, qui reposent sur les lois mécaniques et acoustiques des systèmes (modélisation des causes), permettent généralement d'agir sur les paramètres de contrôle de la production sonore qui ont un sens physique. Il est alors possible de jouer par exemple sur la pression présente à l'embouchure d'un instrument à vent ou encore sur la force exercée par l'archet sur un instrument à cordes.

Les modèles de signaux, habituellement classés en modèles additifs (signal constitué par une somme de sinusoïdes [AS84, ?, MA89, DP91], avec une éventuelle adjonction de bruit [GL88, Ser89, SS90]), modèles soustractifs (signal "sculpté" à partir d'un bruit blanc ou d'un signal complexe), modèles source/filtre (signal constitué par une excitation filtrée par un résonateur, méthode LPC [MG76], "Linear Predictive Vocoder" [AH71]), et modèles non-linéaires (modulation de fréquence, "waveshaping" [Arf79, Yst89]), autorisent l'action directe

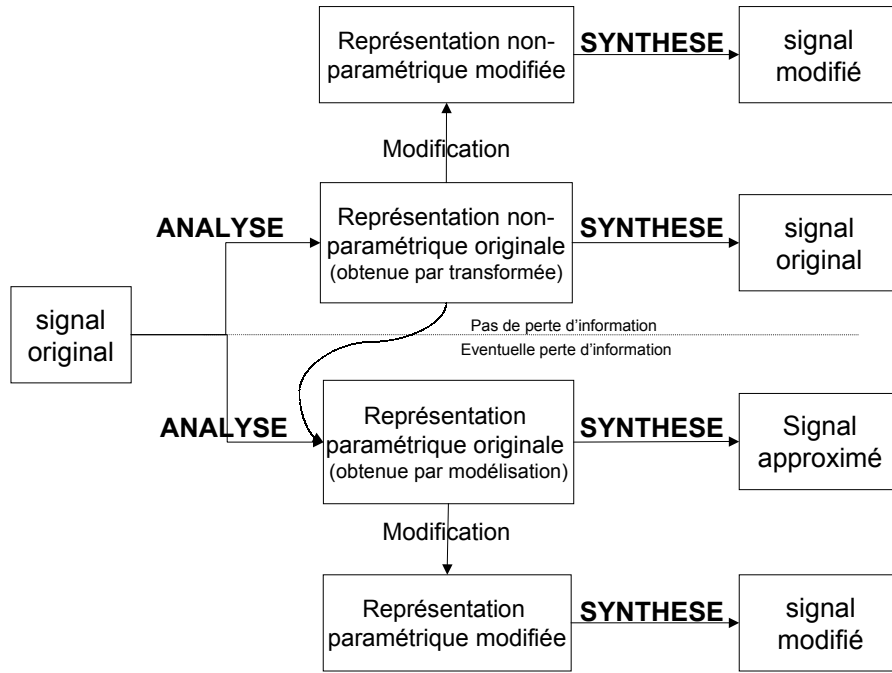


Figure 2.3 – Une vision du concept d’analyse/synthèse basé sur les différentes représentations

sur le caractère temporel ou fréquentiel du signal. Il est donc possible de jouer par exemple sur l’évolution temporelle de paramètres tels que l’amplitude et la fréquence instantanée des sinusoides, ou encore la période fondamentale d’un train d’impulsions et les résonances d’un filtre variant dans le temps.

Toute l’information d’un signal peut être conservée dans une représentation paramétrique si l’on adjoint au signal synthétisé grâce au modèle $s'(t)$, le signal ”résiduel” $s''(t)$, obtenu par différence entre le signal original $s(t)$ et le signal synthétisé. Ces deux signaux contiennent en effet toute l’information originale puisque $s(t) = s'(t) + s''(t)$. C’est une technique très employée pour effectuer des décompositions multiples du signal original tout en conservant la totalité de l’information. On peut en effet réaliser une autre modélisation sur le signal résiduel et obtenir de la sorte un nouveau signal résiduel.

Grâce à ces modèles, il est très simple d’effectuer une dilatation-p en interpolant les paramètres de contrôle (tels que la pression ou la force dans les modèles physiques) ou les paramètres d’évolution temporelle des signaux (tels que l’amplitude et la fréquence des sinusoides [DF98] ou encore la période fondamentale du train d’impulsions et les ”zéros” du filtre récursif), qui sont alors des fonctions explicites du temps, sans modifier les autres paramètres.

Soit D la fonction de dilatation, x et y les signaux originaux et modifiés, \vec{p}_x et \vec{p}_y les vecteurs des paramètres d’analyse et de synthèse, on effectue la synthèse du signal dilaté avec les paramètres suivants :

$$\vec{p}_y(t) = \vec{p}_x(D^{-1}(t))$$

La transposition-p est également très simple. Elle consiste à modifier les paramètres physiques correspondant aux notes des instruments (longueur de corde, de colonne d’air) ou les paramètres agissant sur les fréquences des sinusoides issus des modèles de signaux (fréquence

des partiels pour les modèles additifs [DF98], période fondamentale du train d'impulsions pour les méthodes LPC).

L'inconvénient majeur des représentations paramétriques réside dans la relation forte que doivent entretenir le signal et son modèle. Dès lors que le modèle n'est plus adapté au signal original, le signal synthétisé (sans modification des paramètres) peut être très différent de l'original si de l'information perceptivement audible n'est pas prise en compte dans le modèle. D'autre part un signal synthétisé à partir de paramètres modifiés ne reflète pas toujours la transformation désirée car ces paramètres ne correspondent pas à la signification que le modèle paraît leur donner. Par exemple, l'information d'un signal constitué principalement de transitoires est très mal représentée (et même souvent perdue) dans un modèle additif. Il en résulte un signal de synthèse d'autant moins convaincant que l'on modifie entre-temps les paramètres, devenus très discutables quant à leur signification.

En outre, les modèles sont parfois trop simplifiés pour s'adapter correctement à la complexité des signaux issus du monde réel : dans la méthode LPC, un train d'impulsions pour un son voisé et un bruit aléatoire pour un son non-voisé rendent la parole indistincte ou "floue" [MC90].

Les méthodes basées sur des représentations paramétriques sont donc par nature peu robustes pour des transformations-p, c'est-à-dire qu'elles se comportent de façon inattendue face à un signal arbitraire. Nous ne les développerons donc pas dans cette étude.

Représentations non-paramétriques

La représentation non-paramétrique ne repose pas sur l'utilisation explicite d'un modèle. Elle consiste en un ensemble de données, qui sont issues de la décomposition sur une famille de vecteurs bien choisis. La transformation mathématique associée porte généralement le nom de "transformée".

Cette transformée peut ne pas être inversible et ainsi aboutir à une perte d'information (spectrogramme, Wigner-Ville lissé [Fla98]), ce qui la rend inadaptée à la modification des signaux.

Dans le cas où elle est inversible, la représentation peut être critique, c'est-à-dire contenir exactement la même quantité de données que le signal original, ou bien être redondante, c'est-à-dire contenir plus de données que le signal original.

Bien qu'elles soient les plus employées, les représentations redondantes (Gabor, Fourier à Court Terme, Ondelettes) possèdent une caractéristique particulière due à leur redondance d'information qui doit rester cohérente : une version arbitrairement modifiée d'une représentation redondante ne correspond généralement pas à la transformée d'un signal réel. En d'autres termes, on ne retrouve pas la représentation modifiée par transformée inverse puis transformée directe. Cette propriété est due à l'existence d'un noyau reproduisant [KMMG87].

Les représentations non-paramétriques sont parfois à la base de représentations paramétriques, comme c'est le cas pour les modèles additifs : les paramètres sont extraits d'une représentation temps-fréquence.

Les représentations non-paramétriques sont parfois considérées comme faisant appel à un modèle implicite (par exemple, une somme de sinusoides de fréquence fixe variant lentement en amplitude et en phase au cours du temps pour la représentation de Fourier à court terme). Cette interprétation permet d'expliquer l'échec de la transformation lorsque le signal n'est pas adapté au modèle sous-jacent.

2.1.2 Propriétés des transformation-p

Propriétés de la dilatation-p

Les algorithmes de dilatation-p possèdent tous la même particularité : ils dilatent la durée d'un signal sans en modifier ses fréquences, c'est-à-dire que le signal original se trouve ralenti ou accéléré dans le cas d'un signal analogique, ou encore que le nombre d'échantillons en sortie est plus important ou plus faible que le nombre d'échantillons en entrée dans le cas d'un signal numérique.

Ces algorithmes sont parfois appelés algorithmes de "modification de l'échelle temporelle" ou TSM (de l'anglais "Time Scale Modification").

Propriétés de la transposition-p

Contrairement aux algorithmes de dilatation-p, les algorithmes de transposition-p fournissent le même nombre d'échantillons en sortie qu'en entrée. Seulement, chaque échantillon est modifié de manière à ce que toutes les fréquences soient transposées.

Cette transposition doit être réalisée avec précautions : pour un signal échantillonné, la fréquence la plus élevée pouvant être représentée correspond à la moitié de la fréquence d'échantillonnage F_e (voir [Cau41, Nyq28, Sha48]).

Si l'on transpose vers les aigus ($\alpha > 1$) un signal comportant des fréquences proches de $F_e/2$, les fréquences transposées peuvent subir un repliement spectral si la fréquence d'échantillonnage reste inchangée. Il est nécessaire d'appliquer un filtrage passe-bas afin d'éliminer les fréquences indésirables à l'origine d'artefacts de recouvrement. On perd évidemment dans ce cas l'information spectrale contenue entre $\frac{F_e/2}{\alpha}$ et $F_e/2$.

Si l'on transpose vers les graves ($\alpha < 1$), la bande de fréquences comprise entre $\alpha F_e/2$ et $F_e/2$ se retrouve vide d'information puisqu'il n'y en avait pas dans la bande originale correspondante $[F_e/2, \frac{F_e/2}{\alpha}]$. Des techniques de régénération de hautes fréquences existent pour le signal de parole (recouvrement spectral ou copie spectrale [ML95]) mais on peut s'interroger sur l'utilité et la validité de telles méthodes dans le cas de signaux musicaux.

Pour une transposition-p dans un but de dilatation-p (grâce au rééchantillonnage), il est important de conserver les fréquences supérieures à $F_e/2$ puisque cette information est utilisée au moment du rééchantillonnage afin de conserver le spectre utile du signal dilaté s'étendant jusqu'à $F_e/2$.

2.1.3 Discussion sur la modification des formants

Toutes les méthodes de transposition-p présentées ici dilatent la totalité du spectre, menant ainsi à une modification des formants dans le cas de la voix ou de certains instruments. Cette modification est voulue pour notre transposition-p, puisqu'après l'opération de rééchantillonnage, permettant d'obtenir la dilatation-p désirée, les formants retrouvent leurs positions initiales.

D'autre part, l'existence de formants repose sur une hypothèse faite sur le signal : le signal original doit être modélisé, grâce à un modèle source/filtre par exemple, afin d'extraire les paramètres représentant les formants. Il s'agit donc d'utiliser une représentation paramétrique, que nous ne développons pas ici, mais le lecteur intéressé pourra consulter [BJ95, Dud02].

La transposition-p sans modification de formants ne rentre donc pas dans le cadre de cette étude car d'une part, elle ne fournit pas la transformation requise et d'autre part, elle est basée sur une représentation paramétrique.

2.1.4 Bilan des représentations

Nous ne développons pas les méthodes de transformation-p basées sur des représentations paramétriques du signal, aussi nombreuses et variées qu'il existe de modèles différents. En effet, nous supposons qu'elles ne donnent pas de résultats assez satisfaisants sur la totalité des sons que l'on peut rencontrer sur une bande-son de film (à moins qu'il n'existe un jour un modèle valide pour tous les sons), car elles ne sont pas assez robustes.

Nous nous concentrons dans cet exposé uniquement sur les modifications des représentations non-paramétriques, les seules qui semblent être adaptées à la transformation-p d'un signal arbitraire. Ces méthodes ne font a priori aucune hypothèse sur le signal, elles sont donc toutes adaptées à la musique et a fortiori à la parole.

Cependant, nous étudions quelques méthodes utilisant des représentations paramétriques "résiduelles" en tant que décomposition du signal original (l'information incorrectement modélisée est portée par le résidu).

Pour effectuer cette classification, nous décidons de rester le plus proche possible des dénominations employées dans la littérature. Ainsi, nous distinguons les "méthodes temporelles" des "méthodes fréquentielles" selon le type de représentation sur lequel elles agissent¹. Nous introduisons également un troisième type de méthode, appelé "temps-fréquence", dont la représentation n'est ni purement temporelle, ni strictement fréquentielle (au sens communément admis, à savoir la Représentation de Fourier à Court Terme), mais une sorte de généralisation de cette dernière représentation.

Nous décidons de traiter des problèmes de dilatation-p et de transposition-p au sein de chacune de ces classes de méthodes pour plusieurs raisons : d'abord parce que les méthodes temporelles n'autorisent que la dilatation-p (la transposition-p y est impossible par une approche non-paramétrique), et ensuite parce que dilatation-p et transposition-p sont des approches théoriques équivalentes dans les méthodes fréquentielles et temps-fréquences.

1. Tout algorithme "temporel" possède cependant une interprétation "fréquentielle" et vice-versa du fait de la dualité temps/fréquence.

2.2 Méthodes temporelles

Nous conservons le terme "méthodes temporelles" généralement adopté dans la littérature pour caractériser les méthodes de transformation-p faisant appel **uniquement à la représentation temporelle** du signal. Il s'agit de segmenter le signal original en grains temporels que l'on réorganise ensuite différemment le long de l'axe temporel.

Elles possèdent la caractéristique qu'à part un changement de gain, **aucun des grains temporels n'est modifié**. Il en résulte que seule la dilatation-p est réalisable.

En effet, il est a priori impossible d'obtenir une transposition-p à partir de méthodes temporelles non-paramétriques² puisque la lecture de grains temporels originaux à la vitesse originale conserve les fréquences. Des grains conservés intacts ne peuvent donc pas donner lieu à un changement de fréquence.

Par contre, des grains temporels lus à des vitesses différentes de l'originale mènent bien à une transposition-p si ces grains ont été réagencés de manière à conserver au final la durée originale, mais il s'agit alors de méthodes temporelles de dilatation-p suivies (ou précédées) de rééchantillonnage.

Ces méthodes sont parfois appelées méthodes de "collage" ou de "raccord" (en anglais "splicing methods" [Mal79], "cut-and-splice methods" [MC90] ou encore "splice methods" [Lar95]), méthodes "d'édition automatique de la forme d'onde" ("automatic waveform editing" [SVW94]) ou encore "méthodes par granulation temporelle" [Roa96].

Une technique associée à ces méthodes peut être réalisée manuellement : par exemple, un raccourcissement d'un son sur support optique peut être réalisé en retirant des morceaux du film et en recollant les morceaux restant. Inversement, on peut allonger ce même son en dupliquant certains morceaux. Cette technique est également réalisable avec des supports magnétiques, ou encore sous forme numérique à l'aide de logiciels d'édition sonore.

Nous laissons bien évidemment de côté ces techniques manuelles, bien trop fastidieuses dans le cas d'un traitement d'une bande-son entière, pour nous concentrer uniquement sur les techniques automatiques.

2.2.1 Principe général des méthodes temporelles

Le principe commun à toutes ces méthodes, illustré en figure 2.4, consiste à décomposer le signal temporel original $s(t)$ en une série de grains temporels $g_i(t)$, centrés sur différentes marques de lecture L_i , par l'application de fenêtres temporelles $h_i(t)$ (voir équation 2.1), puis à recomposer le signal dilaté par sommation de ces mêmes grains temporels $g_i(t)$ centrés sur différentes marques d'écriture E_i , en prenant soin de normaliser le signal résultant (voir équation 2.2) pour éviter les éventuelles modulations d'amplitude.

2. La méthode PSOLA [MC90], souvent classée parmi les méthodes temporelles, n'est pas vraiment une méthode non-paramétrique. En effet, elle est basée sur l'hypothèse forte d'un signal monophonique de type source/filtre, dans lequel la composante excitatrice (impulsion glottal dans le cas de la voix) possède une énergie bien localisée dans le temps. De plus, des décisions du type signal voisé/non-voisé doivent être prises. Pour cette raison, nous ne l'exposerons pas directement dans cette classification, mais du fait de ses nombreuses similarités avec les méthodes temporelles exposées en section 2.2, et de son fonctionnement à l'aide de marques de correspondances similaires aux méthodes fréquentielles par interpolation des spectres de la section 2.3.2, nous en donnons une interprétation en annexe C.

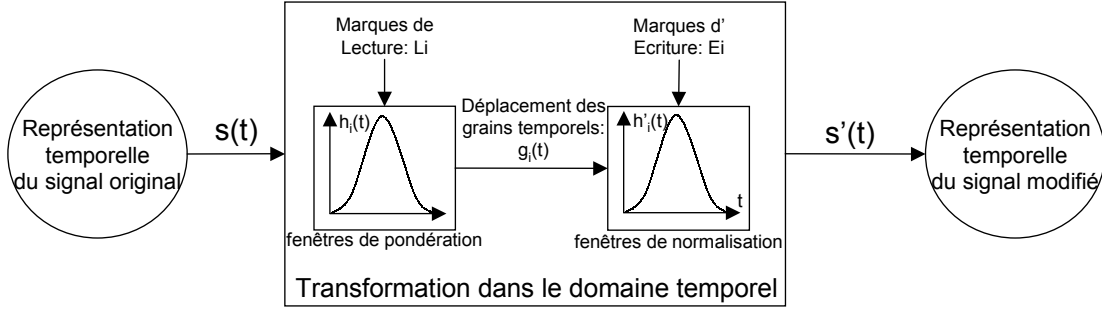


Figure 2.4 – Illustration de la transformation des méthodes temporelles

Les paramètres des méthodes temporelles généralement utilisés sont les suivants :

- Les marques de lecture L_i pour chaque itération, disposées de manière régulière ou non.
- Les marques d'écriture E_i pour chaque itération, disposées de manière régulière ou non.
- Les fenêtres temporelles $h_i(t)$ de support temporel H_i , qui peuvent être identiques ou non à chaque itération.

Les formules pour le principe général des méthodes temporelles sont les suivantes :

⇒ Formule de granulation temporelle :

$$g_i(t) = h_i(t)s(t + L_i) \quad (2.1)$$

Interprétation : application d'une fenêtre de pondération centrée sur la marque de lecture.

⇒ Formule de construction temporelle pour la dilatation-p :

$$Dp[s](t) = \frac{\sum_i g_i(t - E_i)}{\sum_i h_i(t - E_i)} \quad (2.2)$$

Interprétation : sommation des grains temporels centrés sur la marque d'écriture et normalisation pour éviter les modulations d'amplitude.

La formule précédente s'écrit donc, en prenant en compte l'équation 2.1:

$$Dp[s](t) = \frac{\sum_i h_i(t - E_i)s(t - E_i + L_i)}{\sum_i h_i(t - E_i)} \quad (2.3)$$

Dans ce type de méthode, il peut être contraignant de devoir normaliser lors de la construction temporelle³.

Il est donc plus simple de faire en sorte que la somme des fenêtres de granulation temporelle centrées sur les marques d'écriture soit égale à l'unité :

$$\sum_i h_i(t - E_i) = 1 \quad (2.4)$$

3. Dans les appareils mécaniques, cela implique d'agir sur le niveau en synchronisme avec les différentes vitesses de rotation, et dans leurs équivalents numériques, cela ajoute un surcroît de calculs toujours néfaste à l'efficacité. En pratique, seule la méthode SOLA (section 2.2.3) fait appel à cette normalisation.

De cette manière, la formule de construction temporelle pour la dilatation-p de l'équation 2.3 se simplifie de la manière suivante :

$$Dp[s](t) = \sum_i h_i(t - E_i)s(t - E_i + L_i) \quad (2.5)$$

L'idée de base de ce type de méthode consiste à **dilater localement** le signal temporel par déplacement des grains temporels, et à **réitérer cette opération** aussi souvent que nécessaire (périodiquement ou non selon les méthodes).

Pour chaque itération (c'est-à-dire pour chaque valeur de i), une dilatation-p locale est réalisée grâce à la modification de l'espacement entre marques de lecture et d'écriture, entraînant la duplication du grain temporel $g_i(t)$ lorsque $L_{i+1} = L_i$. La dilatation-p globale se fait ainsi itération après itération.

Pour exposer clairement le principe de cette méthode, nous nous focalisons sur une seule de ces itérations consistant à dilater localement le signal. Nous remarquons que plusieurs interprétations peuvent être données. Elles sont représentées par la figure 2.5.

Représentation par fenêtrage

L'interprétation des méthodes temporelles exposée à travers les formules consiste à considérer la dilatation locale comme une granulation temporelle du signal original par application de fenêtres de pondération sur des marques de lecture L_i , et un collage (sommation) de ces grains temporels en des marques d'écriture E_i décalées par rapport aux marques de lecture. Ce décalage, réalisé sur une durée K , fait apparaître la duplication ou la suppression d'un segment de durée K .

Représentation par segmentation

Une autre interprétation possible consiste à découper le signal original en un certain nombre de segments, puis à dupliquer (pour une élongation, soit $\alpha > 1$) ou supprimer (pour une contraction, soit $\alpha < 1$) un de ces segments dont la durée est notée K .

La représentation associée, que nous appelons "représentation par segmentation", symbolise cette segmentation, avant et après l'étape de duplication ou suppression. Elle possède l'avantage de faciliter la compréhension du principe de dilatation locale. Cette interprétation est celle retenue pour exposer la méthode HARMO dans la section 3.4.

Représentation par mixage

La "représentation par mixage" consiste à mettre en parallèle deux portions identiques du signal original, mais décalés d'une durée K . On indique sur chacune d'elles les valeurs de pondération, et l'on construit le signal résultant par sommation. Cette représentation est très intuitive car elle correspond à une opération qui peut être réalisée en pratique par mixage de bandes magnétiques. De plus, elle généralise la représentation par "segmentation", qui ne montre qu'un seul des deux segments originaux sans indiquer les pondérations appliquées.

Diagramme d'entrée/sortie

La représentation dans un diagramme d'entrée/sortie, proposé par Laroche [Lar95] et illustrée en figure 2.6, permet d'observer le temps (ou le nombre d'échantillons) écoulé en sortie en fonction du temps (ou du nombre d'échantillons) écoulé en entrée. La droite de pente unitaire indique un résultat sans traitement ($\alpha = 1$) alors que la droite de pente α indique l'objectif idéal de traitement. Cette droite "idéale" est approximée par une succession de segments parallèles à la

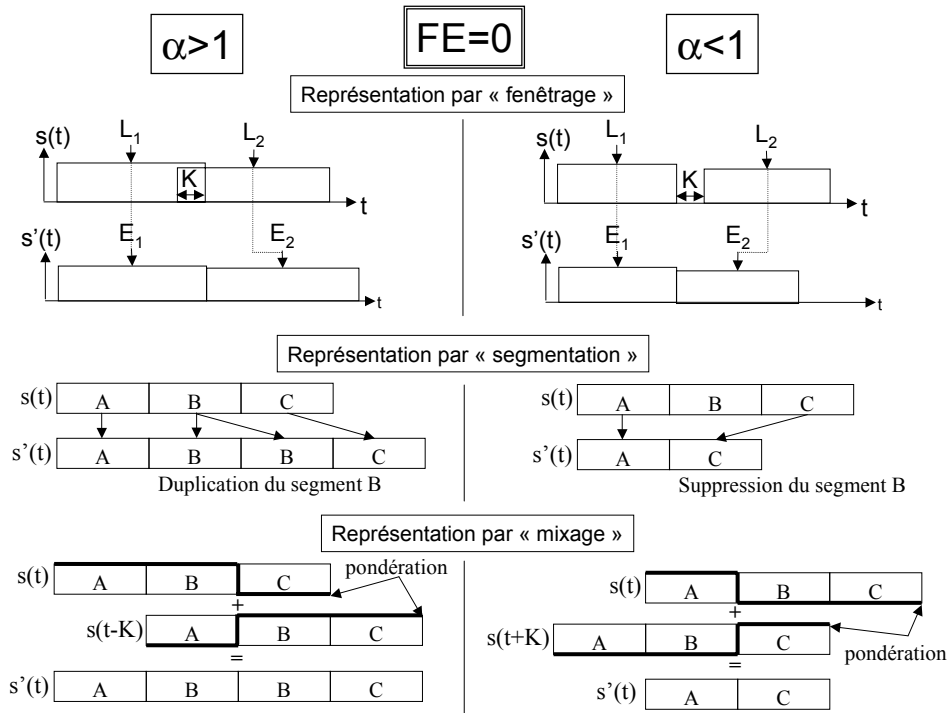


Figure 2.5 – Différentes représentations des méthodes temporelles

droite de pente 1 (correspondants aux segments recopiés tels quels) et de segments verticaux (correspondants aux segments dupliques) pour $\alpha > 1$ ou horizontaux (correspondants aux segments supprimés) pour $\alpha < 1$. Cette représentation permet de visualiser le phénomène d'anisochronie sur lequel nous reviendrons en section 3.1.

Dans cette représentation, alors qu'il n'y a pas d'ambiguïté pour définir quel segment a été supprimé, on remarque qu'il existe deux possibilités de définir le segment qui a été dupliqué. En effet, si la duplication est parfaite, on ne peut plus distinguer le segment original du segment dupliqué.

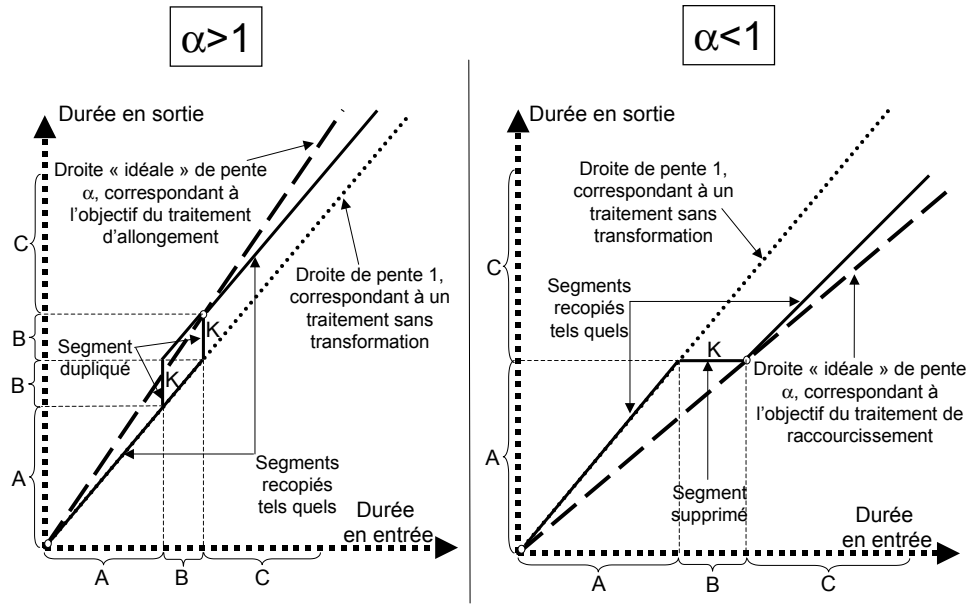


Figure 2.6 – Illustrations du diagramme d'entrée/sortie

Utilisation du fondu-enchaîné

Il se crée généralement, pour une élongation comme pour une contraction, une discontinuité de la forme d'onde au point de raccordement des 2 fenêtres, engendrant un "clac" (impulsion très brève). De manière à atténuer cet artefact, on utilise des fenêtres qui ne sont plus rectangulaires⁴, et qui donc proposent un fondu-enchaîné ("crossfade" en anglais) naturel. On évite de cette manière les discontinuités d'ordre 0 (discontinuité du signal) et on atténue les discontinuités d'ordres supérieurs (discontinuités des dérivées du signal).

On peut remarquer que dans le cas des sons réels, les discontinuités d'amplitude (variation de l'amplitude d'une composante sinusoïdale) sont généralement bien "gommées" (l'artefact est atténué mais on ne peut généralement pas l'éliminer totalement) par l'utilisation du fondu-enchaîné (voir figure 2.7), par contre les discontinuités de désynchronisation (décalage de la phase d'une composante sinusoïdale) provoquent quand même une déformation importante de la forme d'onde (voir figure 2.8).

L'utilisation d'un fondu-enchaîné de longueur inférieure à K modifie légèrement l'approche par "segmentation" : pour une élongation ($\alpha > 1$), on insert un segment, combinaison des deux segments adjacents entre lesquels il prend place. Pour une contraction ($\alpha < 1$), on substitue les deux segments adjacents par un seul, qui est une combinaison des deux précédents. L'utilisation d'un fondu-enchaîné de longueur supérieure à K ne permet plus de raisonner facilement selon la représentation par "segmentation".

La figure 2.9 montre l'utilisation d'un fondu-enchaîné linéaire dont la longueur correspond à K , et la figure 2.10 les diagrammes d'entrée/sortie correspondants.

Pour faciliter la compréhension du fondu-enchaîné, on introduit la notion de "fenêtre équivalente". Il s'agit de la fenêtre rectangulaire correspondante à un fondu-enchaîné de durée nulle. La limite de ces fenêtres est située au point d'intersection des fenêtres de granulation

4. Une discussion sur les caractéristiques de ces fenêtres sera faite en section 3.4.5.

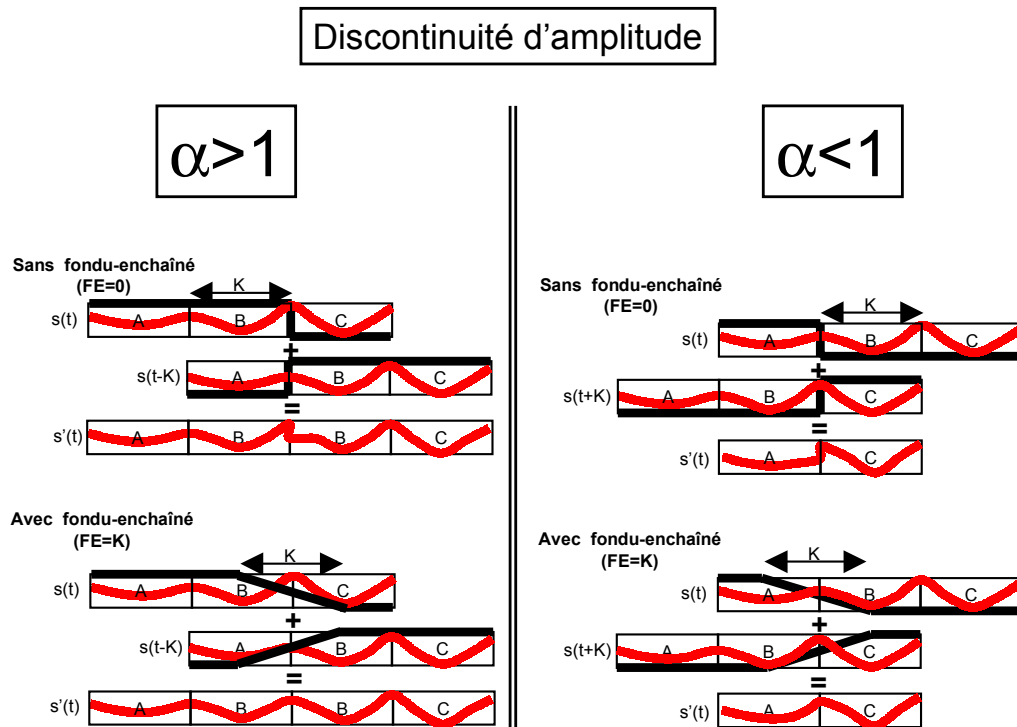


Figure 2.7 – Discontinuités d'amplitude de la forme d'onde

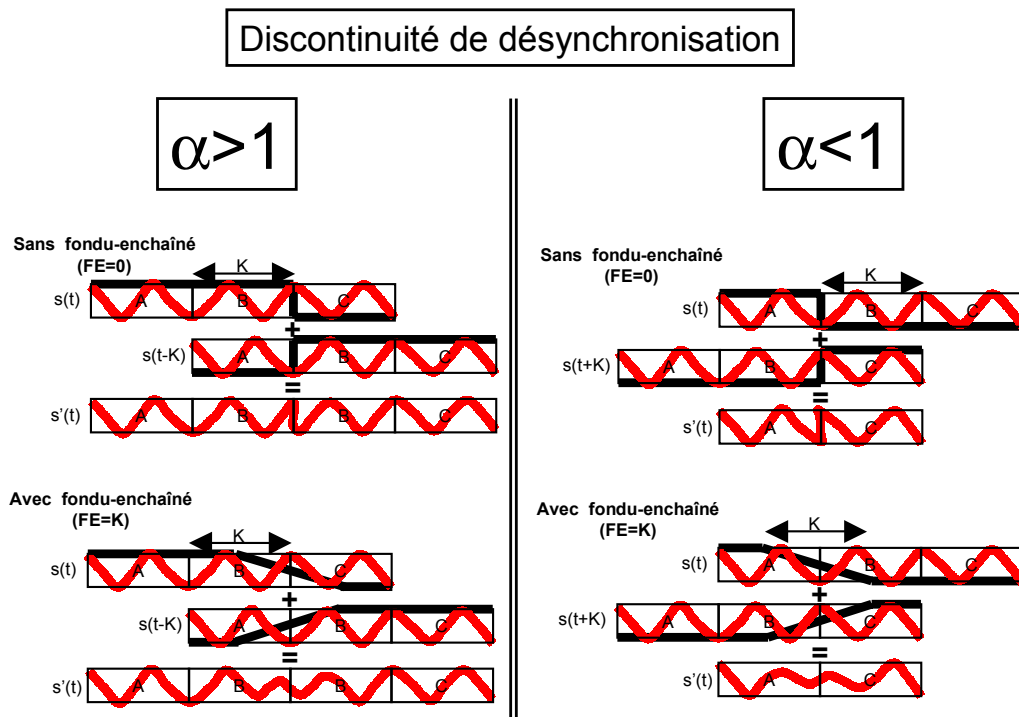


Figure 2.8 – Discontinuités de désynchronisation de la forme d'onde

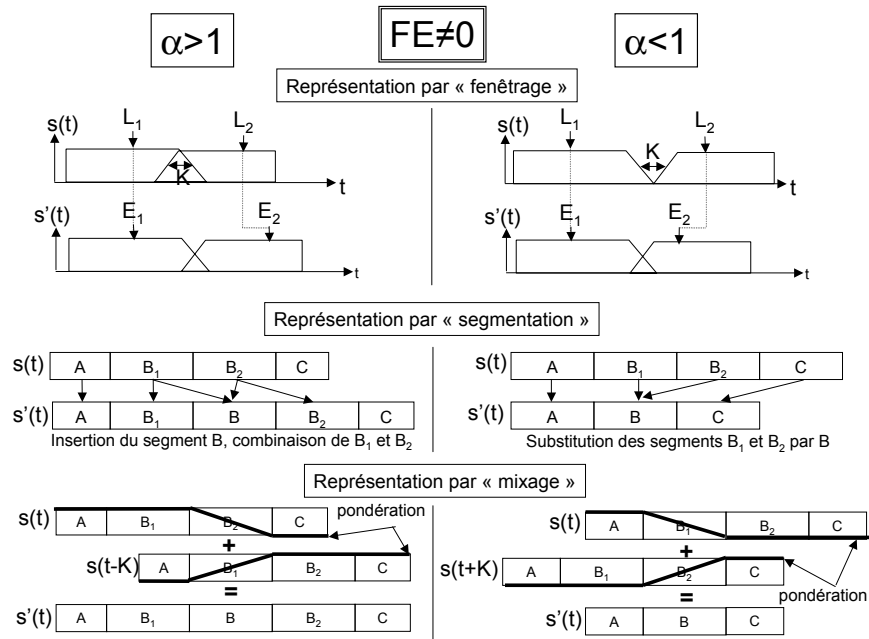


Figure 2.9 – Différentes représentations des méthodes temporelles avec fondu-enchaîné linéaire

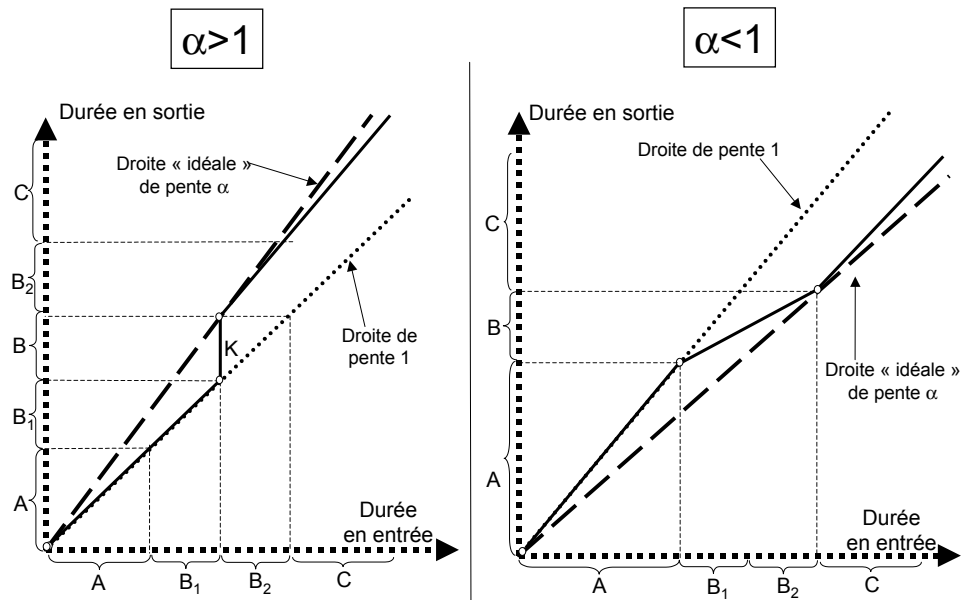


Figure 2.10 – Illustrations du diagramme d'entrée/sortie avec fondu-enchaîné linéaire

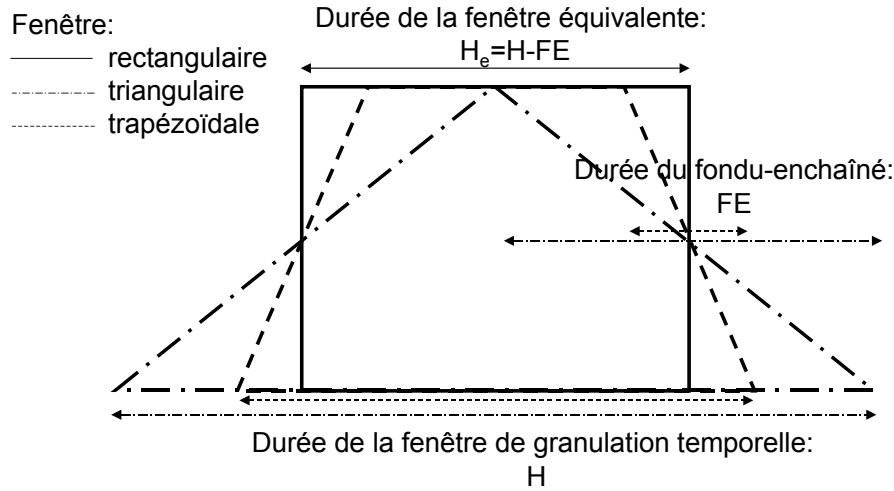


Figure 2.11 – Schéma d'équivalence entre les fenêtres de pondération

temporelle lors de l'écriture, qui se trouve être le point à mi-hauteur de la fenêtre si celle-ci est symétrique. La figure 2.11 illustre les relations entre les fenêtres.

Soit H la durée de la fenêtre de granulation temporelle associée à un fondu-enchaîné (fenêtre trapézoïdale par exemple). On définit la durée H_e de la fenêtre équivalente par l'équation suivante :

$$H_e = H - FE$$

Dans la suite de ce travail, nous appelons "segment inséré" le segment B en référence à la figure 2.9, qui correspond au segment inséré entre les segments B_1 et B_2 (pour $\alpha > 1$) à partir desquels il est créé, ou à la place de ceux-ci (pour $\alpha < 1$). Sa durée vaut K .

Discussion

Dans ce type de méthode, il n'y a "a priori" aucune hypothèse de faite sur le signal, en ce sens que le signal d'origine n'a pas à être modélisé. Ces méthodes n'utilisent donc aucun paramètre de synthèse, puisque la seule information nécessaire pour construire le signal dilaté est une combinaison de signaux originaux pondérés par des fenêtres temporelles. Cependant, nous verrons que ces méthodes peuvent donner de très bons résultats auditifs lorsque le signal est quasi-harmonique et varie lentement dans le temps. Elles donnent encore de bons résultats lorsque le signal est légèrement inharmonique, polyphonique, ou lorsque des transitoires sont présents, mais au prix d'une analyse préalable.

2.2.2 Méthodes "aveugles"

Historique

Le principe que nous allons exposer trouve sa source dans le brevet de French et Zinn datant de 1928 [FZ28]. Sa mise en pratique revêt la forme soit d'un appareil acoustique, soit d'un appareil électromagnétique. Le brevet de Freund [Fre35], reposant sur le même principe, est décliné quant à lui sous la forme d'un appareil mécanique utilisant un support optique (l'article de Gabor [Gab46] décrit précisément un appareil de ce type), et ceux de Schüller [Sch44] et de Fairbanks *et al.* [FEJ59] sous la forme d'un appareil mécanique utilisant une bande magnétique (une version automatique de ce que réalisait Garvey manuellement pour son expérience d'intelligibilité [Gar53]).

C'est sous cette dernière forme, à savoir un magnétophone à têtes tournantes, que de nombreuses machines de dilatation-p et de transposition-p sont apparues : Pierre Schaeffer créa le phonogène universel (conçu et réalisé par Jacques Poullin) en 1951 [Pal93], la compagnie allemande AEG (producteur du premier magnétophone moderne) commercialisa le "Magnetophon-Special" [Sch54], Springer le Tempophon [Spr55], Telefonbau und Normalzeit le "Tempo-Regulator" [Sco67], Lexicon le "Varispeech" [Lee72]. Fairbanks *et al.* écrivirent un article scientifique décrivant précisément ce type d'appareil et la méthode employée [FEJ54].

Des machines équivalentes aux appareils analogiques ont ensuite été réalisés de manière numérique dans le contexte de la parole [Lee72] aussi bien que de la musique [BOGC68] [Tru94]. Celles-ci prennent le nom de "système de registres à décalages" [Lee72], "système RAM ("Random Access Memory")" [Lee72], "granulation temporelle" [Roa96], ou encore "méthode de la mémoire tampon circulaire" [Lar95, Ben88].

Principe et interprétations des méthodes "aveugles"

On qualifie ces méthodes d'"aveugles" car elles agissent indépendamment du signal qu'elles modifient.

Les caractéristiques des variables pour les méthodes "aveugles" sont les suivantes :

→ Les marques de lecture sont régulièrement espacées :

$$L_i = iL$$

→ Les marques d'écriture sont aussi régulièrement espacées :

$$E_i = iE$$

→ Les fenêtres temporelles sont toutes identiques :

$$h_i(t) = h(t)$$

Les marques de lecture et d'écriture sont reliées par la fonction de dilatation :

$$E_i = D(L_i) = \alpha L_i$$

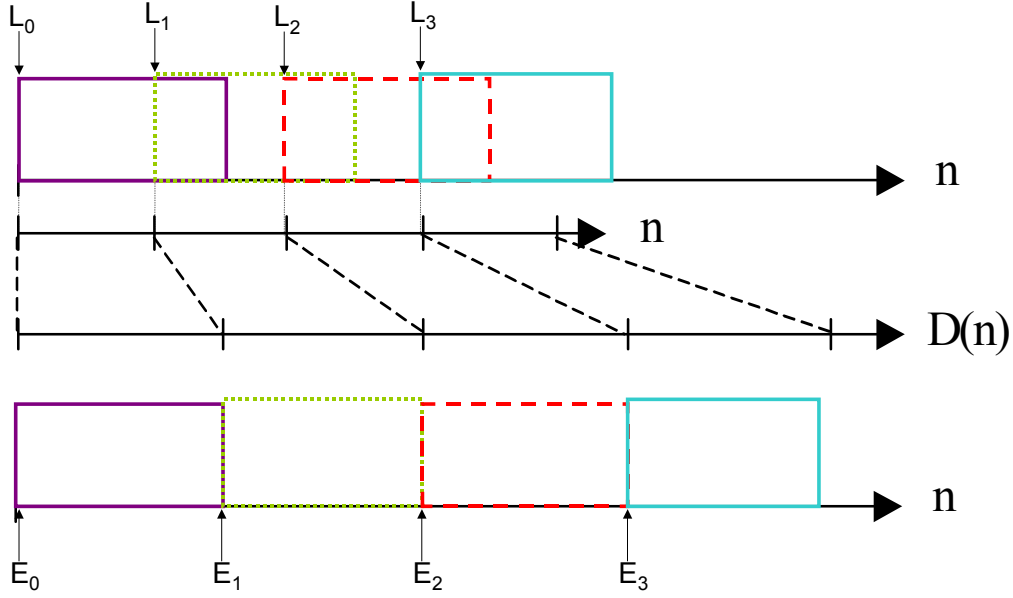


Figure 2.12 – Schéma de positionnement des fenêtres pour une méthode aveugle à fenêtres rectangulaires juxtaposées ($\alpha = 1,5$)

d'où

$$E = \alpha L \quad (2.6)$$

Les formules pour les méthodes "aveugles" sont les suivantes :

⇒ Formule de granulation temporelle pour les méthodes "aveugles" :

$$g_i(t) = h(t)s(t + iL) \quad (2.7)$$

⇒ Formule de construction temporelle pour la dilatation-p des méthodes "aveugles" :

$$Dp[s](t) = \frac{\sum_i g_i(t - iE)}{\sum_i h(t - iE)} \quad (2.8)$$

La formule précédente s'écrit donc, en prenant en compte l'équation 2.7:

$$Dp[s](t) = \frac{\sum_i h(t - iE)s(t - iE + iL)}{\sum_i h(t - iE)} \quad (2.9)$$

La figure 2.12 illustre le cas d'une dilatation-p pour $\alpha = 1,5$, en utilisant des fenêtres de granulation temporelle rectangulaires.

Cet exemple nous montre clairement qu'un segment de durée constante K est inséré à chaque itération, avec :

$$K = \Delta E_i - \Delta L_i = (E_i - E_{i-1}) - (L_i - L_{i-1}) = E - L = L(\alpha - 1) \quad (2.10)$$

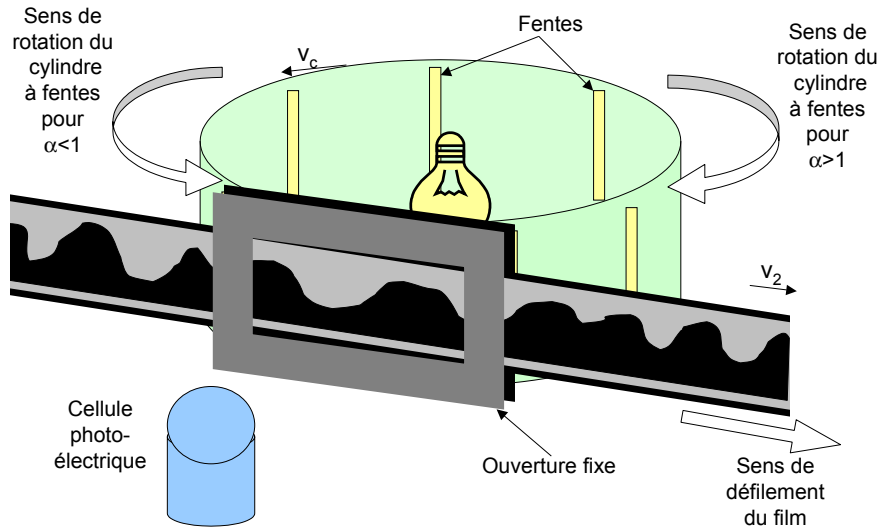


Figure 2.13 – Schéma de l'appareil optique

Par convention, $K > 0$ indique la durée d'un segment dupliqué ($\alpha > 1$, on est dans le cas d'une élongation temporelle), alors que $K < 0$ indique que le segment de durée $|K|$ est supprimé ($\alpha < 1$, on est dans le cas d'une contraction temporelle).

Appareils mécaniques

Les appareils mécaniques peuvent être déclinés sous plusieurs formes : acoustique [FZ28], optique [Fre35] et magnétique [Sch44]. Les principes de toutes ces méthodes sont similaires, seule la représentation physique de l'information change.

Nous donnons un exemple concret basé sur un appareillage optique, issu de [Gab46] et schématisé en figure 2.13, pour illustrer la technique de base des méthodes "aveugles". L'appareillage magnétique est quant à lui étudié en détail dans l'annexe D.

Prenons un signal sinusoïdal de fréquence f_0 enregistré sous forme optique sur une bande-film à la vitesse v_1 .

Supposons que la bande défile ensuite à la vitesse $v_2 = v_1$. La lecture de cette bande à travers une fente fixe nous donne la fréquence originale f_0 . Cependant, si la fente se déplace à la vitesse v_f , la vitesse relative bande-fente, appelée vitesse locale, devient $v_{locale} = v_2 - v_f$. Nous sommes confrontés au phénomène de lecture à vitesse variable, ainsi la fréquence obtenue devient $f_1 = \frac{v_{locale}}{v_1} f_0 = \frac{v_2 - v_f}{v_1} f_0 = \frac{v_1 - v_f}{v_1} f_0$.

Pour conserver la fréquence initiale f_0 , il est nécessaire de modifier la vitesse de lecture v_2 de sorte que la vitesse de lecture locale corresponde à la vitesse de l'enregistrement, $v_{locale} = v_1$, soit $v_2 = v_1 + v_f$. Par cette modification de vitesse de lecture, on compense la modification de fréquence induite par le déplacement de la fente.

Jusqu'ici, nous n'avons rien gagné puisque le signal obtenu sera certes entendu à la fréquence initiale désirée, mais il durera aussi le même temps car la fente se déplace le long de la bande.

Supposons maintenant que le film se déplace devant une ouverture fixe, telle que la fente ne soit effective (i.e. reçoive la lumière issue de la bande) que durant son passage devant cette

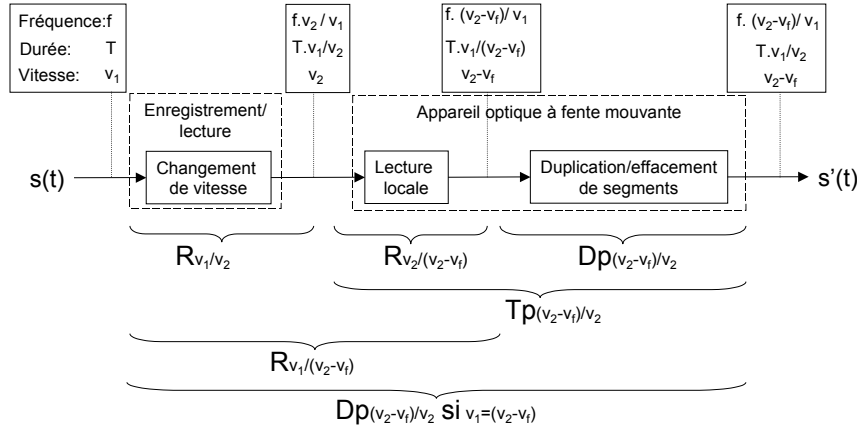


Figure 2.14 – Synoptique de fonctionnement de l'appareil optique de dilatation-p

ouverture. Pour obtenir un enregistrement continu, laissons apparaître une nouvelle fente au moment où la première fente quitte l'ouverture, puis une troisième et ainsi de suite. Nous disposons ainsi des fentes tout autour d'un cylindre rotatif. La vitesse de la fente v_f correspond donc à la vitesse linéaire du cylindre v_c :

$$v_c = v_f$$

Si $v_c < 0$ (la fente se déplace dans le sens opposé à celui du défilement de la bande), la nouvelle fente lit une seconde fois une partie du signal que la première fente avait déjà lu. Un segment est donc dupliqué, entraînant un allongement du signal. De cette manière, on réalise une élongation puisque le signal de sortie est plus long que l'original (la bande défile à la vitesse $v_2 = v_1 + v_c$, inférieure à la vitesse originale v_1 , donc dure plus longtemps) alors que la fréquence de la sinusoïde reste la même (la vitesse locale $v_{locale} = v_2 - v_c$ est égale à la vitesse initiale v_1).

Si $v_c > 0$ (la fente se déplace dans le sens de défilement de la bande), la partie de bande située entre les 2 fentes au moment du relai n'est pas lue. Un segment du signal est donc supprimé, entraînant ainsi un raccourcissement du signal. De cette manière, on réalise une contraction puisque le signal de sortie est plus court que l'original (la bande défile à la vitesse $v_2 = v_1 + v_c$, supérieure à la vitesse originale v_1 , donc dure moins longtemps) alors que la fréquence de la sinusoïde reste la même (la vitesse locale $v_{locale} = v_2 - v_c$ est égale à la vitesse initiale v_1).

La figure 2.14 représente le synoptique de fonctionnement d'un tel appareil optique, dans lequel sont représentées les fréquences et durées en chaque point de la chaîne. On y indique également les opérations de dilatation-p, transposition-p et rééchantillonnage qui sont effectuées.

Il est possible d'éviter les discontinuités d'amplitude, résultant de l'apparition d'une nouvelle fente, en réalisant un fondu-enchaîné. Cela est effectué en donnant à l'ouverture fixe une forme plus étendue, simple (triangulaire, trapézoïdale [Gab46]) ou complexe (exponentielle [Fre35]), permettant de prendre en compte la lumière émise par plusieurs fentes au même instant. Cette forme d'ouverture doit être adaptée à la distance entre les fentes, de manière à ce qu'à chaque instant la composante continue du signal reste constante (cet aspect est discuté en section 3.4.5).

La figure 2.15 schématise la forme et la position des fenêtres de lecture et d'écriture dans le cas d'une ouverture fixe de forme triangulaire.

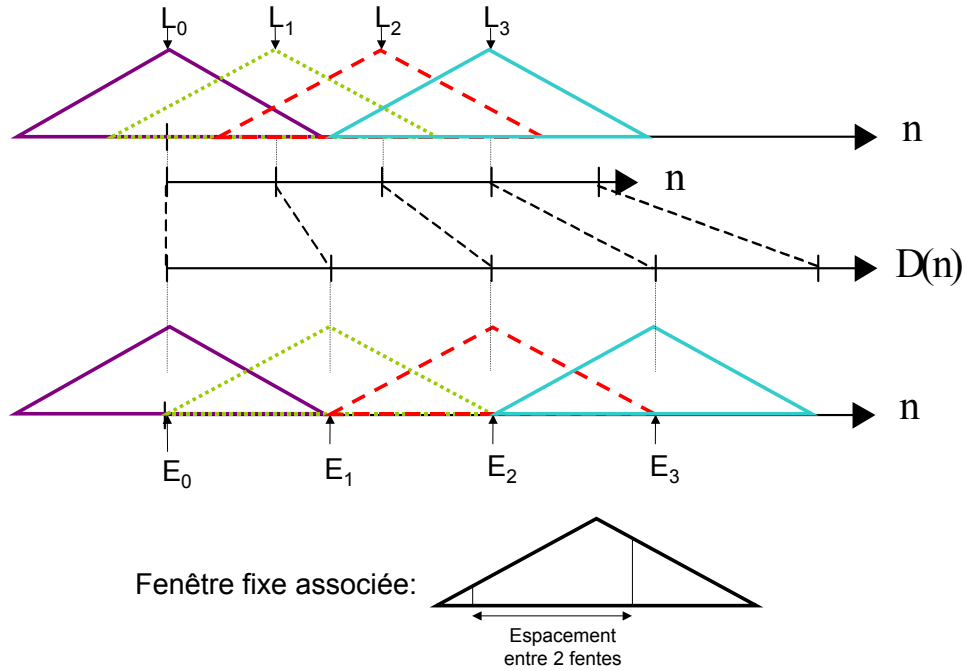


Figure 2.15 – Schéma de positionnement de fenêtres triangulaires pour un appareil à 2 fentes simultanées ($\alpha = 1,5$)

La forme des fenêtres de granulation correspond à la forme de l'ouverture fixe, et la durée H des fenêtres correspond au rapport de la longueur de l'ouverture L par la vitesse de rotation du cylindre contenant les fentes v_c :

$$H = \frac{L}{v_c}$$

On remarque dans ce schéma qu'à chaque instant, le signal de sortie est construit à partir de 2 grains temporels, correspondant aux 2 fentes visibles depuis la fenêtre fixe, mais que chaque échantillon du signal d'entrée participe à trois grains temporels différents.

La figure 2.16 schématise la forme et la position des fenêtres de granulation dans le cas d'une fenêtre fixe de forme trapézoïdale.

On peut ici faire les mêmes remarques que pour la fenêtre triangulaire. De plus, les marques de lecture ont été consciemment placées au passage à mi-hauteur de la fenêtre pour pouvoir faire le parallèle avec la figure 2.12 : on peut donc interpréter les fenêtres triangulaires ou trapézoïdales comme des fenêtres rectangulaires dont les extrémités ont subi une modification (rotation autour de leur point à mi-hauteur) afin de réaliser un fondu-enchaîné.

Inconvénients des appareils mécaniques

De nombreux inconvénients sont liés à ces techniques :

- Difficultés de mise en œuvre mécanique.
- Technologie analogique obsolète.
- Difficulté de modification du fondu-enchaîné.
- Impossibilité de modifier la durée du segment dupliqué/effacé en cours de traitement.
- Défauts audibles et rédhibitoires de discontinuités sur les sons purs.

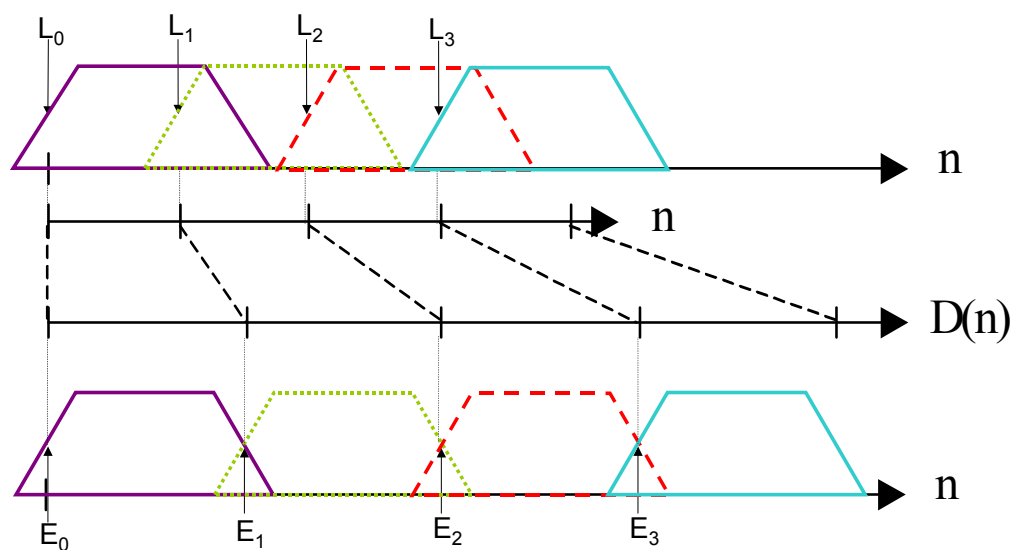


Figure 2.16 – Schéma de positionnement de fenêtres trapézoïdales pour un appareil à 2 fentes simultanées ($\alpha = 1,5$)

- Redoublement des consonnes plosives [Fre35] et présence d'échos lorsque les segments dupliqués sont trop longs [FEJ54] [FZ28] pour $\alpha > 1$.

Équivalents numériques

Des équivalents aux appareils mécaniques sont réalisés sous forme numérique. Une mémoire tampon (ou "buffer") circulaire remplace la bande, des pointeurs informatiques font office de têtes d'enregistrement et de lecture, les fréquences d'échantillonnage correspondent aux vitesses de lecture, et le rééchantillonnage numérique se substitue à la lecture à vitesse variable. Une simple addition pondérée des pointeurs de lecture génère le fondu-enchaîné.

Cette méthode étant réalisée sans analyse du signal, c'est-à-dire de manière "aveugle", les dilatations locales (duplication ou suppression d'un segment) sont répétées périodiquement. Il en résulte un diagramme d'entrée/sortie totalement prévisible, où, pour un taux de dilatation donné, les longueurs des segments dupliqués/supprimés sont constants (voir figure 2.17).

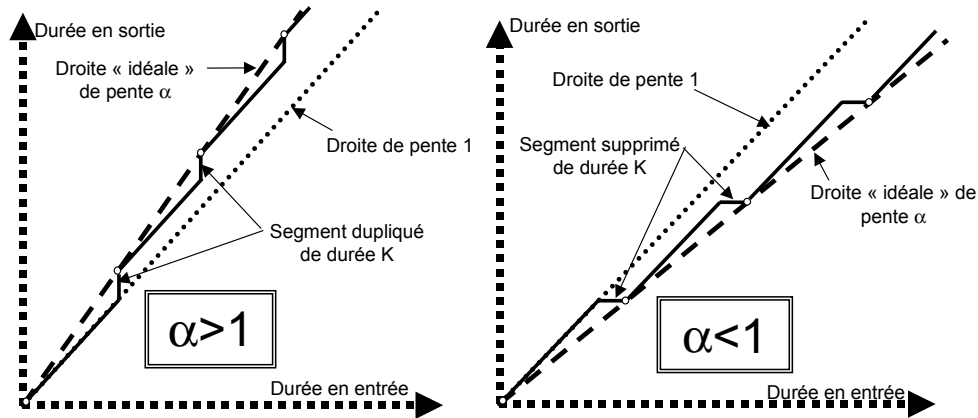


Figure 2.17 – Diagramme d'entrée/sortie standard des méthodes "aveugles" (appareils mécaniques et équivalents numériques)

Ainsi, toutes les descriptions de dilatation- p qui font appel à la granulation sont des dérivés numériques des appareils mécaniques, certes plus souples que ces derniers, mais ne fournissant pas de résultats auditifs supérieurs.

Paramètres modifiables par l'utilisateur

Les paramètres sur lesquels l'utilisateur peut agir sont :

- L'intervalle de temps E entre les marques d'écriture. L'intervalle de temps L entre les marques de lecture est déduit de E par l'équation 2.6 :

$$L = \frac{E}{\alpha}$$

- La durée K du segment dupliqué/supprimé (dans le cas où $FE = 0$), qui est lié à E par l'équation suivante :

$$K = E - L = E - \frac{E}{\alpha} = E(1 - \frac{1}{\alpha}) \quad (2.11)$$

- La durée H et la forme des fenêtres de granulation temporelle. Cette durée est égale à E dans le cas où les fenêtres de granulation temporelle équivalentes se juxtaposent lors de l'écriture.

Variation des paramètres en fonction de α

Selon les implantations utilisées, une variation de α entraîne la variation d'un paramètre ou d'un autre.

Par exemple, dans l'implantation par "shift register" de Lee [Lee72], la durée E séparant les marques d'écriture est équivalente à la durée H_e pour une fenêtre de granulation temporelle rectangulaire. Cette durée, correspondant à ce que Lee appelle les "intervalles conservés", est constante. Il s'ensuit que la durée K , correspondant à ce qu'il appelle les "intervalles supprimés", varie avec α selon l'équation suivante :

$$K = E(1 - \frac{1}{\alpha}) \quad (2.12)$$

D'autre part, dans l'implantation "RAM" de Lee [Lee72], la durée K est constante. C'est également le cas dans toutes les machines mécaniques (optiques et magnétiques) où elle est déterminée par construction de l'appareil.

Il s'ensuit que les marques de lecture E_i sont séparées d'une durée E donnée par l'équation suivante, tirée de l'équation 2.11 :

$$E = \frac{K}{1 - \frac{1}{\alpha}}$$

Avantages et inconvénients

L'avantage majeur de ce type de méthode est la très faible puissance de calcul nécessaire au traitement, utilisée presque exclusivement pour réaliser le fondu-enchaîné qui atténue les discontinuités.

Cependant, les méthodes de collage "aveugles" analogiques comme numériques possèdent les mêmes défauts auditifs :

- Discontinuités d'amplitude sur des sons purs modulés (sons [24, 25]), généralement gommés par l'utilisation d'un fondu-enchaîné (son [26]). La figure 2.18 représente ces sons et montre la discontinuité d'amplitude.
- Discontinuités de désynchronisation sur des sons purs (sons [27, 28]), légèrement gommés par l'utilisation d'un fondu-enchaîné (son [29]) mais qui demeurent des défauts rédhibitoires. La figure 2.19 représente ces sons et montre la discontinuité de désynchronisation.
- Redoublement des consonnes plosives [Fre35] ou présence de pré-échos [FEJ54, FZ28] lorsque les segments dupliqués sont trop longs pour $\alpha > 1$ (sons [30, 31] et figure 2.20). Plus généralement, il y a redoublement (resp. suppression) de transitoire lorsque celui-ci se situe dans la zone dupliquée (resp. supprimée) lorsque $\alpha > 1$ (resp. $\alpha < 1$).

On peut trouver dans [Ita98, KIS⁺99, IK99, Ita00] une discussion de ce type de méthode dans le cadre d'une application à des fonctions de vitesse variable dans un système de reproduction vidéo.

En conclusion, les artefacts liés à la discontinuité de désynchronisation sont des défauts rédhibitoires qu'il est indispensable d'atténuer, notamment grâce à l'utilisation de méthodes adaptatives.

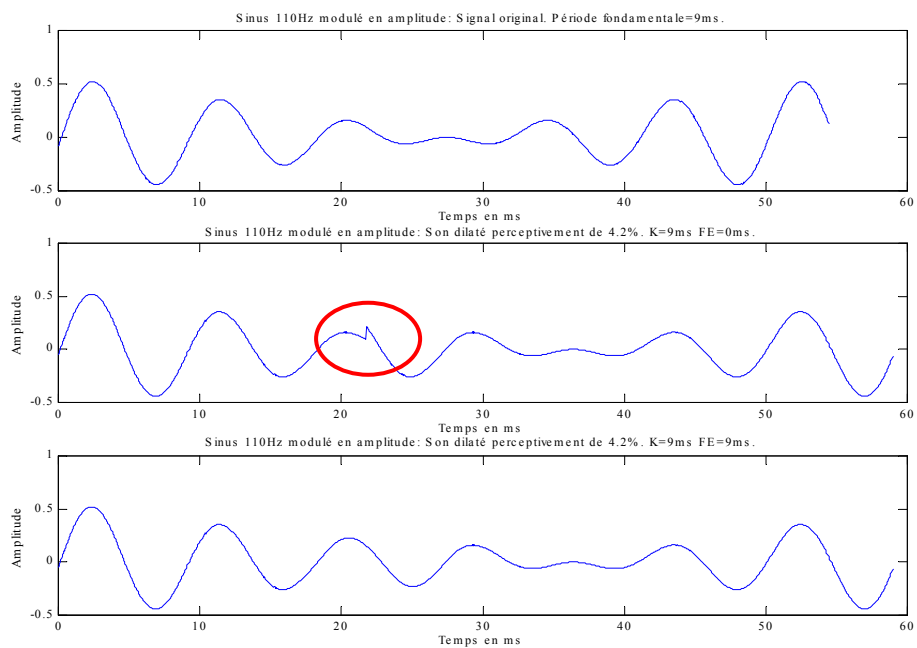


Figure 2.18 – Mise en évidence de la discontinuité d'amplitude

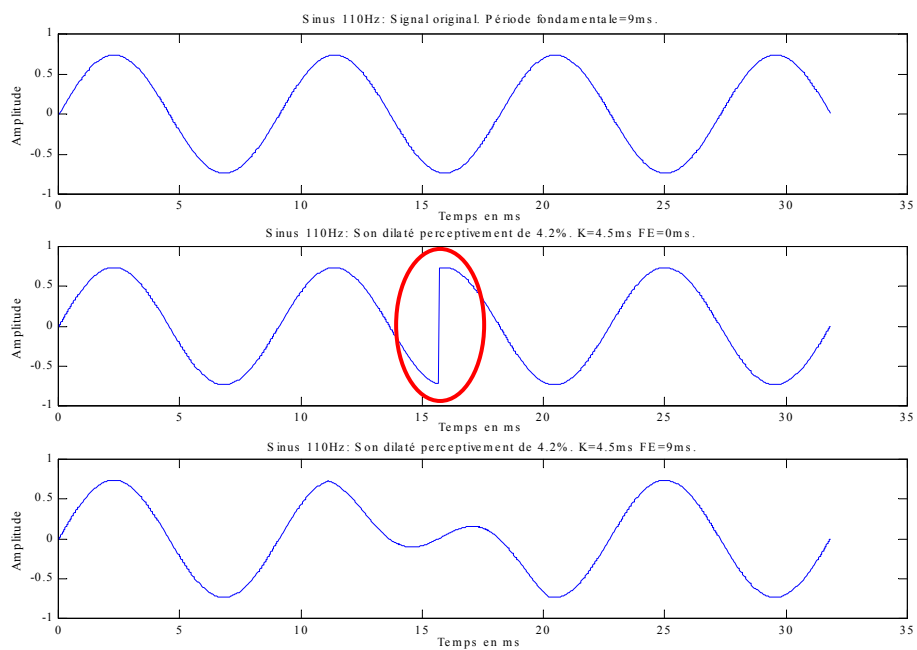


Figure 2.19 – Mise en évidence de la discontinuité de désynchronisation

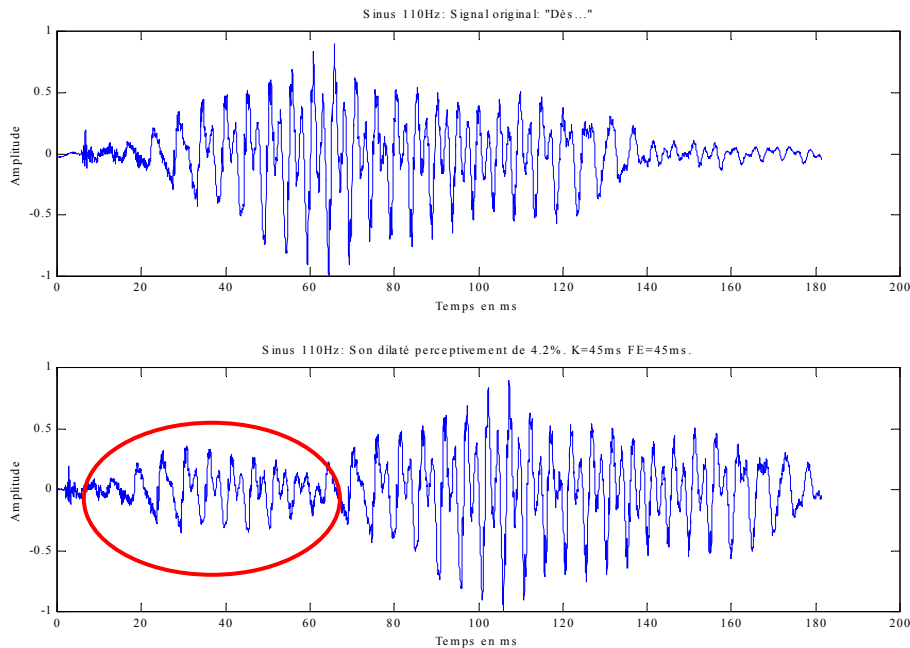


Figure 2.20 – Mise en évidence du redoublement ou pré-écho

2.2.3 Méthodes adaptatives

Les méthodes décrites précédemment ne donnent pas satisfaction sur la plupart des sons. Certains paramètres doivent en effet être adaptés au type de signal à traiter. C’est pourquoi de nouvelles méthodes ont vu le jour, permettant de faire varier des paramètres qui étaient arbitrairement fixés, tels que la durée K des segments insérés ou encore le positionnement de ces segments.

Historique

Scott [Sco67] remarque les discontinuités d’amplitude et de désynchronisation résultant de l’utilisation d’une méthode de collage ”aveugle” et propose de les minimiser en déplaçant manuellement les formes d’ondes pour assurer le synchronisme, cela revenant à faire varier la durée du segment inséré.

Neuburg [Neu78] propose d’adapter la durée du segment inséré à la période fondamentale du signal. Cette technique ne nécessite pas de repérer la fermeture glottale comme c’est le cas dans [Sco67], mais nécessite cependant l’utilisation d’un détecteur de période fondamentale. Ce dernier est réalisé par autocorrélation. Cette détection de période fondamentale peut être réalisée grâce à d’autres techniques [Gol62, RCRM76, Ter02].

Principe des méthodes adaptatives

L’aspect adaptatif de toutes ces méthodes provient de l’estimation, explicite (TDHS) ou implicite (SOLA, WSOLA, EDSOLA), de la période fondamentale P_i au voisinage des marques de lecture L_i . Ce paramètre est alors utilisé pour ajuster les marques de lecture, d’écriture, et fixer la durée des fenêtres de granulation temporelle.

Méthode TDHS ("Time Domain Harmonic Scaling")

En 1979, Malah [Mal79] introduit la méthode TDHS (dilatation d'harmonique dans le domaine temporel). Cette technique consiste à ajuster la fenêtre de granulation temporelle ainsi que les marques de lecture et d'écriture, aux fréquences du signal.

La taille des fenêtres de granulation temporelle est donc variable, proportionnelle à la période fondamentale locale du signal P_i . Ce facteur de proportionnalité dépend uniquement de α .

Les caractéristiques des variables pour les méthodes TDHS sont les suivantes :

→ Les marques de lecture sont placées aux instants suivants :

$$L_i = L_{i-1} + \frac{P_i}{\alpha - 1}$$

→ Les marques d'écriture sont déduites des marques de lecture par la fonction de dilatation :

$$E_i = D(L_i) = \alpha L_i = E_{i-1} + \frac{\alpha}{\alpha - 1} P_i$$

→ Les longueurs des fenêtres équivalentes de granulation temporelle sont données par :

$$H_e = \frac{\alpha}{\alpha - 1} P_i$$

Soit, pour des fenêtres triangulaires :

$$H = 2 \cdot \frac{\alpha}{\alpha - 1} P_i$$

→ La période fondamentale P_i est déterminée à l'aide d'un algorithme de détection de période fondamentale.

Les formules pour les méthodes TDHS sont les suivantes :

⇒ Formule de granulation temporelle :

$$g_i(t) = h_i(t)s(t + L_i) \quad (2.13)$$

⇒ Formule de construction temporelle pour la dilatation-p :

$$Dp[s](t) = \sum_i g_i(t - E_i) \quad (2.14)$$

La formule précédente s'écrit donc, en prenant en compte l'équation 2.13:

$$Dp[s](t) = \sum_i h_i(t - E_i)s(t - E_i + L_i) \quad (2.15)$$

La durée K_i du segment inséré équivalent (correspondant au segment inséré si la fenêtre de granulation temporelle est rectangulaire) est donnée par :

$$K_i = |\Delta E_i - \Delta L_i| = \left| \frac{\alpha}{\alpha - 1} P_i - \frac{1}{\alpha - 1} P_i \right| = |(E_i - E_{i-1}) - (L_i - L_{i-1})| = P_i$$

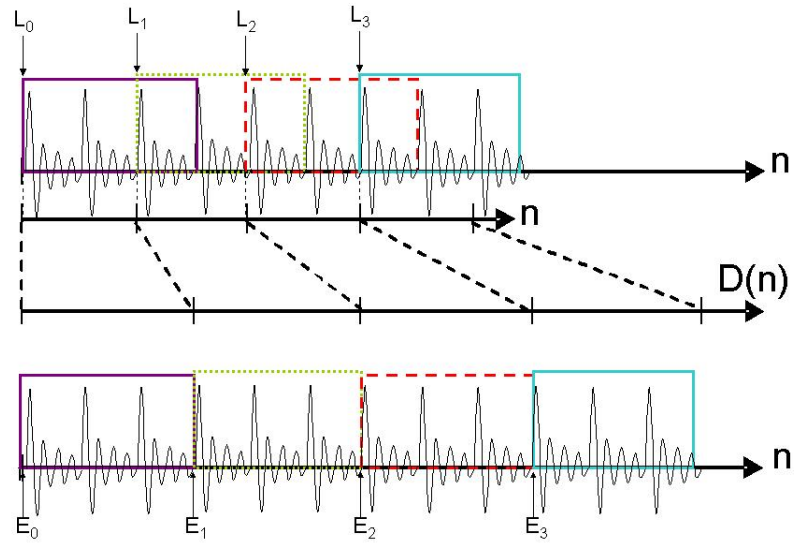


Figure 2.21 – Méthode TDHS : schéma de positionnement pour des fenêtres d'écriture rectangulaires ($FE=0$) et juxtaposées ($\alpha = 1,5$)

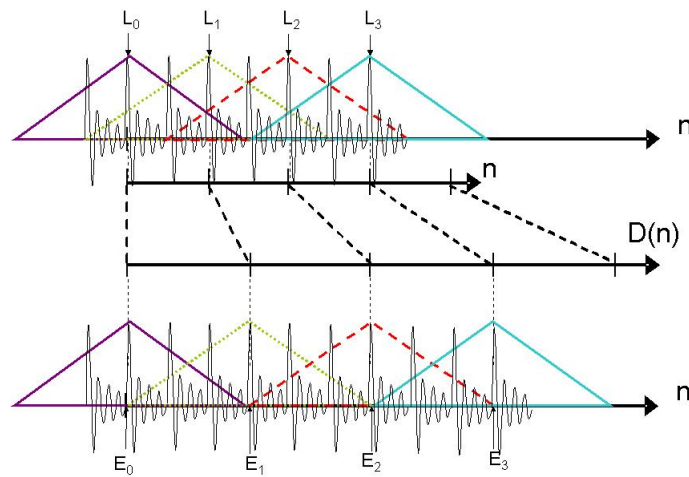


Figure 2.22 – Méthode TDHS : schéma de positionnement pour des fenêtres d'écriture triangulaires ($\alpha = 1,5$)

Ainsi, on remarque que la durée du segment inséré correspond à une période fondamentale du signal, ce qui explique l'absence de discontinuité de désynchronisation.

La méthode TDHS repose sur une estimation assez précise de la période fondamentale locale P_i (dont la méthode de détection n'est pas explicitée dans l'article de Malah). Mais Cox *et al.* [CCJ83] détaillent le fonctionnement d'un détecteur de périodicité, basé sur une modification de la fonction d'autocorrélation avec filtrage passe-bas (FIR de 8 coefficients coupant à 1 kHz) du signal.

Méthode SOLA ("Synchronous OverLap-Add")

En 1985, Roucos et Wilgus [RW85] proposent une méthode de dilatation-p appliquée à la parole qu'ils nomment SOLA (addition-recouvrement synchronisé). Son formalisme est exposé dans un contexte temps-fréquence. Il est issu de la méthode de Griffin et Lim [GL84] dont le principe est de produire un signal temporel à partir du spectre d'amplitude à court terme dilaté (voir la section sur les méthodes fréquentielles 2.3), en reconstruisant de manière itérative un spectre de phase à court terme qui lui soit cohérent. Le choix de l'estimation initiale du signal est important pour la rapidité de convergence de l'algorithme. Roucos et Wilgus suggèrent d'utiliser comme estimation initiale, le signal original retardé d'une durée K . Cette estimation initiale est explicitée dans le domaine temporel et ne nécessite pas d'itération supplémentaire. Cette méthode est donc beaucoup plus efficace en terme de puissance de calculs, et semble donner des résultats au moins aussi bons que ceux de Griffin et Lim pour la parole.

Nous décrivons ici une interprétation de l'algorithme proposé. L'aspect adaptatif de cette méthode se révèle lors du placement des marques d'écriture. En effet, les marques de lecture et la durée des fenêtres de granulation temporelle restent toujours fixes.

Les caractéristiques des variables pour les méthodes SOLA sont les suivantes :

- Les marques de lecture sont toutes placées aux instants arbitraires suivants :

$$L_i = iL$$

- Les marques d'écriture sont ensuite déduites des marques de lecture par la fonction de dilatation avec cependant une tolérance Γ_i qui permet d'éviter la discontinuité de désynchronisation et qui est calculée à partir du comportement de la fonction de corrélation (voir équation 2.19).

$$E_i = D(L_i) + \Gamma_i = i\alpha L + \Gamma_i$$

Il est à noter que les images des marques de lecture par la fonction de dilatation ne correspondent donc généralement pas aux marques d'écriture : $E_i \neq D(L_i)$. Ceci peut être à l'origine d'un problème d'anisochronie (irrégularité rythmique) étudié en section 3.1.

- Les fenêtres de granulation temporelle sont toutes identiques et de durée constante :

$$h_i(t) = h(t)$$

$$\text{Supp } h = H$$

Les formules pour les méthodes SOLA sont les suivantes :

- ⇒ Formule de granulation temporelle :

$$g_i(t) = h(t)s(t + iL) \tag{2.16}$$

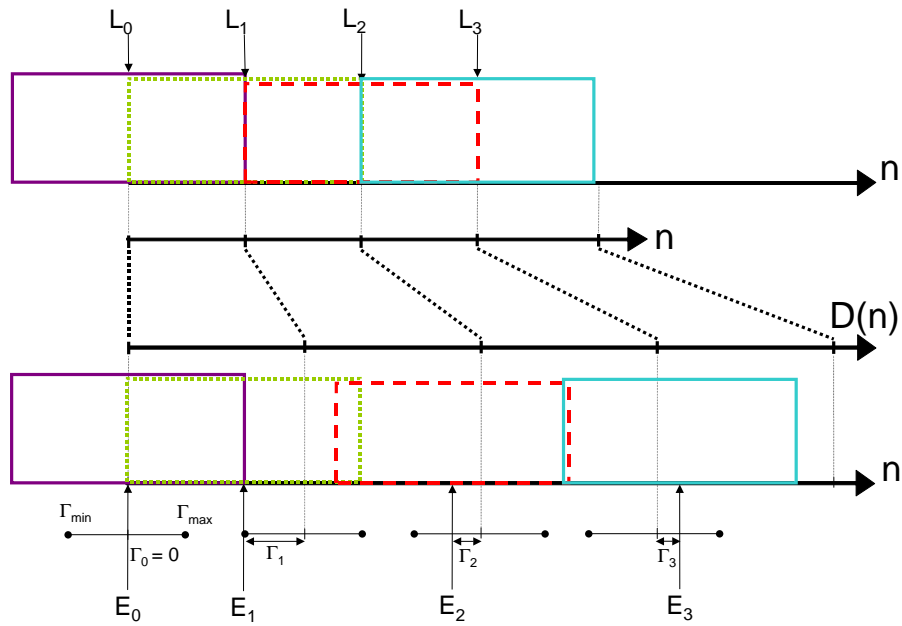


Figure 2.23 – Schéma de positionnement des fenêtres pour une méthode SOLA à fenêtres rectangulaires ($FE=0$) et juxtaposées ($\alpha = 1,5$)

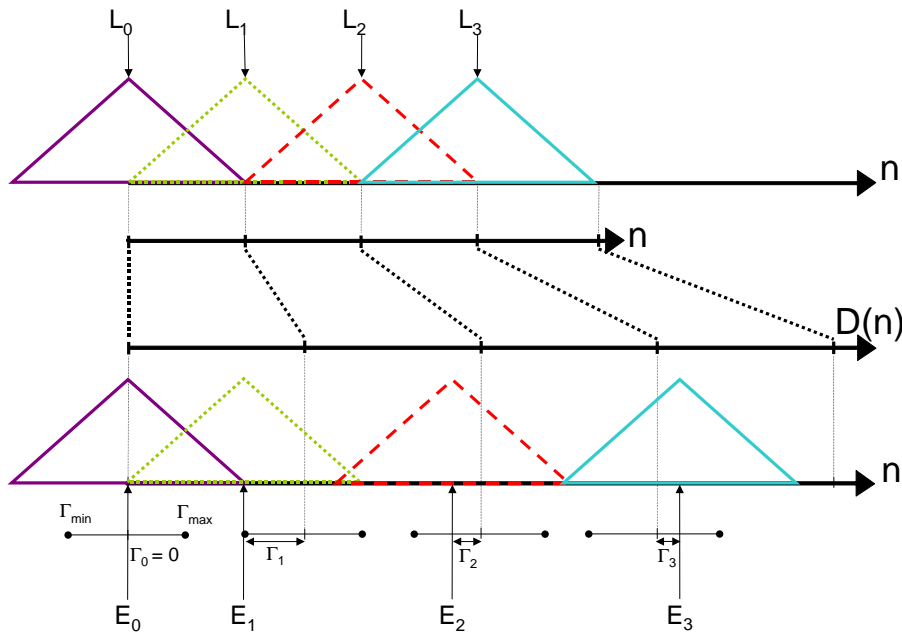


Figure 2.24 – Schéma de positionnement des fenêtres pour une méthode SOLA à fenêtres triangulaires ($\alpha = 1,5$)

⇒ Formule de construction temporelle pour la dilatation-p :

$$Dp[s](t) = \frac{\sum_i g_i(t - E_i)}{\sum_i h(t - E_i)} \quad (2.17)$$

La formule précédente s'écrit donc, en prenant en compte l'équation 2.16:

$$Dp[s](t) = \frac{\sum_i h(t - i\alpha L - \Gamma_i) s(t - i\alpha L - \Gamma_i + iL)}{\sum_i h(t - i\alpha L - \Gamma_i)} \quad (2.18)$$

Les figures 2.23 et 2.24 illustrent cette méthode dans le cas de fenêtres rectangulaires et triangulaires.

Le paramètre Γ_i , dont les valeurs sont comprises entre Γ_{min} (généralement négatif) et Γ_{max} (généralement positif), est évalué selon un critère de maximisation de l'intercorrélation normalisée entre le signal de sortie $s'(t)$ déjà constitué et le grain temporel $g_i(t - E_i)$ à ajouter :

$$CN_{s'g_i}(\Gamma) = \frac{\sum_{t=D(L_i)}^{D(L_i)+N_c-1} s'(t) g_i(t - D(L_i) - \Gamma)}{\left[\sum_{t=D(L_i)}^{D(L_i)+N_c-1} s'^2(t) \sum_{t=D(L_i)}^{D(L_i)+N_c-1} g_i^2(t - D(L_i) - \Gamma) \right]^{1/2}} \quad (2.19)$$

où N_c représente la durée sur laquelle est estimée la corrélation.

Cette maximisation assure que la procédure "OLA" (recouvrement-addition) moyenne le grain temporel à insérer avec la région la plus similaire dans le signal reconstruit.

La durée K_i du segment inséré équivalent (correspondant au segment inséré si la fenêtre de granulation temporelle est rectangulaire) est donnée par :

$$K_i = |\Delta E_i - \Delta L_i| = |(E_i - E_{i-1}) - (L_i - L_{i-1})| = L(\alpha - 1) + \Gamma_i - \Gamma_{i-1}$$

Cette durée est ajustée de manière à ce que les segments de signal se recouvrant soient le plus similaires possible, ce qui revient à insérer une ou plusieurs périodes fondamentales dans le cas d'un signal quasi-périodique.

Des valeurs de Γ_i toujours nulles revient donc à insérer, comme dans les méthodes aveugles, des segments de durée constante $L(\alpha - 1)$.

Makhoul et El-Jaroudi [MEJ86] utilisent la méthode SOLA et testent différentes formes de fenêtres pour le fondu-enchaîné : le simple moyennage (fenêtre constante), une fonction cosinusoidale croissante, et une fonction linéaire (rampe). Les deux dernières donnent un résultat satisfaisant, bien que la fonction linéaire soit plus simple à calculer. La réverbération associée à ce traitement est attribuée à l'impossibilité de synchroniser plus finement que la période d'échantillonnage ($125\mu s$ pour $Fe = 8$ kHz). Le signal est donc suréchantillonné d'un facteur 2 avant d'être dilaté-p puis sous-échantillonné. Cette technique ne devient donc plus nécessaire dès que la fréquence d'échantillonnage dépasse 16 kHz.

Wayman et Wilson [WW88] utilisent la méthode SOLA dans le but d'une compression de données du signal. Leur innovation réside dans le transfert du paramètre Γ_i pour chaque itération afin de reconstruire le signal plus fidèlement lors de la décompression. Cependant, il n'amène aucun élément nouveau quant à la technique de dilatation-p proprement dite.

WSOLA ("Waveform Similarity OverLap-Add")

En 1993, Verhelst et Roelands [VR93] proposent une méthode qu'ils nomment WSOLA (addition-recouvrement par similarité de la forme d'onde). Cette méthode s'inspire fortement de la méthode SOLA, et sa théorie est également développée dans un contexte temps-fréquence.

Ils constatent que dans la méthode SOLA, le dénominateur de l'équation 2.18 ne peut être rendu constant à cause de la tolérance accordée à travers Γ_i , ce qui alourdit les calculs. Ils proposent de rendre régulière la position des grains d'écriture afin de rendre constant ce dénominateur, et de compenser la tolérance ainsi disparue sur les marques d'écriture par une tolérance sur les marques de lecture.

L'interprétation qu'il donne de cette tolérance sur les marques de lecture est la recherche d'une similarité de la forme d'onde. Le critère de similarité de la forme d'onde peut s'exprimer ainsi : la forme d'onde du signal de sortie à un instant t doit être le plus similaire possible à celle du signal original dans un voisinage de l'instant $D^{-1}(t)$.

Les caractéristiques des variables pour les méthodes WSOLA sont les suivantes :

→ Les marques d'écriture sont placées aux instants arbitraires suivants :

$$E_i = iE$$

→ Les marques de lecture sont ensuite déduites des marques d'écriture avec une tolérance Γ_i qui permet d'éviter la discontinuité de désynchronisation :

$$L_i = D^{-1}(E_i) + \Gamma_i = iE/\alpha + \Gamma_i$$

Il est à noter que les images des marques d'écriture par la fonction de dilatation inverse ne correspondent généralement pas aux marques de lecture : $L_i \neq D^{-1}(E_i)$. Ceci peut être à l'origine d'un problème d'anisochronie (irrégularité rythmique) étudié en section 3.1.

→ Les fenêtres de granulation temporelle sont toutes identiques et de taille constante :

$$h_i(t) = h(t)$$

$$\text{Supp } h_i = H$$

La caractéristique de la méthode WSOLA est que la somme des fenêtres de granulation temporelle centrées sur les marques d'écriture est égale à l'unité. Il n'y a donc pas besoin de normaliser la construction temporelle.

Les formules pour les méthodes WSOLA sont les suivantes :

⇒ Formule de granulation temporelle :

$$g_i(t) = h(t)s(t + L_i) = h(t)s(t + iE/\alpha + \Gamma_i) \quad (2.20)$$

⇒ Formule de construction temporelle pour la dilatation-p :

$$Dp[s](t) = \sum_i g_i(t - iE) \quad (2.21)$$

La formule précédente s'écrit donc, en prenant en compte l'équation 2.20:

$$Dp[s](t) = \sum_i h(t - iE)s(t - iE + iE/\alpha + \Gamma_i) \quad (2.22)$$

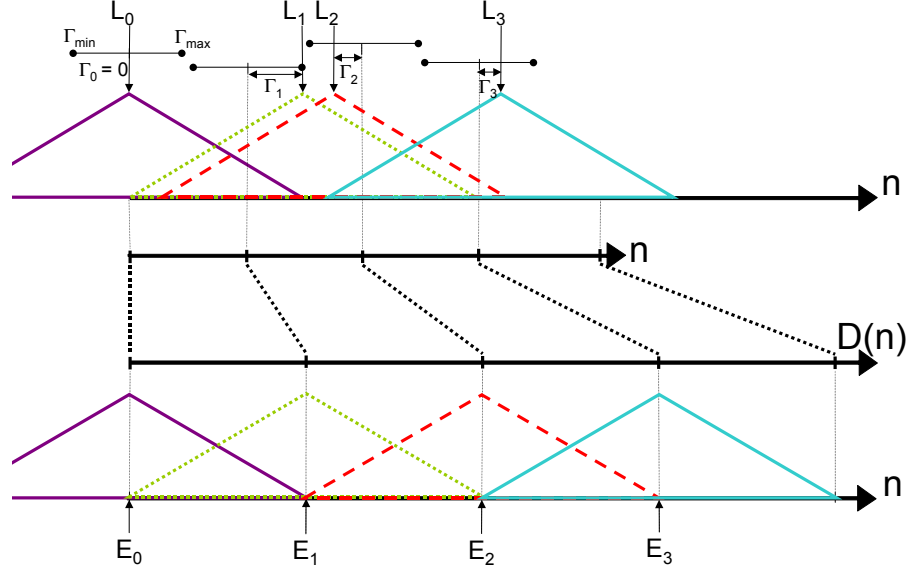


Figure 2.25 – Schéma de positionnement des fenêtres pour une méthode WSOLA à fenêtres triangulaires ($\alpha = 1,5$)

La figure 2.25 illustre cette méthode dans le cas de fenêtres triangulaires.

La valeur Γ_i est choisie par maximisation d'une mesure de similarité (comme l'inter-corrélation) entre le grain temporel qui suit naturellement le précédent grain temporel sélectionné et le grain temporel centré autour de la marque de lecture correspondant à la marque d'écriture désirée. En pratique, la mesure de similarité est effectuée sur les signaux en ne tenant pas compte des fenêtres de pondération $h_i(t)$.

Il en résulte la formule de la fonction de corrélation suivante :

$$C(\Gamma) = \sum_{t=0}^{H-1} s(t + D^{-1}(E_{i-1}) + \Gamma_{i-1} + E) \cdot s(t + D^{-1}(E_i) + \Gamma)$$

La mesure de similarité généralement utilisée est plutôt la fonction de corrélation normalisée [Lar93, VR93, SVW94, Ver00], donnée par :

$$CN(\Gamma) = \frac{C(\Gamma)}{\left[\sum_{t=0}^{H-1} s^2(t + D^{-1}(E_{i-1}) + \Gamma_{i-1} + E) \sum_{t=0}^{H-1} s^2(t + D^{-1}(E_i) + \Gamma) \right]^{1/2}}$$

Puisque seules les valeurs relatives de cette corrélation normalisée nous intéressent (on cherche juste à la maximiser), le premier terme du dénominateur devient inutile puisqu'il est constant. L'équation devient donc simplement :

$$CN(\Gamma) = \frac{C(\Gamma)}{\left[\sum_{t=0}^{H-1} s^2(t + D^{-1}(E_i) + \Gamma) \right]^{1/2}}$$

L'intercorrélation normalisée nécessite une puissance de calcul élevée, mais il est possible de l'optimiser en sous-échantillonnant le signal [Lar93, Lar98] ou bien en utilisant la FFT [OS89, Lar93, TL00].

Il est également possible d'utiliser la fonction de différence d'amplitude moyenne ou AMDF ("Average Magnitude Difference Function") [VR93, SVW94, Lar98, PDH99, Ver00], moins chère en terme de calcul, et donc plus adaptée aux contraintes du temps réel [TL00], mais également plus sensible au bruit :

$$AMDF(\Gamma) = \sum_{t=0}^{H-1} |s(t + D^{-1}(E_{i-1}) + \Gamma_{i-1} + E) - s(t + D^{-1}(E_i) + \Gamma)|$$

Les différents paramètres de cet algorithme doivent être choisis judicieusement pour éviter certains artefacts :

- La durée de la fenêtre de granulation doit être fixée au dessus d'un minimum H_{min} correspondant à la période fondamentale la plus grande présente dans le signal, sinon une transposition fréquentielle sera appliquée à toutes les fréquences plus basses.
- Cependant, la durée de la fenêtre de granulation doit aussi être fixée en dessous d'un maximum H_{max} , car les événements au sein du grain temporel ne sont pas dilatés, ce qui peut mener à un problème d'anisochronie.
- La tolérance Γ_i doit être assez grande pour pouvoir embrasser la période fondamentale la plus élevée du signal, afin de dilater correctement les fréquences les plus basses.
- Cependant, la tolérance ne doit pas non plus être trop élevée sous peine d'être confronté au problème d'anisochronie.

Ainsi, des compromis doivent être effectués, mais ces choix sont difficiles à effectuer.

Pour la voix, le choix des paramètres n'est pas vraiment critiques, et de bons résultats peuvent être obtenus avec les valeurs suivantes [SVW94] : $H=20$ ms, $\Gamma_{max} = 5ms$. Le résultat reste sans artefact pour $\alpha < 0,5$ (et même jusqu'à 0,25), mais une réverbération, un bourdonnement et des saccades sont décelés pour $\alpha > 2$, expliqué par la répétition régulière de segments bruités introduisant une corrélation, colorant les portions non-voisées. Un bon compromis sur la durée de la fenêtre et sur la tolérance reste donc possible pour la voix pour des taux de dilatation peu élevés.

En revanche, les résultats sur la musique sont moins bons que sur la voix. Globalement, le choix de la durée de la fenêtre s'avère plus critique (40 à 160 ms selon signal). Des problèmes pour $\alpha > 1$ sont remarqués dans le cas de sources polyphoniques. Les meilleurs résultats sont obtenus pour $\alpha \approx 1$ et $\alpha < 1$, bien que des artefacts soient présents pour des modulations (vibrato) de sources monophoniques. Un bon compromis sur la fenêtre et sur la tolérance n'est plus vraiment possible pour des sons musicaux complexes.

Méthode SOLAFS ("Synchronous OverLap-Add Fixed Synthesis")

Quelques mois avant l'apparition de la méthode WSOLA, Hejna [HMC92] publie un brevet sur une méthode de dilatation-p nommée SOLAFS (addition-recouvrement synchronisé - (fenêtre de) synthèse fixe) qui semble être identique en tous points à cette dernière.

Cependant, l'auteur note que le calcul de la mesure de similarité n'est pas nécessaire à chaque itération. En effet, lorsque la tolérance sur la marque de lecture est telle qu'il est autorisé

de sélectionner le grain temporel suivant naturellement le dernier grain du signal de sortie ($L_i = L_{i-1} + E$), alors il est évident que ce grain est sélectionné et il n'est pas utile d'effectuer une mesure de similarité pour arriver à cette conclusion. L'auteur nomme cette particularité "prédiction".

La figure 2.25 illustre cette méthode. On remarque que les fenêtres résultantes englobent plusieurs fenêtres triangulaires pour lesquelles $\Gamma_i = 0$.

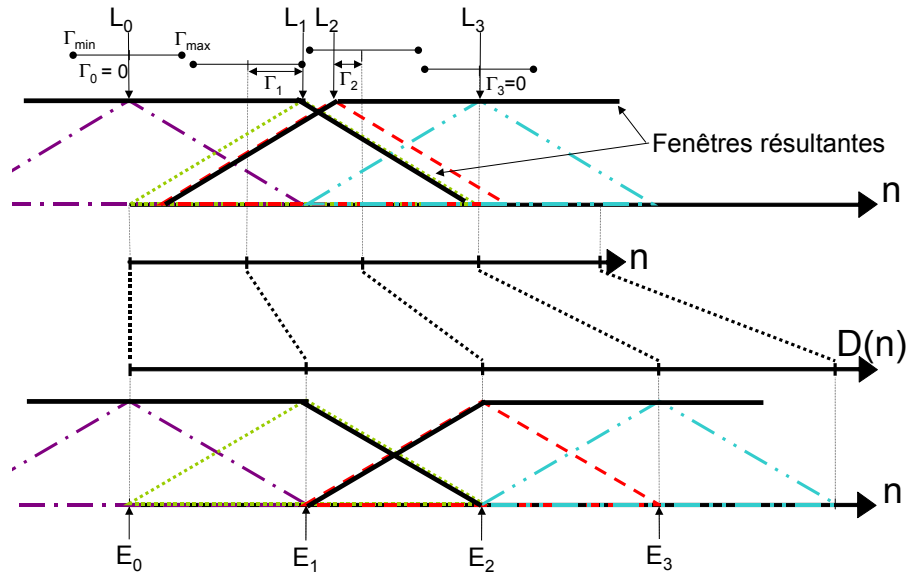


Figure 2.26 – Schéma de regroupement des fenêtres pour une méthode SOLAFS ($\alpha = 1,5$)

La notion de "prédiction" met en évidence la représentation par duplication, et le fait que plusieurs petites fenêtres de lecture peuvent être interprétées comme une seule grande. Ainsi, si l'on observe les grandes fenêtres résultantes, leur durée est dépendante du signal. Dans ce cas, il devient très adapté d'étudier la transformation à travers un diagramme d'entrée/sortie.

Méthode EDSOLA ("Edge Detection Synchronized OverLap-Add")

Nous appelons EDSOLA (Détection de transitoires - addition-recouvrement synchronisé) les méthodes qui réalisent une détection de transitoire (explicite ou non) afin d'éviter un redoublement audible de ce dernier [Dat87, Lar93, LKK97, Lar00, PBKM00]. Elles possèdent la possibilité de décaler le segment inséré en fonction du signal, ce qui n'est pas le cas dans les méthodes que nous avons étudiées jusqu'ici.

Les algorithmes utilisés sont des extensions de WSOLA (elle-même méthode dérivée de SOLA) où l'on ne s'intéresse plus qu'aux grains temporels qui se déplacent relativement aux précédents. Ainsi, de nombreuses estimations de similarité sont évitées.

La véritable différence avec les méthodes WSOLA ou SOLAFS réside dans le fait que dans ces dernières, les grains temporels de lecture sont placés à des marques de lecture fixées (à une tolérance Γ_i près) puisque les fenêtres sont toutes de taille fixe. Il en résulte que lorsqu'une estimation de similarité doit être effectuée, le grain temporel de référence est entièrement déterminé, même s'il comporte un transitoire, ce qui mène parfois à un redoublement audible. Ce n'est pas le cas des méthodes EDSOLA, pour lesquelles la mesure de similarité peut être retardée afin

d'avoir la possibilité de l'effectuer dans une partie plus stationnaire du signal, dans les limites bien entendu des déformations rythmiques engendrées par cette attente d'insertion de segment. La figure 2.27 en donne une illustration : une zone transitoire se trouve à l'emplacement où devrait avoir lieu une insertion de segment (comparé à la figure 2.26). L'insertion est donc retardée afin de ne pas dupliquer la zone transitoire.

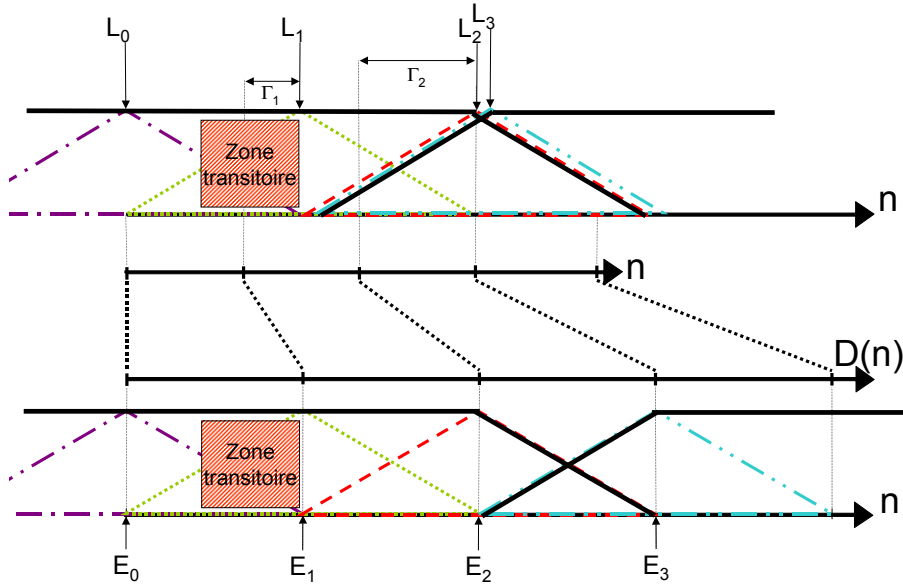


Figure 2.27 – Schéma de regroupement des fenêtres pour une méthode EDSOLA ($\alpha = 1,5$)

En 1987, Dattorro [Dat87] présente succinctement l'algorithme employé dans la machine Lexicon 2400, dont le fonctionnement se révèle être une sorte de généralisation des méthodes WSOLA ou SOLAFS : en effet, avec ces méthodes seul le choix de la durée du segment inséré est possible, mais lorsque la "prédiction" n'est plus valide (le grain temporel "naturel", c'est-à-dire celui qui suit naturellement le précédent, n'est pas accessible même avec la tolérance) on impose d'utiliser un grain du voisinage de la marque $D^{-1}(E_i)$. Si ce grain comporte un transitoire ou une sifflante, il est probable que celui-ci soit auditivement dupliqué/supprimé [Dat98].

La méthode proposée par Dattorro est basée sur un pré-traitement réalisé sur un signal sous-échantillonné d'un facteur 4, dans lequel une mesure de confiance estime à la fois le degré de similarité de la forme d'onde (identique à SOLAFS et WSOLA) et la probabilité de présence de transitoire. Si un transitoire est détecté, la duplication est retardée et une nouvelle mesure de confiance est réalisée à l'instant suivant. Le segment inséré est donc généralement un segment très périodique.

L'algorithme de Dattorro [Dat87] est une méthode permettant d'éviter de dupliquer ou supprimer des transitoires de manière implicite. En effet, l'indice de confiance permet de choisir le point d'insertion qui favorise l'insertion d'un segment le plus périodique possible. Le fondu-enchaîné est proportionnel ici à la distance du saut (K).

En 1993, Laroche [Lar93] évoque lui aussi la possibilité d'optimiser la position des points d'insertion. En 2000, il met en pratique cette méthode dans un brevet [Lar00]. La détection de transitoire est réalisée par comparaison d'énergie entre de courts segments successifs, et la

détection de périodicité est réalisée avec un seuillage adaptatif. Le choix du point d'insertion est basé sur critère de périodicité (on n'insère pas si la mesure de similarité est trop faible). Néanmoins, lorsque la limite d'anisochronie est atteinte, l'insertion devient inévitable, même si la valeur du maximum de la corrélation est faible. Ainsi, il est possible qu'il existe un transitoire dans le segment inséré.

2.2.4 Méthodes recourant à des décompositions préalables

Les méthodes décrites dans cette partie n'offrent aucune idée nouvelle quant à la technique de dilatation-p en elle-même. Elles doivent leur originalité au fait qu'elles appliquent l'une des techniques précédemment décrites à une décomposition temporelle⁵ du signal original.

Grâce à une telle décomposition, il est possible de transformer, indépendamment ou non, mais de manière plus adaptée, chacun des signaux temporels issu de la décomposition, comme illustré en figure 2.28.

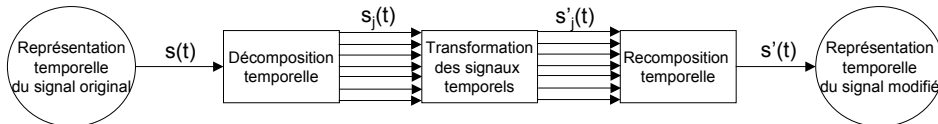


Figure 2.28 – Illustration de la transformation de la représentation temporelle décomposée

Décomposition en sous-bandes fréquentielles

Une décomposition en sous-bandes fréquentielles est une décomposition du signal temporel $s(t)$ en N_{sb} signaux temporels $s_j(t)$ ($j \in [1, N_{sb}]$) appelés "sous-bandes" et correspondant à différentes bandes de fréquences.

La représentation obtenue (qui est une somme de représentations temporelles) reflète grossièrement l'aspect "analyseur de fréquence" de l'oreille.

Nous considérons ici qu'une décomposition temporelle en sous-bandes fréquentielles n'est pas une transformation du domaine fréquentiel mais bien une transformation temporelle puisque le signal d'origine et les signaux résultants sont tous des représentations temporelles.

D'une part, il n'est réalisé aucune modification au niveau de la représentation fréquentielle du signal (même si l'on peut utiliser le domaine fréquentiel pour réaliser la décomposition), d'autre part cette décomposition est généralement réalisée dans le domaine temporel par le produit de convolution (ou filtrage) adéquat, lorsque le nombre de sous-bandes n'est pas trop important.

Ainsi, on reste bien dans nos conditions d'appellation "méthodes temporelles" selon lesquelles on utilise uniquement la représentation temporelle du signal.

Il semble important de pouvoir retrouver le signal original lorsqu'aucune modification n'est effectuée sur les sous-bandes. On parle alors de méthode de décomposition à reconstruction parfaite.

La technique généralement utilisée consiste en l'utilisation d'un banc de filtres réalisé avec des structures à réponse impulsionnelle finie (FIR), seules capables de conserver une phase linéaire, nécessaire à une haute qualité sonore. Le signal issu de la j^{ieme} sous-bande est alors donné par :

$$s_j(t) = (s * q_j)(t)$$

avec $*$ désignant le produit de convolution, et q_j la réponse impulsionnelle du filtre correspondant à la j^{ieme} sous-bande.

5. On appelle décomposition temporelle la décomposition du signal original en un certain nombre de signaux temporels qu'il suffit de sommer pour retrouver le signal original. Chacun de ces signaux temporels possède une caractéristique propre : bande de fréquence donnée dans le cas d'une décomposition en sous-bandes fréquentielles, sinusoïde ou bruit dans le cas d'une décomposition sinus+bruit...

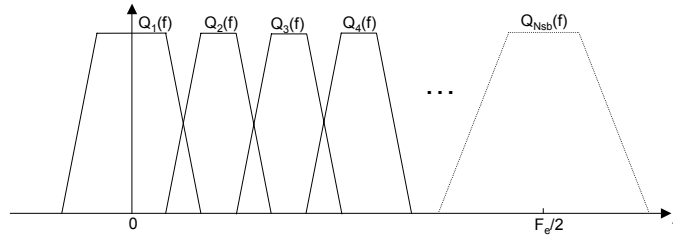


Figure 2.29 – Illustration de la Réponse en Fréquence des filtres utilisés pour la décomposition en sous-bandes

Ces filtres sont tels que la somme de leur réponse en fréquence est égale à l'unité. Soit Q_j , la réponse en fréquence complexe du j^{ieme} filtre, on a :

$$\sum_{j=1}^{N_{sb}} |Q_j(f)| = 1 \quad \forall f \in [0, F_e/2]$$

Comme tous ces filtres ne peuvent pas être parfaitement juxtaposés, ce qui impliquerait une pente infinie et donc une réponse impulsionnelle infinie, ils doivent donc se recouvrir partiellement, comme illustré en figure 2.29.

Toutes les méthodes décrites auparavant peuvent être appliquées telles quelles aux différents signaux temporels $s_j(t)$ issus de la décomposition en sous-bandes fréquentielles, avant de recomposer un signal modifié $s'(t)$.

"Subband WSOLA"

En 1994, Spleesters *et al.* [SVW94] proposent une extension à l'audio en général de leur méthode WSOLA, développé à l'origine pour la parole [VR93].

L'idée de cette amélioration provient de la constatation que WSOLA fonctionne bien sur la voix car le signal est généralement périodique, mais le signal audio est autrement plus complexe.

Il leur vient donc l'idée de décomposer le signal original $s(t)$ en 16 sous-bandes de largeur de bande égale $s_j(t)$, de traiter indépendamment chacune de ces sous-bandes par la méthode WSOLA qui leur fournit 16 signaux temporels dilatés $s'_j(t)$, puis de reconstituer un signal large bande en sommant toutes ces sous-bandes dilatées :

$$s'(t) = \sum_{j=1}^{16} s'_j(t)$$

Ils réalisent la décomposition à l'aide d'un banc de 16 filtres à phase linéaire. Le banc de filtres de synthèse n'est pas utilisé, bien que celui-ci devrait être nécessaire pour restreindre la bande passante des sous-bandes aux fréquences adéquates, car l'algorithme WSOLA est non-linéaire, et donc susceptible d'introduire des fréquences indésirables.

L'optimisation des paramètres de la méthode mène à employer de longues fenêtres de granulation temporelle pour les basses fréquences et de courtes fenêtres pour les hautes fréquences. Cette technique est donc sensée être une solution au problème des sons inharmoniques.

Leurs expérimentations montrent qu'il subsiste toujours des défauts mais cette méthode semble s'avérer quand même meilleur que WSOLA. Pour une valeur de tolérance élevée, il

est observé une perte de puissance et de pureté sur les aspects dynamiques et rythmiques du son.

L'expérience d'un simple décalage temporel entre les sous-bandes sans autre modification ($\alpha = 1$) montre une perte de synchronisation "inter-bande" qui mène à un artefact. L'application de méthodes temporelles agissant indépendamment sur les différentes sous-bandes mène donc à une désynchronisation des phases des sous-bandes.

Cette désynchronisation des phases est généralement catastrophique pour des signaux transitoires. De plus, elle peut donner de mauvais résultats dans le cas où une composante, située entre les 2 fréquences centrales des filtres de décomposition, sont traitées différemment dans les 2 sous-bandes. Il peut en résulter des interférences destructives générant une modulation d'amplitude pour cette composante.

SASOLA ("Subband Analysis Synchronous OverLap-and-Add")

En 1998, Tan et Lin [LT98, TL00] décrivent une méthode nommée SASOLA (Analyse en sous-bandes et addition-recouvrement synchronisé). Cette méthode est extrêmement similaire à la méthode "Subband WSOLA".

Elle est basée sur une décomposition en 8 ou 17 sous-bandes, réalisée par l'utilisation de filtres miroir en quadrature qui possèdent la propriété de reconstruction parfaite [SB86]. Chacune des sous-bandes est ensuite traitée indépendamment.

Pour $\alpha < 1$, la longueur des fenêtres de granulation temporelle est fixée à 40 ms pour toutes les sous-bandes. En revanche, pour $\alpha > 1$, de grandes fenêtres de granulation provoquent de l'écho et du "bégaiement"; leur durée est donc réduite en divisant par α pour la première sous-bande, et par 2α pour les suivantes.

La mesure de similarité utilisée est l'autocorrélation normalisée, calculée dans le domaine fréquentiel par FFT.

Cette méthode permet de résoudre en partie le problème des sons inharmoniques et polyphoniques. En effet, au lieu de chercher une durée K qui soit un compromis pour tous les partiels présents dans le son original, on cherche N_{sb} paramètres K_j qui soient des compromis uniquement au sein de chaque sous-bande, dans lesquelles le signal est plus "prédictible".

Pour $\alpha > 1,5$, les échos et bégaiements observés avec la méthode SOLA disparaissent avec l'utilisation de SASOLA.

Cependant, des problèmes de phase entre les canaux fréquentiels apparaissent. En effet, un transitoire est bien respecté lorsque tous les canaux sont bien en phase. Or cette méthode crée intrinsèquement des déphasages menant à un étalement du transitoire.

Selon les auteurs, l'augmentation du nombre de sous-bandes (de 8 à 17 canaux) améliore la qualité sonore, "bien qu'une légère distorsion de phase soit audible".

Décomposition hybride

Dans ce type de méthode, la décomposition n'est plus uniquement fondée sur une discrimination fréquentielle du signal. Des aspects perceptifs du signal sont pris en compte tels que la partie tonale, le bruit ou encore les transitoires. Le principe est donc de réaliser des modélisations successives du signal. Chaque composante modélisée est soustraite au signal original, ce qui fournit un signal résiduel sur lequel une nouvelle modélisation peut être effectuée. Le signal résiduel final, communément appelé résidu, est souvent assimilé à du bruit. S'il est sommé à tous les autres signaux modélisés, on doit obtenir le signal original. L'utilisation d'une décomposition hybride permet d'optimiser le traitement sur les différents signaux modélisés.

Daudet [Dau00] remarque la possibilité d'utiliser une décomposition hybride (tonal/transitoire/stochastique) pour effectuer une dilatation-p.

Duxbury, Davies et Sandler [DDS02] évoquent également une telle décomposition mais constatent que les transitoires ne sont plus verrouillés en phase avec les parties périodiques, ce qui mène à des distorsions d'amplitude.

2.2.5 Bilan des "méthodes temporelles"

La caractéristique des méthodes temporelles selon laquelle aucun des grains temporels n'est modifié, est une raison pour laquelle les résultats obtenus avec ce type de méthode sont généralement de très haute qualité. En effet, la forme d'onde d'un signal sonore est extrêmement fragile, c'est-à-dire qu'une modification infime de celle-ci peut être décelée par l'oreille. Les grains temporels contiennent des détails acoustiques riches qui sont difficilement reproduits par d'autres types de méthodes [Ver00].

Ainsi, en conservant intacts les grains temporels, on s'assure d'une qualité sonore parfaite pendant leur durée. Bien sûr, des défauts audibles sont parfois présents, mais on peut dire que ceux-ci sont dus aux conséquences du recouvrement des grains, et non aux grains eux-mêmes. Pour de faibles taux de dilatation ($\alpha \simeq 1$), relativement peu de raccords sont présents, et donc susceptibles de produire des artefacts, expliquant ainsi la bonne qualité généralement obtenue [SVW94].

D'autre part, les méthodes temporelles ne font pas d'hypothèse explicite sur le signal, ce qui explique leur robustesse vis-à-vis de l'éclectisme des signaux à traiter.

Les méthodes aveugles possèdent des défauts rédhibitoires sur les sons harmoniques.

Toutes les méthodes adaptatives sont très similaires entre elles puisqu'elles se basent toutes sur une estimation d'un indice de périodicité pour sélectionner la durée du segment inséré. Elles permettent ainsi d'éviter les discontinuités de désynchronisation sur les signaux harmoniques.

En présence de bruit, l'estimation de périodicité renvoie une valeur aléatoire (la corrélation d'un signal décorrélé est stochastique), ce qui ne provoque généralement aucun artefact audible et permet même d'éviter une insertion régulière de segments.

Lorsque le signal est inharmonique, il n'est pas toujours possible de trouver un bon compromis sur la durée K pour qu'aucun des partiels ne subissent de discontinuité. C'est ainsi le cas pour la plupart des notes de piano, où la raideur de la corde est à l'origine de l'inharmonicité du signal [VC93]. On entend alors des discontinuités sur certains partiels aussi bien sur des sons de synthèse (sons [32, 33]) que sur des sons réels (sons [34, 35]). On est confronté au même problème dans le cas des sons polyphoniques, cependant les résultats sur ce type de sons sont généralement meilleurs car les défauts de discontinuités sont souvent masqués par les autres sons.

L'insertion de segments extrêmement courts comparés à la période fondamentale peut ne pas provoquer de discontinuité de désynchronisation audible, mais dans ce cas, la période fondamentale augmente (ou diminue) de la durée du segment dupliqué (ou supprimé), ce qui signifie une transposition de la fréquence. On réalise alors une sorte de rééchantillonnage (dilatation ET transposition) au lieu d'une dilatation-p.

De même, l'insertion d'un segment plus court (mais du même ordre de grandeur) qu'une période du signal original mène à une discontinuité de désynchronisation pour cette fréquence. Il en résulte que la fréquence minimale recherchée par l'estimation de périodicité doit être inférieure à la fréquence la plus basse du signal.

Il est cependant délicat d'insérer/supprimer des segments trop longs car il en résulte une déformation rythmique traitée en section 3.1.

Ces méthodes souffrent parfois de défauts liés à la répétition ou l'amputation de transitoires d'attaques, car ces transitoires peuvent se trouver à l'intérieur du segment inséré. Il est parfois possible de résoudre ce problème en faisant varier le paramètre que l'on nomme "point d'insertion" (en référence au point d'insertion du segment inséré dans la représentation par

”segmentation”).

La dilatation-p obtenue par une méthode de collage conserve généralement la durée des transitoires et ne modifie pas la position des formants. En conséquence de quoi, la transposition-p associée transpose les transitoires et modifie les formants, c’est pourquoi elle n’est habituellement pas utilisée pour effectuer une transposition-p dans un but musical (changement de note d’un instrument).

En conclusion, on peut dire que les méthodes temporelles donnent d’excellents résultats sonores pour de faibles taux de dilatation (+/-20%). Elles sont de plus assez robustes face à différents types de signaux. Cependant, elles montrent leurs limites dans le cas de sons inharmo- niques, et des compromis entre anisochronie et discontinuité de désynchronisation doivent être réalisés pour des séquences rythmiques accompagnées de basses fréquences.

2.3 Méthodes fréquentielles

Nous conservons le terme "méthodes fréquentielles"⁶ généralement adopté dans la littérature pour caractériser les méthodes de transformation-p faisant appel à une **modification de la Représentation de Fourier à Court Terme** (ou RFCT), obtenue grâce à une Transformée de Fourier à Court Terme (ou TFCT, dont la traduction anglaise est "Short Time Fourier Transform" ou STFT).

Les méthodes fréquentielles sont un cas particulier de méthodes dites "temps-fréquence" (abordées en section 2.4), ayant la propriété de posséder un nombre important (une puissance de 2 afin de pouvoir bénéficier de l'algorithme de FFT : typiquement 512, 1024 ou 2048) de sous-bandes fréquentielles de *largeur constante*. Elles sont généralement assimilées à un système d'analyse/synthèse décrit dans la suite : le **Vocodeur de Phase**.

2.3.1 Principe général des méthodes fréquentielles

Le principe de ces méthodes repose sur la modification d'une représentation du signal décrivant la répartition de l'information en fonction du temps et de la fréquence. La transformation consiste à étirer "l'image" (les valeurs complexes fonction du temps et de la fréquence) de cette représentation selon sa dimension temporelle pour effectuer une dilatation-p, ou selon sa dimension fréquentielle pour effectuer une transposition-p.

Cette transformation est illustrée en figure 2.30 à travers le simple étirement (dilatation mathématique) du spectrogramme (ensemble des modules, soit les racines carrées de la puissance, de chacune des valeurs complexes correspondant aux points de l'image). Malheureusement, les résultats sonores associés sont de très mauvaise qualité (sons [36, 37, 38]) car le spectrogramme n'offre qu'une partie de l'information du signal original. L'autre partie, appelée phasogramme (ensemble des phases, soit les arguments, de chacune des valeurs complexes correspondant aux points de l'image), contient une information extrêmement importante, qui est plus difficilement interprétable, et surtout plus difficilement modifiable. De plus, chacune des valeurs d'une telle image ne peut être modifiée indépendamment des autres, ce qui complique la transformation. Nous présentons donc les approches permettant de résoudre les problèmes rencontrés.

Les méthodes fréquentielles ont été beaucoup décrites dans la littérature. Pour une description générale et théorique, on pourra se référer à [Por76, Por81b, ML95, VH96, Lar98, Ver00, AKZ02], pour une description plus intuitive à [Moo78, Dol86, Moo90, Roa96] et pour une description de l'implantation à [Por76, GS85, Moo90, DGBA00, AKZ02].

Afin de faire le lien avec les méthodes temporelles dans le cas de la dilatation-p, nous pouvons présenter les méthodes fréquentielles sous la forme de méthodes temporelles avec décomposition en sous-bandes, dont les grains temporels sont modifiés avant la sommation temporelle. Cette manière originale de présenter les méthodes fréquentielles est illustrée en figure 2.31.

Décomposition en sous-bandes de largeur constante et modification (réalisée dans le domaine fréquentiel) des grains temporels, sont les deux caractéristiques de ce que nous appelons les

6. Nous pensons que le terme "méthodes fréquentielles" est utilisé par un abus de langage, qui provient sûrement de la manière d'implanter ce genre de méthode sous forme de FFT/IFFT (comme par exemple le "vocodeur de phase") et ce, de façon quasi-systématique car pratique et rapide. En effet, une méthode à proprement parler "fréquentielle" devrait agir uniquement sur des données issues de ce domaine, c'est-à-dire indépendamment de toute évolution temporelle.

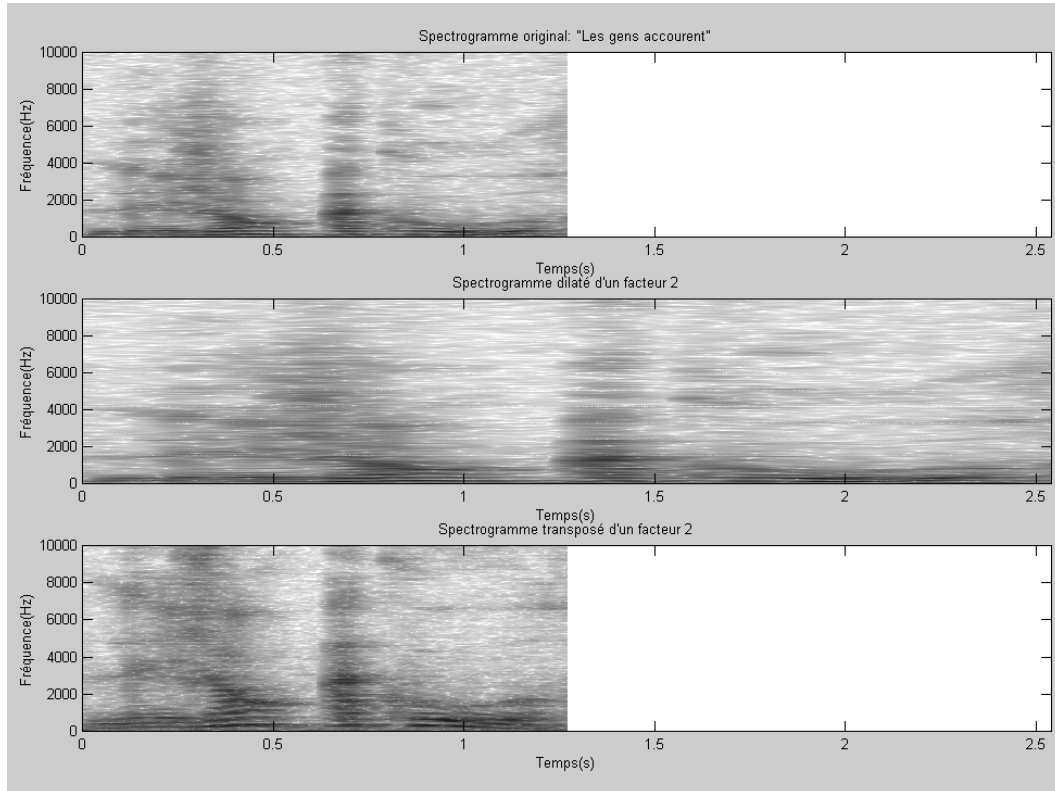


Figure 2.30 – *Spectrogramme original, dilaté et transposé d'une voix d'homme*

méthodes fréquentielles.

La décomposition en sous-bandes peut être effectuée conjointement avec la transformation (Transformée de Fourier) qui permet de passer dans le domaine fréquentiel. Ainsi, décomposition en sous-bandes, granulation temporelle et Transformée de Fourier sont avantageusement remplacés par l'opération appelée "Analyse à Court Terme", alors que Transformée de Fourier Inverse, sommation des grains et recomposition des sous-bandes sont remplacés par l'opération appelée "Synthèse à Court Terme".

On aboutit finalement à une transformation dont le principe est de modifier les coefficients de ce qui est appelé Transformée de Fourier à Court Terme.

Introduction à la Transformée de Fourier à Court Terme

Dans cette partie, nous formalisons la Transformée de Fourier à Court Terme (TFCT).

La solution au problème de la transformation-p de facteur α est évidente lorsqu'il s'agit de traiter un signal $s(t)$ composé d'une sinusoïde tronquée de fréquence ω et de support T . En effet, le signal dilaté s'écrit de la même manière que le signal original, c'est-à-dire $s'(t) = \sin(\omega t)$, mais sur un support temporel dilaté de α par rapport à l'original, soit αT . Le signal transposé est obtenu en multipliant la fréquence de la sinusoïde, c'est-à-dire $s'(t) = \sin(\alpha \omega t)$ avec le même support temporel que l'original.

Il semble donc intéressant de pouvoir décomposer le signal original en une somme de sinusoïdes afin de pouvoir effectuer le traitement adéquat sur chacune de ces composantes.

Cependant, l'utilisation de la transformée de Fourier n'est pas souhaitable sur la globalité

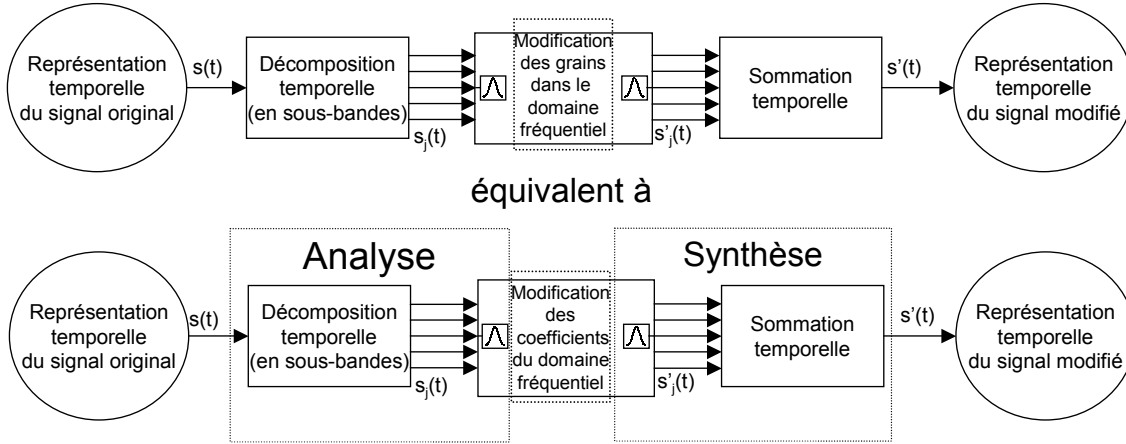


Figure 2.31 – Illustration de la transformation des grains temporels

du signal, car elle ne prendrait pas en compte l'évolution temporelle qui se produit pour des signaux réels. Il est donc nécessaire d'utiliser des transformations adaptées à ces signaux comme les transformations temps-fréquence [Coh95].

Nous nous intéressons ici à un type de transformation particulier, largement utilisé dans le cadre du traitement des sons, que l'on appelle la Transformée de Fourier à Court Terme (TFCT) [Gui02].

Il s'agit de rendre compte des évolutions des caractéristiques spectrales au court du temps en effectuant des transformées de Fourier non pas sur toute la durée du signal, mais sur un ensemble de "tranches" temporelles. En se plaçant dans le cadre d'un signal continu $s(t)$, on obtient alors l'opération suivante :

$$S_n(\Omega) = \int_{t_n}^{t_{n+1}} s(t) e^{-j\Omega t} dt$$

Cependant, d'une part le signal est coupé brutalement, ce qui entraîne un élargissement du spectre, et d'autre part la variable n est discrète donc la covariance en temps n'est pas assurée (la translation du signal de t_0 quelconque n'entraîne pas une translation équivalente de S_n puisque n est discret).

Pour remédier à ces deux problèmes, d'une part on applique au signal une fenêtre d'analyse h régulière et à support compact, généralement centrée à l'instant τ , et d'autre part on déplace cette fenêtre de façon continue. De plus, afin de satisfaire la relation de covariance, nous translatons également en temps l'exponentielle complexe, ce qui nous mène à l'opération suivante :

$$S(\tau, \Omega) = \int_{-\infty}^{+\infty} s(t) \bar{h}(t - \tau) e^{-j\Omega(t - \tau)} dt \quad (2.23)$$

On peut noter la présence du conjugué de la fenêtre d'analyse \bar{h} . Ce conjugué est optionnel et ne change rien en pratique car les fenêtres utilisées sont généralement réelles ($\bar{h} = h$, ce que nous supposons à partir de maintenant), mais cela nous permet d'avoir une notation correcte pour interpréter la TFCT en tant que produit scalaire.

Représentation de Fourier à Court Terme discrète

La TFCT telle que nous l'avons définie permet de simplifier les calculs formels et offre une représentation conjointe en temps et en fréquence. Cependant, en pratique, nous avons affaire

à des signaux temporels discrets $s(n)$, et nous utilisons des Transformées de Fourier Discrètes (TFD) où la variable fréquentielle est elle aussi discrète. Il est donc nécessaire d'échantillonner la TFCT à la fois en temps et en fréquence.

Cauchy [Cau41], Nyquist [Cau41] et Shannon [Sha48] ont montré que l'échantillonnage de l'axe temporel, qui entraîne une périodisation de l'axe fréquentiel, est réalisé correctement (pas de perte d'information) lorsque la fréquence d'échantillonnage F_e utilisée est supérieure à deux fois la plus haute fréquence F_{max} présente dans le signal continu. La condition pour éviter le recouvrement fréquentiel s'écrit donc :

$$F_e \geq 2F_{max}$$

En ce qui concerne l'échantillonnage de l'axe fréquentiel, qui entraîne une périodisation de l'axe temporel, on peut montrer que le recouvrement temporel est évité lorsque la longueur H de la fenêtre h est inférieure ou égale à la longueur N de la TFD [Dol86]. La condition pour éviter le recouvrement temporel s'écrit donc :

$$N \geq H$$

L'échantillonnage minimal sans perte d'information étant atteint pour $N = H$, nous nous placerons toujours dans ce cas. Les valeurs discrètes des fréquences réduites sont indexées par k et sont données par :

$$\Omega_k = 2\pi \frac{k}{N} \quad k \in [0, N-1] \quad (2.24)$$

Pour ce qui concerne ces questions d'échantillonnage, on pourra consulter par exemple [OS75].

La discrétisation de la TFCT précédente, alors nommée TFCT discrète, est donc définie par :

$$S(p, \Omega_k) = \sum_{m=-\infty}^{+\infty} s(m)h(m-p)e^{-j\Omega_k(m-p)} \quad (2.25)$$

qui s'écrit encore (avec le changement de variable $i = m - p$) :

$$S(p, \Omega_k) = \sum_{i=-\infty}^{+\infty} s(i+p)h(i)e^{-j\Omega_k i} \quad (2.26)$$

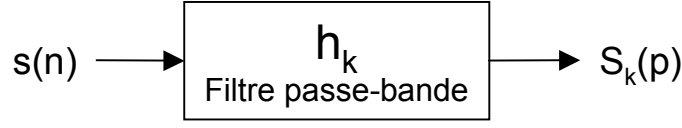
Cette définition de la TFCT discrète est conforme aux définitions données entre autres dans [VR93, ML95, Lar98, Ver00]. Elle est parfois appelée "**convention passe-bande**" ou "référence temporelle glissante", ce qui traduit fidèlement la manière dont elle est implantée [CR83, Cap93].

De nombreux auteurs ([SR73, Por76, All77, AR77, Mal79, Cro80, GL84, Dol86, AKZ02]) préfèrent utiliser la définition connue sous le terme "**convention passe-bas**" (ou "référence temporelle fixe"), dont l'écriture est dans certains cas moins lourde que la précédente. Elle est donnée par l'équation suivante :

$$S_{LP}(p, \Omega_k) = \sum_{m=-\infty}^{+\infty} s(m)h(p-m)e^{-j\Omega_k m}$$

Si la fenêtre h est symétrique ($h(t) = h(-t)$), ce que nous supposons à partir de maintenant, la relation entre ces deux définitions s'écrit :

$$S_{LP}(p, \Omega_k) = S(p, \Omega_k)e^{-j\Omega_k p}$$

Figure 2.32 – *Interprétation en banc de filtres passe-bande*

L'équation 2.25 peut être interprétée de plusieurs façons différentes, selon la manière d'agencer les termes de la somme [Cap93, Dep99].

– *Interprétation en banc de filtres passe-bande*

Dans l'interprétation en banc de filtres, on considère l'évolution temporelle de l'amplitude en sortie de chaque filtre. En d'autres termes, on fixe la variable k et on observe le signal complexe en fonction de p .

La TFCT de l'équation 2.25 peut s'écrire de la manière suivante :

$$\begin{aligned}
 S_k(p) &= \sum_{m=-\infty}^{+\infty} s(m)[h(m-p)e^{-j\Omega_k(m-p)}] \\
 &= \sum_{m=-\infty}^{+\infty} s(m)h_k(m-p) \\
 &= (s * h_k)(p)
 \end{aligned}$$

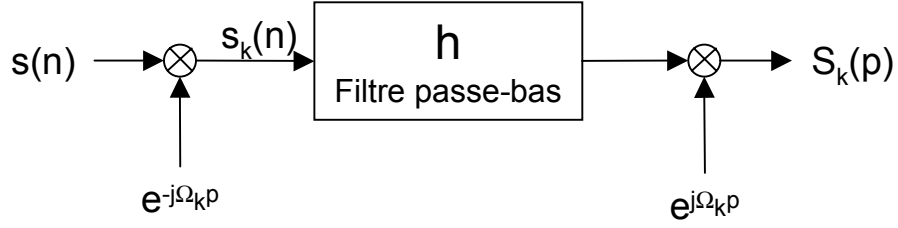
avec $h_k(p)$ la fenêtre modulée à la fréquence Ω_k , et $*$ le produit de convolution. On peut donc interpréter la TFCT comme la convolution du signal original par h_k , c'est-à-dire le filtrage du signal original par une série de filtres passe-bande, dont les fréquences sont centrées en Ω_k . La réponse impulsionnelle h_k de chacun de ces filtres est obtenue en modulant la fenêtre h par une exponentielle complexe de fréquence Ω_k . Le terme de convention passe-bande prend donc ici tout son sens.

Le schéma 2.32 représente le diagramme de cette interprétation.

– *Interprétation en banc de filtres passe-bas*

Dans cette deuxième interprétation en banc de filtre, on considère une modulation du signal original, un filtrage passe-bas, puis une démodulation comme l'indiquent le schéma 2.33 et les équations suivantes :

$$\begin{aligned}
 S_k(p) &= \sum_{m=-\infty}^{+\infty} [s(m)e^{-j\Omega_k(m-p)}]h(m-p) \\
 &= e^{j\Omega_k p} \sum_{m=-\infty}^{+\infty} [s(m)e^{-j\Omega_k m}]h(m-p) \\
 &= e^{j\Omega_k p} \sum_{m=-\infty}^{+\infty} s_k(m)h(m-p)
 \end{aligned}$$

Figure 2.33 – *Interprétation en banc de filtres passe-bas*

$$= e^{j\Omega_k p} (s_k * h)(p)$$

Le signal $s_k(t)$ correspond à la modulation du signal original de sorte que la fréquence Ω_k se retrouve centrée en 0, et h joue le rôle de la réponse impulsionnelle d'un filtre passe-bas dont la longueur, c'est-à-dire le nombre de points de sa réponse impulsionnelle, est égale à N .

Nous pouvons remarquer que si l'on opte pour la convention passe-bas, le terme de démodulation disparaît, laissant place uniquement à un filtrage passe-bas du signal modulé, d'où le nom de la convention.

– *Interprétation en Transformée de Fourier*

Dans l'interprétation en Transformée de Fourier, on considère le spectre local du signal à un instant donné. En d'autres termes, on fixe la variable p et on observe les fréquences Ω_k . L'image temps-fréquence est alors vue comme une succession de spectres locaux, comme nous l'indique l'équation suivante :

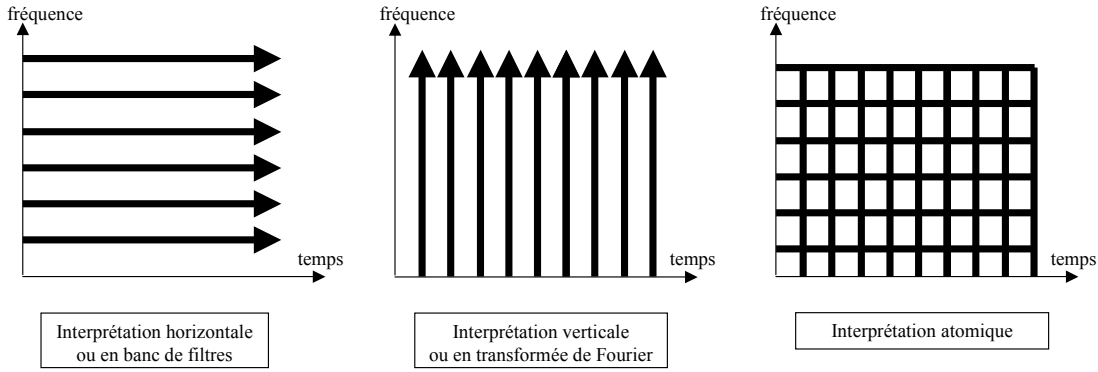
$$\begin{aligned} S_p(\Omega_k) &= \sum_{m=-\infty}^{+\infty} [s(m)h(m-p)]e^{-j\Omega_k(m-p)} \\ &= e^{j\Omega_k p} \sum_{m=-\infty}^{+\infty} s_p(m)e^{-j\Omega_k m} \\ &= e^{j\Omega_k p} F[s_p](\Omega_k) \end{aligned}$$

avec F l'opérateur de Fourier et $s_p(n)$ le signal à court terme original fenêtré autour de l'instant p . Il s'agit donc de fenêtrer le signal original grâce à une fenêtre glissante et d'en effectuer la transformée de Fourier avant de moduler le signal par une exponentielle complexe. Cela revient à déplacer l'origine temporelle du signal, d'où l'appellation "référence temporelle glissante".

On remarque que dans la convention passe-bas, le terme de modulation disparaît. La TFCT discrète est alors vue comme la succession des spectres d'un signal fenêtré dont l'origine est fixe, d'où son appellation "référence temporelle fixe".

– *Interprétation atomique*

Dans cette interprétation, on fixe à la fois les variables p et k . Il n'y a plus d'axe privilégié.

Figure 2.34 – *Interprétations de la Représentation de Fourier à Court Terme*

Puisque nous supposons que h est réel ($\bar{h}(n) = h(n)$), on a :

$$\begin{aligned}
 S_{p,k} &= \sum_{m=-\infty}^{+\infty} s(m)[h(m-p)e^{-j\Omega_k(m-p)}] \\
 &= \sum_{m=-\infty}^{+\infty} s(m)\bar{h}_{p,k}(m) \\
 &= \langle s, h_{p,k} \rangle
 \end{aligned}$$

avec $h_{p,k}(m) = h(m-p)e^{j\Omega_k(m-p)}$ l'atome d'analyse.

On peut donc représenter la TFCT discrète comme une décomposition (rôle du produit scalaire) du signal sur une famille d'atomes $h_{p,k}$ qui sont des sinusoides fenêtrées.

Dans le cas où la fenêtre h est une gaussienne, la TFCT est appelée "Transformée de Gabor".

Toutes ces interprétations sont équivalentes entre elles, mais peuvent donner lieu à des implantations différentes, certaines plus efficaces que d'autres d'un point de vue puissance de calcul.

La figure 2.34 schématise les différentes interprétations de la RFCT.

Analyse

L'interprétation en banc de filtres mène à une implantation assez intuitive, mais très lourde en calculs, surtout lorsque N est élevé. Il s'agit en effet d'effectuer en parallèle N filtrages fournissant l'amplitude et la phase des sinusoides de fréquences Ω_k pour $k \in [0, N-1]$.

L'interprétation en transformée de Fourier est importante car elle permet une implantation efficace grâce à la FFT. En effet, en prenant une fenêtre centrée h de support fini $H = N$ ($h(n) = 0$ pour $n < -N/2$ et $n \geq N/2$), l'équation 2.25 se simplifie de la sorte :

$$S_p(\Omega_k) = e^{j\Omega_k p} \sum_{m=p-N/2}^{p+N/2-1} s(m)h(m-p)e^{-j\Omega_k m}$$

Ce qui donne, avec le changement de variable adéquat ($i = m - p + N/2$) :

$$\begin{aligned}
 S_p(\Omega_k) &= e^{j\Omega_k p} \sum_{i=0}^{N-1} s(i - N/2 + p)h(i - N/2)e^{-j\Omega_k(i+p-N/2)} \\
 &= e^{j\Omega_k \frac{N}{2}} \sum_{i=0}^{N-1} s(i - N/2 + p)h(i - N/2)e^{-j\Omega_k i} \\
 &= e^{j\pi k} \sum_{i=0}^{N-1} s(i - N/2 + p)h(i - N/2)e^{-j\Omega_k i} \\
 &= (-1)^k \sum_{i=0}^{N-1} s(i - N/2 + p)h(i - N/2)e^{-j\Omega_k i} \tag{2.27}
 \end{aligned}$$

Cette formulation est totalement adaptée à une implantation sous forme de FFT puisque l'on reconnaît la DFT sous forme d'une sommation de 0 à $N - 1$. Le terme $e^{j\pi k} = (-1)^k$ peut être réalisé en pratique simplement en effectuant une rotation circulaire de $N/2$ au signal fenêtré (cela revient à échanger les moitiés droite et gauche de ce signal) avant d'en faire la FFT [Cro80].

Dans la convention passe-bande, le terme précédant la somme est $e^{j\pi k}$. Dans cette convention, lorsqu'une sinusoïde de fréquence Ω_k est présente dans le k^{ieme} canal, la phase dans ce canal est identique à la phase du signal.

Dans la convention passe-bas, le terme précédant la somme est $e^{-j\Omega_k(p-N/2)}$. Dans cette convention, lorsqu'une sinusoïde de fréquence Ω_k est présente dans le k^{ieme} canal, la phase dans ce canal est constante au cours du temps puisqu'il s'agit de la différence de phase entre le signal analysé et la sinusoïde analysante.

C'est ce que nous allons montrer à travers l'exemple suivant.

Exemple d'analyse pour une seule sinusoïde

Soit un signal constitué d'une exponentielle complexe $s(n) = e^{j\psi(n)}$ dont la phase est linéaire ($\psi(n) = \omega n + \Psi$, la fréquence ω est constante), on calcule la TFCT discrète à partir de l'équation 2.26 :

$$\begin{aligned}
 S(p, \Omega_k) &= \sum_{i=-\infty}^{+\infty} e^{j\psi(i+p)}h(i)e^{-j\Omega_k i} \\
 &= e^{j\psi(p)} \sum_{i=-\infty}^{+\infty} h(i)e^{-j(\Omega_k - \omega)i} \\
 &= e^{j\psi(p)} \hat{h}(\Omega_k - \omega) \tag{2.28}
 \end{aligned}$$

avec \hat{h} la transformée de Fourier de h .

Si h est symétrique, \hat{h} est réel. Donc, la phase de la TFCT correspond à la phase du signal pour toutes les fréquences où la transformée de Fourier de la fenêtre est positive (lorsqu'elle est négative, la phase de la TFCT correspond à l'opposé de la phase du signal).

Pour un signal réel $s(n) = \cos(\psi(n))$, la TFCT s'écrit :

$$S(p, \Omega_k) = e^{j\psi(p)} \hat{h}(\Omega_k - \omega_0) + e^{-j\psi(p)} \hat{h}(\Omega_k + \omega_0)$$

Ici, la phase de la TFCT correspond à la phase du signal pour les fréquences positives, et à l'opposé de la phase du signal pour les fréquences négatives.

Nous montrons par les 3 exemples suivants comment la TFCT peut être représentée et interprétée grâce aux informations de module et de phase.

La figure 2.35 représente le signal temporel d'une sinusoïde de fréquence $2\pi 5/N$ (centrée sur le canal 5) ainsi que son spectrogramme⁷ (module de la TFCT). La fenêtre utilisée est une gaussienne de longueur $N = 64$ donnée par :

$$h(t) = e^{-\frac{t^2}{2\sigma^2}}$$

échantillonnée et tronquée en $N/2$ à 10^{-8} (d'où $\sigma^2 = -\frac{N^2}{8\ln(10^{-8})}$). On remarque un étalement de l'énergie dans au moins 6 canaux adjacents de part et d'autre du maximum, traduisant la forme en fréquence de la fenêtre d'analyse représentée à droite.

Cette figure montre également 3 phasogrammes (phase de la TFCT) : le phasogramme de la DFT (obtenu simplement par l'application de la FFT), le phasogramme de la TFCT dans la convention passe-bande (phasogramme de la DFT avec le terme correctif $(-1)^k$), et le phasogramme de la TFCT dans la convention passe-bas (phasogramme de la DFT avec le terme correctif $e^{-j\Omega_k(p-N/2)}$).

Le phasogramme de la TFCT dans la convention passe-bande montre que pour la fenêtre temporelle gaussienne, dont la partie réelle de la transformée de Fourier est positive, les phases de tous les canaux, dans la partie stationnaire du signal, sont identiques à la phase du canal centré sur la sinusoïde (le canal 5 dans notre cas), qui est elle-même identique à la phase du signal original.

Pour le phasogramme de la TFCT dans la convention passe-bas, on peut remarquer la valeur constante de la phase dans le canal 5 signifiant que la sinusoïde "analysée" (signal original) possède la même phase que la sinusoïde "analysante" (fréquence centrale du filtre) dans ce canal. Par contre, l'interprétation des phases dans les canaux adjacents est moins immédiate puisqu'il s'agit également de la différence de phase entre la sinusoïde "analysée" et la sinusoïde "analysante" (dont la fréquence est proportionnelle au numéro de canal), relation qui n'est pas très intuitive.

Il faut noter que dans les canaux où l'énergie est négligeable, les valeurs de la phase sont représentées mais elles ne sont généralement pas significatives.

La figure 2.36 représente les mêmes données avec une fenêtre temporelle de Hanning (arche de cosinus) tronquée et symétrique. On peut voir un maximum d'énergie dans le canal 5, ainsi qu'un étalement de l'énergie dans les canaux adjacents 4 et 6 traduisant la forme en fréquence de la fenêtre d'analyse. La transformée de Fourier d'une telle fenêtre est positive dans les 3 canaux du lobe principal, puis alterne entre valeurs positives et négatives. Il s'ensuit que les phases des canaux correspondant aux valeurs positives sont identiques à la phase du canal centré sur la sinusoïde, alors que les phases des canaux correspondant aux valeurs négatives sont décalés d'une valeur π .

La figure 2.37 représente aussi les mêmes données avec une fenêtre de Hanning tronquée, mais le signal sinusoïdal analysé possède une fréquence située entre les canaux 5 et 6. L'énergie s'étale maintenant sur 4 canaux, pour lesquels les phases sont identiques dans la convention passe-bande.

La forme temporelle de la fenêtre d'analyse h est choisie en réalisant un compromis entre la largeur du lobe fréquentiel principal et l'amplitude des rebonds des lobes fréquentiels

7. Toutes les figures représentent un détail des 16 premiers canaux fréquentiels.

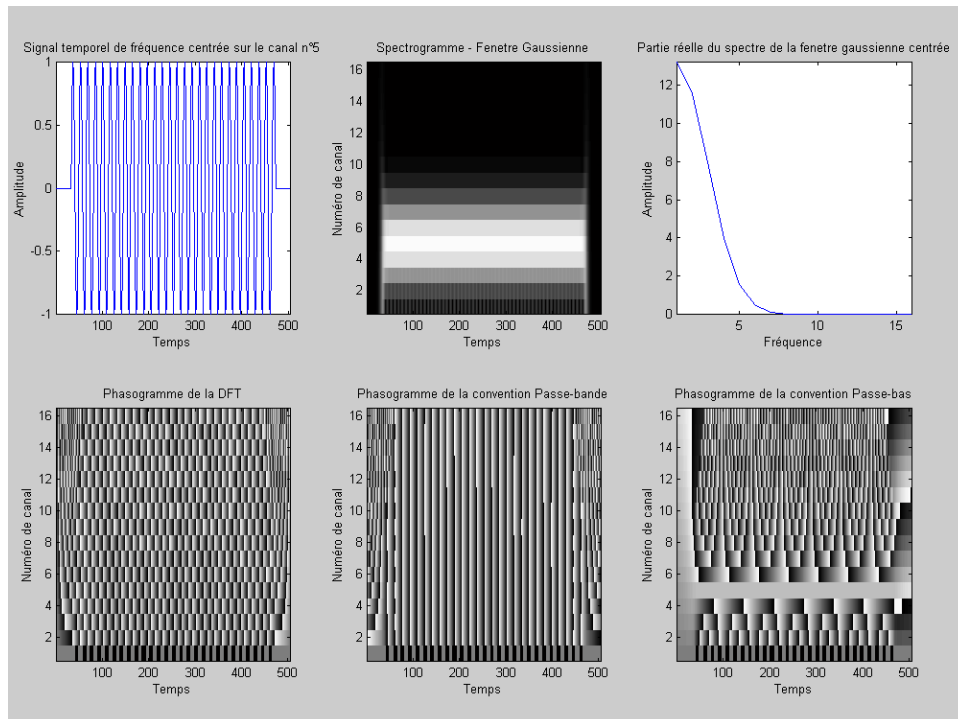


Figure 2.35 – Représentations d'une sinusoïde centrée sur un canal (fenêtre gaussienne)

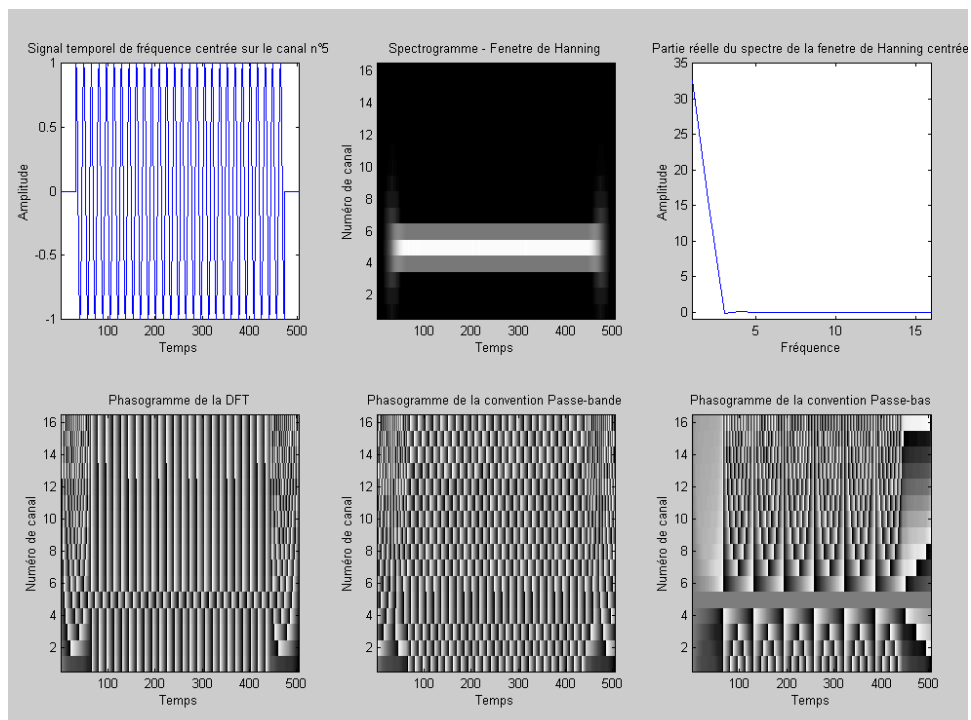


Figure 2.36 – Représentations d'une sinusoïde centrée sur un canal (fenêtre de Hanning)

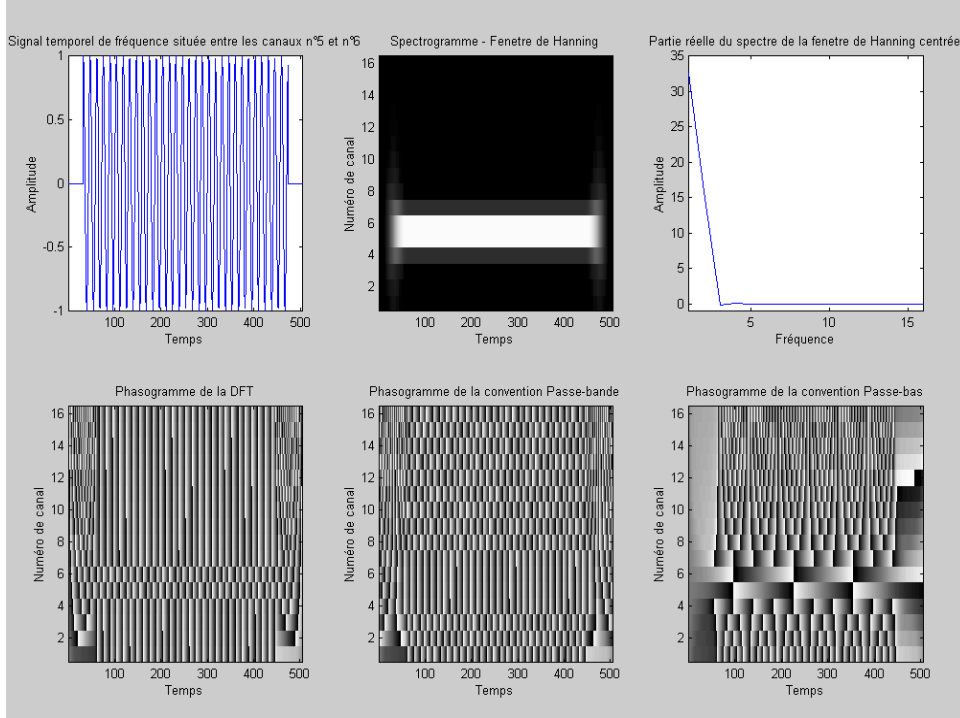


Figure 2.37 – Représentations d’une sinusoïde à cheval entre deux canaux (fenêtre de Hanning)

secondaires. Pour des applications telles que la dilatation-p, une fenêtre souvent utilisée est celle de Hanning ou plutôt Hanning modifiée [GL84], mais il est possible d’utiliser également Hamming, Kaiser, Blackman-Harris...

La TFCT offrant une représentation très redondante, il est possible d’effectuer un sous-échantillonnage de ses données temporelles. Dans l’interprétation en banc de filtre, on explique ce sous-échantillonnage par le fait que les signaux issus de chaque sous-bande sont à bande limitée : dans la convention passe-bas, il est donc possible d’effectuer une réduction de fréquence d’échantillonnage car toutes les sous-bandes sont limitées en fréquence par le filtre passe-bas. Dans l’interprétation en Transformée de Fourier, cela revient à placer la fenêtre d’analyse à des instants temporels discrets $L_i = iL$ où L représente le pas d’analyse.

La TFCT discrète issue de l’équation 2.27 devient alors :

$$S(L_i, \Omega_k) = e^{j\pi k} \sum_{m=0}^{N-1} s(m - N/2 + L_i) h(m - N/2) e^{-j\Omega_k m}$$

Transformation

Les transformations des méthodes fréquentielles sont basées sur la modification de la TFCT discrète.

La transformation de la représentation pour la **dilatation-p** est donnée par :

$$S'_{Dp}(\tau, \Omega) = S\left(\frac{\tau}{\alpha}, \Omega\right) e^{-j\Omega \frac{\tau}{\alpha}} e^{j\Omega \tau}$$

qui peut s’écrire à l’aide du module M et de la phase φ de la TFCT :

$$S'_{Dp}(\tau, \Omega) = \left[M\left(\frac{\tau}{\alpha}, \Omega\right) e^{j\varphi\left(\frac{\tau}{\alpha}, \Omega\right)} \right] e^{-j\Omega \frac{\tau}{\alpha}} e^{j\Omega \tau}$$

Dans l'interprétation en banc de filtres, on peut voir cette transformation comme une modification de l'évolution temporelle (à travers le module M) de la sinusoïde de fréquence ω avec une compensation pour le terme de phase.

Dans l'interprétation en transformée de Fourier, on peut voir cette transformation comme une interpolation temporelle des spectres d'amplitude successifs avec modification adéquate des phases.

Si l'on suppose que le signal $s(t)$ est constitué d'une sinusoïde de fréquence ω , alors la phase de la transformée à cette fréquence correspond à la phase du signal : $\varphi(\frac{\tau}{\alpha}, \omega) = \omega \frac{\tau}{\alpha}$. La représentation modifiée pour la fréquence ω est alors donnée par :

$$S'_{Dp}(\tau, \omega) = M\left(\frac{\tau}{\alpha}, \omega\right) e^{j\omega\tau}$$

On obtient donc bien une sinusoïde de fréquence inchangée ω mais dont le support temporel est modifié.

La transformation de la représentation pour la **transposition-p** est donnée par :

$$S'_{Tp}(\tau, \Omega) = S\left(\tau, \frac{\Omega}{\alpha}\right) e^{-j\frac{\Omega}{\alpha}\tau} e^{j\Omega\tau}$$

qui peut s'écrire à l'aide du module M et de la phase φ de la TFCT :

$$S'_{Tp}(\tau, \Omega) = \left[M\left(\tau, \frac{\Omega}{\alpha}\right) e^{j\varphi(\tau, \frac{\Omega}{\alpha})} \right] e^{-j\frac{\Omega}{\alpha}\tau} e^{j\Omega\tau}$$

Dans l'interprétation en banc de filtres, on peut voir cette transformation comme l'application de l'évolution temporelle (à travers le module M) de la sinusoïde de fréquence $\frac{\Omega}{\alpha}$ à la sinusoïde de fréquence Ω .

Dans l'interprétation en transformée de Fourier, on peut voir cette transformation comme une dilatation locale du spectre d'amplitude avec modification adéquate des phases.

Si l'on suppose que le signal $s(t)$ est constitué d'une sinusoïde de fréquence ω , alors la phase de la transformée à cette fréquence correspond à la phase du signal : $\varphi(\tau, \frac{\omega}{\alpha}) = \frac{\omega}{\alpha}\tau$. La représentation modifiée pour la fréquence ω est alors donnée par :

$$S'_{Tp}(\tau, \omega) = M\left(\tau, \frac{\omega}{\alpha}\right) e^{j\omega\tau}$$

Il faut noter que la modification d'une image obtenue par TFCT (représentation du signal dans les deux dimensions conjointes temps et fréquence) ne mène généralement pas à une image "valide", c'est-à-dire que cette nouvelle image bidimensionnelle ne correspond à la transformée d'aucun signal temporel. Attention, cela ne signifie pas qu'un signal temporel ne peut pas être obtenu à partir de cette image modifiée, mais plutôt que la TFCT du signal obtenu ne correspond pas à l'image modifiée. Cette propriété de la TFCT, et plus généralement des représentations temps-fréquence, s'explique par la présence d'un noyau reproduisant [GKMM89], fonction pour laquelle la valeur d'un point donné d'une image dépend de la valeur

des autres points de son voisinage. En d'autres termes, on ne peut agir de manière cohérente sur une TFCT sans prendre en compte le voisinage de la zone dans laquelle a lieu la modification.

Nous étudions dans la suite en détail les différentes modifications possibles selon les méthodes et les implantations ainsi que diverses améliorations.

Synthèse

Les différentes interprétations des méthodes de synthèse mènent, comme pour l'analyse, à des implantations différentes.

Dans l'interprétation en banc de filtres, le signal dilaté $s'(n)$ est obtenu par sommation des différentes sous-bandes modifiées $s'_k(n)$:

$$s'(n) = \sum_{k=0}^{N-1} S'_k(n)$$

On peut montrer que pour $s(n)$ réel, les signaux $S'_k(n)$ ($k \in [0, N/2]$ suffit alors) sont à valeurs réelles et peuvent être interprétés comme des fonctions cosinus de fréquence Ω_k variant lentement en amplitude et en phase [AKZ02].

On appelle cette méthode la synthèse en banc de filtres, ou encore synthèse en banc d'oscillateurs. Cette synthèse additive peut être implantée de manière efficace grâce à l'utilisation de la FFT inverse [RD92].

Si la TFCT a été sous-échantillonnée, il est nécessaire de sur-échantillonner chaque sous-bande avant de procéder à la sommation.

Dans l'interprétation en transformée de Fourier, le signal dilaté $s'(n)$ est obtenu par addition-recouvrement des grains temporels issus de la transformée inverse. Les pas de synthèse E étant différents des pas d'analyse L , le signal est dilaté.

Ce chevauchement des grains peut être réalisé par une procédure nommée OLA ("OverLap-Add" ou recouvrement simple) donnée par l'équation suivante :

$$s'(n) = \frac{\sum_{i=-\infty}^{\infty} s'_{iE}(n)}{\sum_{i=-\infty}^{\infty} h(iE - n)}$$

avec $s'_{iE}(n)$ les grains temporels centrés en iE et obtenus grâce à la Transformée de Fourier Discrète Inverse (TFDI) de $S'_{iE}(\Omega_k)$ définie par :

$$s'_{iE}(n) = \frac{1}{N} \sum_{k=0}^{N-1} S'_{iE}(\Omega_k) e^{j\Omega_k n}$$

Le chevauchement des grains peut aussi être réalisé par une procédure plus complexe nommée WOLA ("Weighted OverLap-Add" ou recouvrement pondérée), consistant à pondérer par une fenêtre de synthèse v le grain temporel modifié :

$$s'(n) = \frac{\sum_{i=-\infty}^{\infty} v(iE - n)s'_{iE}(n)}{\sum_{i=-\infty}^{\infty} h(iL - n)v(iE - n)} \quad (2.29)$$

Cette procédure possède un avantage de qualité certain lorsque des transformations sont effectuées sur la TFCT discrète. En effet, une modification de la phase d'une transformée de Fourier implique généralement une modification de l'amplitude du signal temporel correspondant, menant ainsi à la présence d'énergie sur les bords du signal temporel à court terme. Il est donc nécessaire d'atténuer cette énergie, absente du signal à court terme original du fait du fenêtrage d'analyse, par l'application d'une fenêtre de synthèse adéquate.

Fenêtres d'analyse et de synthèse peuvent être a priori quelconques du moment que les modulations d'amplitude induites par ces fenêtrages peuvent être compensées par une fonction de normalisation après la synthèse. Cependant, il est intéressant d'utiliser une unique fenêtre lors de l'analyse et de la synthèse, qui prenne en compte la fonction de normalisation ce qui évite le calcul de la compensation de modulation après la synthèse.

Griffin et Lim [GL84] montrent ainsi que l'utilisation d'une fenêtre de Hanning ou Hamming modifiée (périodicité de N au lieu de $N - 1$) supprime la nécessité de normaliser (le terme du dénominateur de l'équation 2.29 disparaît).

Le Vocodeur de Phase

Lorsque les valeurs complexes issues de la TFCT sont exprimées sous forme polaire (module et phase), il est coutume d'appeler ce système "vocodeur de phase".

Le terme anglais "vocoder" est dérivé de "VOIce CODER", un système destiné à réduire la largeur de bande nécessaire à la transmission de la parole sur les lignes téléphoniques [Dud39]. Il consiste à coder l'évolution énergétique de chacune des sous-bandes. En pratique, son utilisation pour cette application se révèle impossible car l'information à transmettre est supérieure à l'information du signal original [GS85]. D'autre part, la qualité est plutôt médiocre car l'information de phase est absente.

Pour corriger ce défaut de qualité, Flanagan et Golden [FG66] introduisent le vocodeur de phase, capable de reconstituer le signal original exact. Depuis, de nombreux auteurs [SR73, Por76, Cro80] ont amélioré cette technique notamment en termes de rapidité de calculs.

Dans ce système, la TFCT continue est mise sous la forme polaire suivante :

$$S(\tau, \Omega) = M(\tau, \Omega)e^{j\varphi(\tau, \Omega)}$$

avec M et φ les modules et phases à court terme de la TFCT.

La fréquence instantanée f , qui est dérivée de la phase, est beaucoup plus intuitive à manipuler que les valeurs de phase. De plus, la phase n'est généralement pas bornée, alors que sa dérivée l'est. On a donc :

$$f(\tau, \Omega) = \frac{1}{2\pi} \frac{\partial \varphi(\tau, \Omega)}{\partial \tau}$$

En pratique, la phase n'est obtenue qu'à 2π près. Il est donc nécessaire d'appliquer un algorithme de "déroulement de phase" (on supprime la discontinuité de phase au cours du temps) avant d'en extraire la fréquence instantanée. D'autre part, pour la TFCT discrète, il

est nécessaire que les instants d'analyse L_i ne soient pas trop éloignés afin qu'il n'y ait aucune indétermination sur la valeur de la phase déroulée [AKZ02].

2.3.2 Méthodes "aveugles" de dilatation-p

Dans ce type de méthode, les pas d'analyse et les pas de synthèse sont constants : $L_i = L$ et $E_i = E$. De plus, les fenêtres d'analyse sont toutes identiques : $h_i = h$. Seules les informations de "bas niveau" que sont le module et la phase de la TFCT discrète sont utilisées pour la transformation. Aucune interprétation de "haut niveau" de ces données n'est réalisée, c'est-à-dire que les valeurs de la TFCT discrètes sont traitées de manière identique pour toutes les fréquences, et sans prendre en considération l'aspect stationnaire ou transitoire du signal.

Principe des méthodes aveugles

Le but de ce type de méthode est d'éviter les discontinuités de phase en manipulant la phase de chacune des composantes.

La dilatation-p consiste à faire évoluer les spectres originaux selon une échelle temporelle dilatée. Pour la TFCT, cela revient donc à diviser la variable temporelle sans modifier la variable fréquentielle : $S'(\tau, \Omega) = S(\tau/\alpha, \Omega)$. Cependant, cette dilatation de l'axe temporel entraîne une incohérence des phases puisque la fréquence instantanée en Ω correspond maintenant à la fréquence instantanée Ω/α . Il est donc nécessaire d'appliquer une compensation de phase adéquate.

Nous distinguons les méthodes qui utilisent des pas de synthèse proportionnels aux pas d'analyse $E = \alpha L$, qui mènent à extrapoler la phase déroulée, des méthodes qui utilisent des pas identiques $E = L$, qui mènent parfois à utiliser plusieurs fois le même spectre d'amplitude.

Méthode par vocodeur de phase classique ($E = \alpha L$)

Portnoff [Por76, Por80, Por81a, Por81b] développe les bases théoriques du vocodeur de phase dans le contexte de la parole, et utilise ces résultats dans une application de dilatation-p. De nombreux auteurs [Cro80, Dol86, AKZ02] ont suivi cette trace, apportant une compréhension accrue de ce système ainsi que des améliorations liées à l'efficacité en termes de calcul.

Le principe mis en oeuvre ici consiste à échantillonner différemment la TFCT discrète entre l'analyse et la synthèse. Les instants d'analyse sont répartis régulièrement (pas L constant), et les instants de synthèse (pas E constant). On obtient alors une évolution temporelle modifiée des spectres d'analyse. Les amplitudes des spectres ne sont pas modifiées afin de conserver les caractéristiques fréquentielles, mais les phases de chacune des composantes sont compensées afin de ne créer aucune discontinuité lors du recouvrement des grains dans l'interprétation en transformée de Fourier, ou pour que l'évolution temporelle des phases corresponde bien à la fréquence analysée dans l'interprétation en banc de filtres.

Une fois la fréquence instantanée calculée aux instants L_i , on construit la TFCT discrète aux instant E_i :

$$S(E_i, \Omega_k) = M(L_i, \Omega_k) e^{j2\pi f(L_i, \Omega_k) E_i}$$

Il suffit ensuite de construire le signal temporel ainsi dilaté par une des méthodes de synthèse étudiées précédemment.

Cette technique revient donc à effectuer une décomposition en sous-bandes (avec M_k et ϕ_k le module et la phase de la k^{ieme} sous-bande) constituées de sinusoides lentement modulées en amplitude et en phase. Le signal dilaté est obtenu en rééchantillonnant les signaux de module

et de phase de chaque sous-bande :

$$s'(t) = \sum_{k=0}^{N-1} M_k(t/\alpha) \cos(\alpha \phi_k(t/\alpha))$$

Le rééchantillonnage de la phase ϕ_k entraînant une modification de la fréquence, il est nécessaire de multiplier cette fonction par α afin de conserver la fréquence originale.

Méthode par interpolation des spectres d'amplitude ($E = L$)

La démarche présentée ici est utilisée par Ellis [Ell91] et Bonada [Bon00a]. Elle est similaire à la démarche employée pour la méthode PSOLA C. Dans cette méthode, les pas d'analyse et de synthèse sont identiques ($E = L$). Les marques d'écriture E_i ne correspondent donc plus aux marques de lecture L_i par la fonction de dilatation. On déduit de E_i une marque de correspondance Lc_i par inversion de la fonction de dilatation :

$$Lc_i = \frac{E_i}{\alpha}$$

Cette marque indique dans l'échelle temporelle originale l'instant auquel aurait dû avoir lieu l'événement synthétisé à l'instant E_i . Comme nous le verrons dans la suite, on remarque que le taux de dilatation α n'est utilisé que pour le calcul de la marque de correspondance, et nulle part ailleurs.

Spectres d'amplitude

Etant donné que le spectre d'amplitude original n'est généralement pas calculé aux instants Lc_i , il est nécessaire d'en donner une valeur estimée par interpolation des spectres d'amplitude calculés aux instants L_j et L_{j+1} tels que $L_j \leq Lc_i \leq L_{j+1}$.

Cette interpolation peut être aussi simple que de prendre le spectre d'amplitude correspondant à l'indice L_j le plus proche de Lc_i auquel cas un même spectre peut être utilisé plusieurs fois [Bon00a]. Une autre solution un peu plus complexe consiste à effectuer une interpolation linéaire entre les spectres d'amplitude correspondants aux indices L_j et L_{j+1} entourant l'indice Lc_i [Ell02a] :

$$M'(E_i, \Omega_k) = \left(\frac{L_{j+1} - Lc_i}{L_{j+1} - L_j} \right) M(L_j, \Omega_k) + \left(\frac{Lc_i - L_j}{L_{j+1} - L_j} \right) M(L_{j+1}, \Omega_k)$$

Spectres de phase

Etant donné que les pas d'analyse et de synthèse sont identiques, la variation de phase entre deux marques de lecture reste identique entre les deux marques d'écriture correspondantes.

Le calcul de la phase à l'instant E_i consiste à dérouler la phase à partir de la phase d'origine spécifiée $\varphi(E_{i-1}, \Omega_k)$. Ce déroulement de phase est effectuée de manière similaire au déroulement de phase entre les instants L_j et L_{j+1} . Ce calcul nécessite comme précédemment un déroulement de phase pour extraire la fréquence instantanée, mais le taux de dilatation α n'est pas utilisé ici.

Avantages et inconvénients par rapport à la méthode classique

Un avantage se profile grâce aux pas d'analyse et de synthèse identiques : il n'y a pas de limitation portée sur α . En effet, la méthode classique doit respecter un échantillonnage minimal de la TFCT discrète afin de pouvoir estimer correctement les variations de phase, et ce pour

l'analyse ET la synthèse (typiquement un recouvrement minimal de 50% est préconisé, mais 75% est plus souvent utilisé). Or pour un facteur d'élongation (resp. de compression) très élevé, la satisfaction de cette contrainte pour la synthèse (resp. l'analyse) entraîne des calculs et un débit de données d'analyse (resp. de synthèse) excessifs.

D'autre part, un bruit blanc (son [40]) traité par un algorithme classique (doté d'une normalisation des fenêtres correcte) souffre d'une modulation d'amplitude (son [41]), atténuée fortement grâce à cette technique (son [42]). La coloration observée est discutée dans la suite.

On remarque dans ces exemples sonores une particularité : la sonie (niveau sonore subjective) du signal dilaté est inférieure à la sonie originale. Cela provient du fait que les signaux traités ont été normalisés à "pleine échelle" (la valeur absolue la plus élevée du signal correspond à $2^{15} - 1$ pour un codage sur 16 bits). Or, le traitement fournit parfois des valeurs plus élevées que la valeur maximale du signal original. L'ensemble du signal est atténué pour éviter l'écèlement, ce qui entraîne une diminution de la sonie.

Avantages et inconvénients des méthodes aveugles

Les méthodes aveugles donnent d'assez bons résultats sonores (surtout pour des taux de dilatation élevés) pour des sons simples qui répondent à l'hypothèse sous-jacente suivante :

Le signal fenêtré doit être considéré comme une somme de sinusoides d'amplitudes et de fréquences (séparées d'au moins F_e/N) fixes. C'est le cas par exemple pour les signaux vocaux (sons [36] et [39]).

Or, cette hypothèse n'est pas respectée pour au moins deux types de sons, menant à deux types d'artefacts :

Sons complexes Une coloration (ou "phasing"), ressemblant à un filtrage en peigne, avec une modulation des fréquences du filtre, se fait entendre sur des sons pour lesquels plus d'une sinusoïde est présente par canal. On en donne un exemple sonore sur un son d'orchestre (sons [88, 44]).

Un autre exemple extrême de ce type est le signal de bruit blanc (son [40]) pour lequel une forte coloration se dégage (son [41]), même avec la méthode d'interpolation des spectres (son [42]).

Cette coloration du bruit, imputée à l'inadéquation entre le signal et l'hypothèse sous-jacente (un bruit est très mal représenté par une somme de sinusoides), peut être réduite par l'augmentation de la taille des fenêtres temporelles d'analyse [Lar98], entraînant la réduction des fenêtres fréquentielles correspondantes, améliorant ainsi la résolution fréquentielle.

Pour des signaux quasi-périodiques comportant plusieurs sinusoides au sein d'un même canal d'analyse, celles-ci ne peuvent être résolues et par conséquent bien traitées, ce qui donne naissance à un "effet choral" [Moo78] (sensation de plusieurs personnes parlant en même temps). On en donne un exemple sonore sur un son d'orchestre (sons [88, 44]).

Sons impulsifs Un étalement des transitoires ou "bavement" des attaques se produit pour des sons de type impulsifs [TL00]. On en donne un exemple sonore sur un son de castagnettes (sons [11, 45]).

On peut interpréter cet étalement par la désynchronisation verticale des phases (par rapport au plan temps-fréquence, c'est-à-dire entre bandes fréquentielles à un instant donné) entraînant la délocalisation de l'énergie du transitoire.

En effet, la modification des phases de la transformée de Fourier d'un transitoire mène lors de sa reconstruction à un étalement de l'énergie sur toute la durée de la fenêtre d'analyse.

On rappelle que la différence entre une impulsion et un bruit blanc, dont les densités spectrales sont similaires, réside dans les relations qu'entretiennent les phases des différents

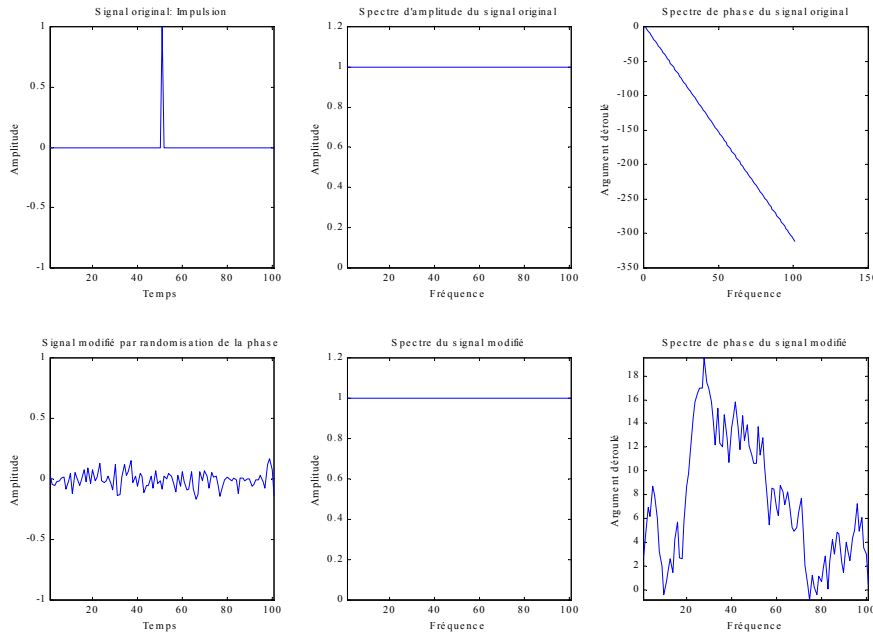


Figure 2.38 – *Illustration de la modification de phase pour un signal impulsionnel*

canaux (relations linéaires en fonction de la fréquence pour l’impulsion, et aléatoires pour la phase). La figure 2.38 illustre cette différence.

Ce phénomène s’affaiblit lorsque la taille de la fenêtre d’analyse diminue, puisque la délocalisation de l’énergie est liée à cette taille.

Un autre phénomène est à l’origine de l’étalement des transitoires : la dilatation- p est appliquée globalement sur la totalité du signal, de manière uniforme. Il est donc naturel que les transitoires fassent également l’objet de la dilatation. On observe ainsi sur la figure 2.30 l’allongement du son [k] dans le mot ”accourent”.

On peut ajouter à cette liste de défauts, caractéristiques des méthodes fréquentielles, le phénomène de réverbération [Por81b, Lar98, TL00]. Ce phénomène serait imputable à la perte de la forme de l’onde temporelle pour des signaux dont la variation lente de phase n’est pas prise en compte lors de l’analyse. De plus, puisque l’on écarte artificiellement des grains contenant des informations de réponses impulsionnelles pour $\alpha > 1$, ces dernières apparaissent plus longues, entraînant ainsi une sensation d’espace plus grand.

En conclusion, on se retrouve face à deux types d’artefacts dont les solutions sont contradictoires : pour améliorer les sons complexes, il faut augmenter la taille de la fenêtre temporelle, ce qui aggrave l’étalement des transitoires, et inversement la diminution de la taille de fenêtre améliore les transitoires mais aggrave les défauts pour les sons complexes. Il est utilisé généralement un compromis entre résolution temporelle et fréquentielle, ne satisfaisant complètement aucun de ces 2 types de sons. D’autres solutions doivent donc être proposées pour améliorer cette méthode.

2.3.3 Méthodes "aveugles" de transposition-p

Principe des méthodes aveugles

La transposition-p consiste à faire évoluer selon l'échelle temporelle originale les spectres dilatés mathématiquement. Pour la TFCT $S(\tau, \Omega)$, cela revient donc à diviser la variable fréquentielle sans modifier la variable temporelle : $S'(\tau, \Omega) = S(\tau, \Omega/\alpha)$. Cependant, cette dilatation de l'axe fréquentiel entraîne une incohérence des phases puisque la fréquence instantanée en Ω correspond à la fréquence instantanée originale en Ω/α . Il est donc nécessaire d'appliquer une compensation de phase adéquate.

Multiplication de la fréquence instantanée

Si le signal original est constitué d'une sinusoïde à la fréquence ω , alors la phase de la TFCT originale en ω est donnée par $\omega\tau$. Si l'on désire transposer cette sinusoïde à la nouvelle fréquence $\alpha\omega$, il faut remplacer l'ancien terme de phase $\omega\tau$ par le nouveau $\alpha\omega\tau$. Ceci est réalisé par une démodulation à la fréquence $\omega\tau$ puis une modulation à la nouvelle fréquence $\alpha\omega\tau$. D'où l'équation suivante :

$$S'(\tau, \alpha\omega) = S(\tau, \omega) e^{-j\omega\tau} e^{j\alpha\omega\tau}$$

En généralisant ce principe au cas où une sinusoïde est présente au sein de chaque canal d'analyse, la transformation sur la TFCT peut s'écrire de la manière suivante :

$$S'(\tau, \Omega) = S\left(\tau, \frac{\Omega}{\alpha}\right) e^{-j\frac{\Omega}{\alpha}\tau} e^{j\Omega\tau} = S\left(\tau, \frac{\Omega}{\alpha}\right) e^{-j\Omega(1-\frac{1}{\alpha})\tau}$$

Dans l'interprétation en banc de filtres, la TFCT est vue comme une série de filtres, grâce auxquels le terme de phase est directement accessible. La transformation est donc immédiate car elle consiste à multiplier par α ce terme de phase.

La sortie de chacun des filtres fournit un signal $s_k(\tau)$ donné par :

$$s_k(\tau) = M(\tau, \Omega_k) e^{j\varphi(\tau, \Omega_k)} = M_k(\tau) e^{j\varphi_k(\tau)}$$

où $M_k(\tau)$ et $\varphi_k(\tau)$ représentent le module et la phase de la TFCT à l'instant τ .

La somme des signaux filtrés reconstitue quasi-parfaitement le signal original (la différence avec l'original est inaudible). Le signal de sortie est donné par la partie réelle du signal temporel suivant :

$$s(\tau) = \sum_{k=0}^{N-1} s_k(\tau) = \sum_{k=0}^{N-1} M_k(\tau) \cos[\varphi_k(\tau)]$$

Il suffit donc, pour effectuer une transposition-p, de multiplier le terme de phase $\varphi_k(\tau)$ par α lors de la reconstruction :

$$s'(\tau) = \sum_{k=0}^{N-1} s'_k(\tau) = \sum_{k=0}^{N-1} M_k(\tau) \cos[\alpha\varphi_k(\tau)]$$

Cette technique est évoquée dans l'article de Flanagan et Golden [FG66] où la TFCT est cependant définie dans la convention passe-bas, ce qui mène à une interprétation des signaux $s_k(\tau)$ comme des sinusoïdes de fréquences fixes $\Omega_k = k/N$ lentement modulées en amplitude et en phase. Dans notre notation, la modulation de phase est contenue dans le terme $\varphi_k(\tau)$.

La reconstruction du signal temporel modifié par ce type de méthode est extrêmement coûteuse puisqu'elle nécessite N oscillateurs.

Décalage de pics dans la TFCT

Laroche et Dolson [LD99b, LD99c] proposent une méthode dans laquelle les spectres ne sont pas globalement dilatés, comme c'est le cas dans les méthodes classiques, mais plutôt décalés en fréquence. Si ce décalage est constant pour toutes les fréquences, la méthode revient à moduler le signal original par une fréquence donnée, ce qui transforme un son harmonique en un son inharmonique [Spr02]. Bien que cette transformation puisse être intéressante musicalement, elle ne nous intéresse pas dans notre contexte. Pour effectuer une transposition-p, il est donc nécessaire d'identifier chacune des composantes sinusoïdales afin de leur appliquer indépendamment le décalage en fréquence adéquat.

Le principe consiste à identifier les pics fréquentiels de la TFCT correspondant à des sinusoïdes de fréquences ω_k , et les décaler aux nouvelles fréquences ω'_k telles que :

$$\omega'_k = \omega_k + \Delta\omega_k$$

Or, le décalage du spectre complexe :

$$S'(\tau, \omega_k + \Delta\omega_k) = S(\tau, \omega_k)$$

induit un décalage à la fois du spectre d'amplitude mais aussi du spectre de phase. Le spectre de phase ne se retrouve alors plus cohérent avec le spectre d'amplitude puisque les valeurs de phases, dont les dérivées correspondent à ω_k , sont utilisées dans des canaux où la fréquence devrait maintenant être $\omega_k + \Delta\omega_k$.

Cependant, si l'on démodule la phase de la TFCT en ω_k par $e^{-j\omega_k\tau}$ (la phase devient alors constante en fonction du temps, ce qui est obtenu directement en utilisant la convention passe-bas) et qu'on la remodule à la bonne fréquence par $e^{j(\omega_k + \Delta\omega_k)\tau}$, alors on compense les phases afin que la nouvelle fréquence soit bien $\omega'_k = \omega_k + \Delta\omega_k$. Il en résulte que la modification de la TFCT peut s'exprimer sous la forme suivante :

$$S'(\tau, \Omega) = S(\tau, \Omega - \Delta\omega_k) e^{j\Delta\omega_k\tau} \quad (2.30)$$

pour toutes les fréquences Ω situées au voisinage de la sinusoïde-cible de fréquence $\omega_k + \Delta\omega_k = \alpha\omega_k$.

La reconstruction est effectuée par FFT inverse, extrêmement moins coûteuse en terme de calculs que la méthode précédente en banc de filtres. D'autre part, il est inutile d'extraire explicitement la fréquence instantanée (évitant ainsi les lourds calculs d'arctangente et de déroulement de phase) puisque la modification consiste uniquement à effectuer une multiplication complexe. Cette dernière remarque est valable si $\Delta\omega_k$ peut être déterminé uniquement à partir des pics fréquentiels du spectre d'amplitude, ce qui est généralement réalisé par une interpolation quadratique (on fait passer une parabole par les 3 valeurs les plus élevées du pic et l'index du maximum indique la fréquence de la sinusoïde).

Cette technique est largement utilisée dans des contextes de transformation musicale [Fav01] puisqu'elle permet de déplacer composantes fréquentielles indépendamment les unes des autres.

2.3.4 Méthodes adaptatives

Nous avons vu que les méthodes fréquentielles classiques doivent faire face à des problèmes dont les solutions sont contradictoires. Nous nous intéressons donc ici aux solutions alternatives envisagées, qui prennent en compte l'information contenue dans le signal afin de réaliser un traitement mieux adapté.

Principe des méthodes adaptatives

Le but de ces méthodes se regroupe selon deux catégories, associées aux types de problèmes rencontrés.

D'une part, pour éviter la coloration, des algorithmes de verrouillage de la phase ont été mis au point. Ils diminuent fortement les artefacts pour lesquels ils ont été conçus, mais souffrent encore de l'étalement des transitoires.

D'autre part, pour éviter l'étalement des transitoires, des algorithmes ont été conçus pour conserver les relations de phase nécessaires à la localisation temporelle de l'impulsion à des instants donnés.

Méthodes à verrouillage de phase

Puckette [Puc95] propose une méthode permettant d'atténuer la coloration et la réverbération. Elle consiste à conserver les relations de phase "verticales" existantes entre les canaux adjacents à un pic fréquentiel. En effet, ces relations entre canaux se perdent au bout d'un certain temps à cause du fonctionnement récursif de l'algorithme, pour lequel les erreurs s'accumulent [SP92]. Pour un signal sinusoïdal pur, toutes les phases des canaux sont identiques (au signe près, selon que la transformée de Fourier de la fenêtre d'analyse soit positive ou négative). L'idée ici est donc d'appliquer la phase du canal où se situe un pic dans les canaux qui lui sont adjacents.

Dans cette méthode, l'auteur propose également un autre type d'échantillonnage de la TFCT discrète que celle utilisée classiquement : dans la méthode classique, la fréquence instantanée est tirée des valeurs de phase calculées aux instants L_i . Ici, une fois le pas de synthèse E choisi, l'utilisation d'une FFT supplémentaire réalisée sur le signal original à l'instant $L_i - E$ permet de connaître la différence de phase à appliquer entre les instants de synthèse E_{i-1} et E_i . Cette astuce permet d'éviter les fastidieux calculs d'arctangente nécessaires aux calculs de la phase au prix d'une FFT supplémentaire.

Cet algorithme complet est implanté de manière très efficace car il ne requiert que des multiplications complexes et aucun déroulement de phase.

Laroche et Dolson [LD97, LD99a, Dol00] proposent une amélioration du principe précédent qui se révèle être trop approximatif, en se rapprochant d'un modèle sinusoïdal : l'algorithme effectue une détection simple de pics et subdivise l'axe fréquentiel en des "régions d'influence" selon la position de ces pics. Ces "régions d'influence" déterminent l'étendue des fréquences pour lesquelles la sinusoïde estimée impose ses phases.

Cet algorithme nommé "verrouillage à phases identiques" permet de ne calculer le déroulement de phase que pour les pics fréquentiels, diminuant ainsi la charge de calculs comparé à la méthode classique.

Une amélioration de cet algorithme est également proposée, nommée "verrouillage à phases proportionnelles", qui permet de tenir compte du passage d'une sinusoïde d'un canal vers un autre. Cette fois, il n'y a plus de gain en calculs comparé à la méthode classique mais la qualité est améliorée.

Méthode par construction itérative de la phase

Griffin et Lim [GL84] proposent un algorithme de synthèse nommé WOLA ("Weighted OverLap-Add") qui permet de minimiser l'erreur quadratique entre la TFCT du signal synthétisé et une TFCT modifiée.

Ils utilisent cet algorithme afin de construire itérativement un signal temporel dilaté à partir de l'information du module de la TFCT modifiée. Les itérations successives consistent à reconstruire une information de phase cohérente avec le module de sorte que le signal synthétisé possède une TFCT la plus proche possible (au sens des moindres carrés) de la TFCT désirée (mais non valide). Cette construction de la phase est rendue possible par l'importante redondance d'information contenue dans le spectrogramme.

Ces travaux sont repris par Roucos et Wilgus [RW85]. Ceux-ci testent la rapidité (ou plutôt la lenteur) de convergence de l'algorithme avec différents signaux comme première estimée. Ils sont également étudiés par Veldhuis [VH96] et Roehrig [Roe90]

Selon Laroche [Lar98], cette méthode permet de diminuer à la fois l'étalement de transitoire et la coloration. Cependant, l'algorithme est très cher en terme de puissance de calculs et sa convergence ne mène pas nécessairement au minimum global de l'erreur quadratique (le résultat n'est pas assuré être le meilleur). De plus, il ne semble pas capable de restituer le signal original lorsque le taux de dilatation est fixé à 1 [Roe90].

Une méthode similaire à cette idée de reconstruction itérative de la phase est proposée en appendice de [QDH95] en utilisant le signal analytique. Il s'agit de construire le signal dilaté $s'(n)$ à partir de l'enveloppe temporelle $a(n)$ et de l'enveloppe spectrale $A(\Omega)$ désirées.

Méthodes utilisant une détection de transitoires

Quatieri, Dunn et Hanna [QDH95] proposent de conserver les relations de phase entre des sous-bandes à un instant précis plutôt que d'imposer l'enveloppe temporelle à travers $M_k(n)$. Cet instant est déterminé par le maximum de l'enveloppe. L'implantation est réalisée par un banc de 21 filtres utilisant une réponse impulsionnelle de type Gabor de 2 ms.

Une des faiblesses de cette méthode réside dans l'hypothèse qu'il n'y a pas plus qu'un changement significatif (un transitoire) dans les caractéristiques de la forme d'onde.

Pour résoudre ce problème, les sous-bandes sont réparties en plusieurs groupes pour lesquels l'événement principal (transitoire) se situe à un moment commun. On autorise de cette manière la présence de transitoires à plusieurs instants différents dans une même période d'observation, sous la condition que ces différents transitoires aient lieu dans des gammes de fréquences différentes.

Les auteurs remarquent qu'une technique d'analyse/synthèse multirésolution pourrait améliorer le problème d'étalement du transitoire.

Duxbury *et al.* [DDS02] exposent une méthode dans laquelle le taux de dilatation est variable. Cette technique, déjà suggérée par Settel et Lippe [SL95], Covell *et al.* [CWS98] et Bonada [Bon00a], permet de ne pas dilater les transitoires. Néanmoins, cette précaution n'est généralement pas suffisante pour garantir l'intégrité de ces derniers. Par conséquent, ils introduisent un verrouillage de phase au moment de l'apparition du transitoire. Cette technique perturbe la continuité des phases par rapport à la technique classique. Elle introduit une légère discontinuité due à une variation rapide de fréquence, qui reste toutefois inaudible car les régions immédiatement avant et après un transitoire bénéficient du masquage temporel [Moo97].

Les positions de début et de fin des transitoires sont détectées grâce à la variation de l'énergie d'un signal spécifique, issu d'une décomposition sinus/transitoire élaborée par les auteurs [DDS01].

Méthodes utilisant l'information de non-stationnarité du signal

Masri [Mas96] interprète les déformations des spectres d'amplitude et de phase lorsque le signal n'est pas stationnaire à l'intérieur d'une fenêtre d'analyse. Il est ainsi capable, dans certaines mesures, de déterminer le type de variation (amplitude ou fréquence) associé au signal.

Cette information peut être exploitée afin d'enchaîner plus continûment les phases des grains temporels successifs dans le cas d'un signal non-stationnaire [BJB01, PR99].

2.3.5 Bilan des "méthodes fréquentielles"

Les méthodes fréquentielles sont des techniques qui demandent une puissance de calculs élevée mais qui sont de plus en plus employées grâce à la rapidité des ordinateurs actuels.

Les résultats sonores des méthodes aveugles sont satisfaisants sur une grande majorité de sons simples (typiquement monophoniques) et sans transitoires marqués. On peut encore obtenir de bons résultats lorsque l'on optimise les paramètres de fenêtrage pour des signaux complexes (polyphoniques) et transitoires particuliers.

Cependant, les valeurs de ces paramètres sont généralement contradictoires et il n'existe pas de compromis acceptables pour tous les types de sons : des artefacts audibles sont produits. D'un côté, il est possible de faire de ces défauts sonores une utilisation musicale intéressante, mais dans un contexte de dilatation-p fidèle au son original, ces imperfections deviennent rédhibitoires. Ceci est regrettable car l'emploi d'une méthode aveugle est bien adaptée au traitement multicanal puisque les relations de phases entre canaux reste inchangée. De plus, la qualité des résultats obtenus avec des taux de dilatation élevés (supérieurs à 20%) et très élevés (supérieurs à 100%) sont largement meilleurs que ceux obtenus avec des méthodes temporelles.

De nombreuses techniques fréquentielles adaptatives ont permis de résoudre un certain nombre de ces problèmes, mais les résultats sonores ne sont toujours pas à la hauteur des exigences de qualité, notamment en ce qui concerne les transitoires.

2.4 Méthodes temps-fréquence

Nous appelons "méthodes temps-fréquence" les méthodes de dilatation-p faisant appel à une représentation temps-fréquence autre que la RFCT.

La RFCT se révèle en effet mal adaptée à un traitement où de bonnes résolutions temporelles et fréquentielles sont essentielles. Une bonne résolution temporelle est obtenue avec des fenêtres temporelles courtes (on peut alors déterminer précisément le début et la fin d'une attaque) alors qu'une bonne résolution fréquentielle est obtenue avec des fenêtres temporelles longues (on peut alors déterminer précisément les fréquences des composantes spectrales présentes). Cette différence entre analyse large bande et bande étroite est schématisée à travers les deux spectrogrammes du même signal temporel de la figure 2.39. Le spectrogramme à large bande montre une bonne précision quant à la localisation des attaques (pour $t=0,7s$ par exemple) sans parvenir à donner d'information précise sur les fréquences présentes, alors que le spectrogramme à bande étroite montre une bonne précision de la localisation des fréquences sans parvenir à donner d'information sur la position des attaques.

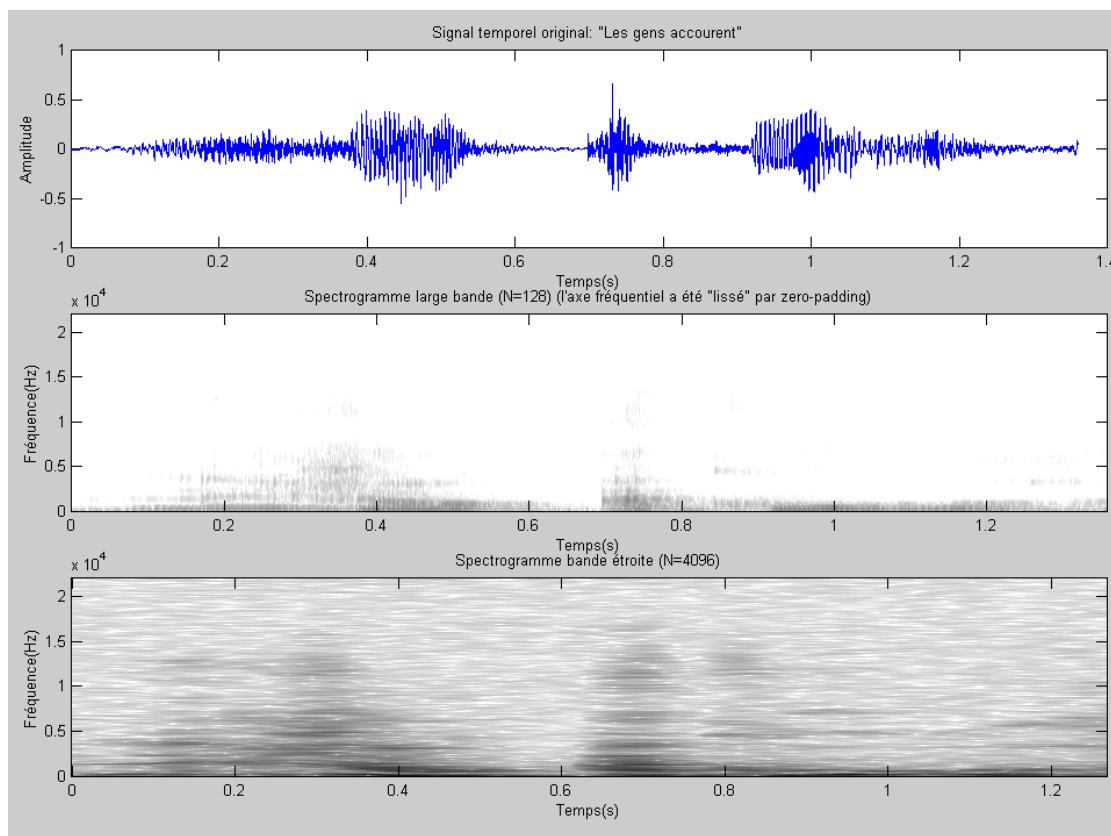


Figure 2.39 – *Signal temporel (haut) et spectrogrammes correspondants : large bande (milieu) et bande étroite (bas)*

On met donc en place une sorte de généralisation des méthodes fréquentielles dans le sens où les bandes de fréquences ne sont plus obligatoirement de largeur constante.

2.4.1 Principe général des méthodes temps-fréquence

Le principe des méthodes temps-fréquence reste le même que celui des méthodes fréquentielles : les modifications apportées à la représentation pour effectuer une dilatation-p

sont du même type que celles utilisées dans les méthodes fréquentielles : modification des phases afin d'éviter les discontinuités lors du recouvrement dans les méthodes de synthèse par OLA, ou interpolation des fonctions d'amplitude et de phase au cours du temps des composantes sinusoïdales dans les méthodes de synthèse en banc de filtres. La différence réside dans le type d'analyse qui est réalisé sur le signal.

L'analyse par TFCT pour obtenir une bonne résolution temporelle doit être réalisée avec une fenêtre étroite, ce qui entraîne une mauvaise résolution fréquentielle. Inversement, une bonne résolution fréquentielle est obtenue avec une fenêtre large, ce qui entraîne une mauvaise résolution temporelle. Cela explique en partie le compromis impossible à faire entre des transitoires qui ne "bavent" pas (petite fenêtre) et une absence de coloration (grande fenêtre).

On envisage donc de réaliser une analyse qui permettrait de tirer parti des spécificités de l'oreille : la résolution fréquentielle est élevée à basse fréquence, mais plus faible à haute fréquence [ZF90, Roa96]. De plus, l'information perceptive caractéristique d'un transitoire est surtout présente dans les hautes fréquences, pour lesquelles le contenu spectral est généralement très énergétique. D'autre part, les basses fréquences ne nécessitent pas une grande précision temporelle puisqu'elles évoluent lentement.

Il semble donc judicieux d'avoir une bonne résolution fréquentielle à basse fréquence, au détriment d'une bonne précision temporelle, et une bonne résolution temporelle à haute fréquence, au détriment d'une bonne précision fréquentielle.

Pour mettre en place un tel type d'analyse, il est nécessaire de comprendre les paramètres régissant résolution fréquentielle et temporelle, et ce à travers les différentes interprétations de l'analyse par TFCT.

Dans l'interprétation en banc de filtres précédemment étudiée, nous notons que les filtres utilisés possèdent tous la même longueur (durée de la réponse impulsionnelle) et que leurs fréquences centrales sont réparties régulièrement. Dans l'interprétation en transformée de Fourier, les grains temporels ont une durée identique car la fenêtre appliquée avant la transformée de Fourier, qui fournit des valeurs aux fréquences multiples d'une fréquence fondamentale, est toujours la même. Dans l'interprétation atomique, les atomes d'analyse sont tous de taille identique et de fréquences proportionnelles.

Les méthodes fréquentielles réalisent à travers la TFCT une analyse à bande de fréquence constante ($\Delta\Omega = C^{te}$). La largeur de bande est liée à la fonction h par le biais de la dualité temps-fréquence : plus le support temporel est étendu, plus le support fréquentiel est étroit. Ainsi, pour une fonction h donnée, la largeur de bande est toujours la même.

Nous proposons maintenant de considérer des fonctions h qui ne sont plus de durée constante H , mais qui dépendent de la fréquence. Cela revient à adapter la durée des filtres ou de la fenêtre d'analyse à la fréquence d'analyse. On obtient donc des atomes $h_{p,k}$ de support fréquentiel variant proportionnellement avec k .

Pour que l'analyse puisse être un processus réversible (reconstruction parfaite), il est nécessaire de prendre des précautions lors du choix de la durée des fenêtres et de la discrétisation des fréquences.

Ces choix mènent à la création des diverses représentations étudiées dans la suite.

2.4.2 Représentation temps-échelle

La représentation temps-échelle est obtenue par une transformée en ondelettes [KMMG87, Tor98], notée TO et donnée par l'équation suivante :

$$TO(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} s(t) \bar{g}\left(\frac{t-b}{a}\right) dt = \langle s, g_{b,a} \rangle$$

avec $g_{b,a} = \frac{1}{\sqrt{a}} g\left(\frac{t-b}{a}\right)$ l'atome ou ondelette d'analyse.

Les paramètres a et b sont les paramètres d'échelle (ou résolution temporelle: durée de l'atome) et de translation temporelle. La fonction $g(t)$ est appelée "ondelette analysante" et doit vérifier la condition d'admissibilité: en pratique, elle doit être à valeur moyenne nulle, c'est-à-dire qu'elle doit osciller. Cette fonction est généralement exprimée sous la forme d'une fenêtre (ou enveloppe) modulée par une sinussoïde (ou une exponentielle complexe). Ainsi, pour une fenêtre h modulée à la fréquence ω , l'atome d'analyse s'exprime sous la forme:

$$g_{b,a} = \frac{1}{\sqrt{a}} h\left(\frac{t-b}{a}\right) e^{j\omega\left(\frac{t-b}{a}\right)}$$

Cette formule est à comparer à la formule d'interprétation atomique de la TFCT dans l'équation 2.27: le paramètre a permet de jouer sur la durée du grain d'analyse, et donc sur la largeur de bande de l'analyse. Plus a augmente, plus le support de $g_{b,a}$ augmente, et plus la largeur de bande diminue. La fréquence de l'ondelette est inversement proportionnelle au paramètre a .

Il s'agit donc d'une analyse pour laquelle la largeur de bande est proportionnelle à la fréquence ($\frac{\Delta\Omega}{\Omega} = C^{te}$).

Dilatation-p

Ellis [Ell92] soumet l'idée d'une analyse de ce type, qu'il nomme CQFB ("Constant-Q Filter-Bank", où Q désigne le facteur de qualité, c'est-à-dire la largeur de bande relative). Il remarque que la dilatation-p par une méthode fréquentielle avec une taille de fenêtre d'analyse H entraîne une transposition des fréquences inférieures à $1/H$. Il explique ce phénomène dans le cas d'un signal de parole, constitué pour les sons voisés d'une succession périodique de transitoires appelés "impulsions glottales" et séparés par une période fondamentale P_0 , par un éloignement temporel relatif de chacune de ces impulsions lorsque $P_0 > H$. On modifie donc la fréquence fondamentale du locuteur. Si la fenêtre d'analyse est trop courte, la distance entre les pulsations glottales est modifiée, entraînant une modification de la fréquence fondamentale (bien que les formants soient préservés). Inversement, si la fenêtre d'analyse est trop longue, les détails temporels sont étalés. Il remarque qu'avec une analyse de type ondelettes, ces phénomènes peuvent être évités puisque la taille des fenêtres temporelles d'analyse augmente lorsque la fréquence diminue, permettant alors de recouvrir plus d'une période à analyser.

Les détails de l'implantation ne sont pas révélés, mais il semble que des ondelettes dyadiques soient utilisées, fournissant des coefficients alimentant des filtres d'une octave plus bas que ceux d'analyse dans le cas d'une dilatation-p de facteur 2 (la période d'échantillonnage critique étant deux fois plus longue, l'élongation est réalisée implicitement).

Transposition-p

Garras et Sommen [GS98] exposent également une méthode qu'ils nomment "Vocodeur de phase à Q constant". L'analyse obtenue est donc une sorte d'analyse par transformée en ondelettes. Celle-ci est réalisée de manière très efficace grâce à un échantillonnage exponentiel du

signal original avant d'effectuer une FFT. La transformée à un instant donné est alors donnée par :

$$T'(\omega) = \sqrt{\log(a)} \int_0^{+\infty} \frac{s(t)}{\sqrt{t}} e^{j\omega \log_a(t)} dt$$

La transposition en elle-même est réalisée classiquement en multipliant la phase par le facteur α et en synthétisant le signal de sortie par un banc de sinusoides.

2.4.3 Représentation basée sur l'échelle des Barks

Hoek [Hoe01] obtient une représentation multirésolution à partir d'un fenêtrage de taille unique. Il propose en effet d'appliquer au spectre d'amplitude à court terme, produit pas la TFCT, une série de filtrages passe-bas. La convolution dans le domaine fréquentiel étant équivalente à une multiplication dans le domaine temporel, ce filtrage revient à réduire artificiellement la taille de la fenêtre temporelle d'analyse. La réponse impulsionnelle du filtre passe-bas, appelée "fonction de noyau variable", est adaptée à la fréquence de manière à refléter le comportement de la réponse en fréquence de l'oreille. Elle est exprimée sous la forme de l'équation aux différences suivante :

$$y_{out}(f) = [1 - w(f)]y_{in}(f) + w(f)y_{out}(f - 1)$$

où $w(f)$ est la "fonction de noyau variable" donnée par :

$$w(f) = 0,4 + 0,26 \arctan(4 \ln(0,1f) - 18)$$

L'effet de ce filtrage "adaptatif" sur le spectre mène à réaliser un fenêtrage dans le domaine temporel dont le support varie avec la fréquence.

Une détection de pics très simple est utilisée (existence d'un pic en Ω_k si $M_{\Omega_k} > M_{\Omega_{k-1}}$ et $M_{\Omega_k} > M_{\Omega_{k+1}}$), et puisque l'information fréquentielle autour du pic est importante (elle représente les modulations de fréquence et d'amplitude associées à la composante sinusoïdale [Mas96]), elle est prise en compte dans la construction d'un vecteur, somme des valeurs complexes des données fréquentielles autour de ce pic.

2.4.4 Représentation adaptative

Bonada [Bon00a] suggère une représentation dans laquelle les largeurs de bande sont constantes par morceaux.

Cette représentation est réalisée par la mise en parallèle de plusieurs TFCT (en nombre K théoriquement quelconque, mais fixé à trois dans ce cas), chacune utilisant un fenêtrage particulier (type et taille de fenêtre différente). Chaque TFCT contient le spectre complet du signal original doté d'une résolution temporelle/fréquentielle spécifique. On obtient donc une analyse multirésolution où les largeurs de bande sont inversement proportionnelles à la taille de fenêtre employée. Le but est d'utiliser à la fois les données spectrales hautes fréquences de la TFCT réalisée avec une petite fenêtre (bonne résolution temporelle) et les données spectrales basses fréquences de la TFCT réalisée avec une grande fenêtre (bonne résolution fréquentielle).

Il existe évidemment une information redondante entre toutes les TFCT qu'il faut éliminer après la transformation. Cette opération est réalisée à l'aide de K filtres passe-tout dont les fréquences de coupure sont variables et adaptées au signal (pour éviter qu'une même sinusoïde

se retrouve dans deux canaux différents). Ces filtres variant dans le temps sont appliqués aux TFCT avant la reconstruction par OLA.

La méthode de transformation employée repose sur l'interpolation simple des spectres d'amplitude ($E = L$, voir section 2.3.2) avec verrouillage de phase (pour améliorer les sons complexes) et détection de transitoires (pour améliorer l'étalement de transitoire).

Le verrouillage de phase utilisé est classique (voir section 2.3.4) mais il est adapté à la méthode d'interpolation simple des spectres.

La détection de "changements rapides" (transitoires) est réalisée par un critère portant sur plusieurs indices (énergie dans un banc de filtres, coefficients cepstraux sur l'échelle des Mels, et leurs dérivées). Lorsqu'un transitoire est détecté, la phase originale est imposée dans les canaux supérieurs à une fréquence de coupure déterminée par l'analyse, ainsi que dans les canaux où il n'existe pas de pics fréquentiels stables. Pour $K > 1$, différents types de transitoires peuvent être détectés grâce aux différentes TFCT dont les résolutions temporelles et fréquentielles sont optimisées (fenêtres longues pour détecter des changements rapides dans les basses fréquences comme les coups de grosse caisse, fenêtres courtes pour détecter des changements rapides dans les hautes fréquences comme les frappes de cymbales).

On peut donc voir cet algorithme comme une méthode temporelle pour laquelle les grains sont modifiés uniquement en vue d'assurer la continuité des phases des composantes sinusoïdales. Pour cela, une décomposition en sous-bandes de largeurs arbitraires est effectuée dans laquelle les relations de phase sont conservées sauf dans les sous-bandes contenant une sinusoïde : la phase y est alors calculée à la manière du vocodeur à verrouillage de phase.

2.4.5 Représentation temps-fréquence multi-résolution

D'une part, nous avons vu que la TFCT offrait une représentation temps-fréquence de résolution temporelle/fréquentielle fixe (la fenêtre d'analyse est constante quelle que soit la fréquence analysée). D'autre part, nous avons vu que la TO possédait une résolution fréquentielle proportionnelle à la fréquence (fenêtre d'analyse courte pour les hautes fréquences, longue pour les basses fréquences). Or, aucune de ces deux représentations n'est capable de fournir une information précise simultanément en temps et en fréquence.

Le principe de la représentation temps-fréquence multi-résolution consiste à extraire simultanément de l'information de ces deux types d'analyse : pour chaque échelle (c'est-à-dire chaque taille de fenêtre), un spectre complet est calculé pour toutes les fréquences, ou réciproquement, chaque fréquence est calculée avec toutes les tailles de fenêtres. On peut voir cette analyse comme une série de TFCT avec des fenêtres de taille différentes, ou de manière équivalente comme une série de TO avec des fréquences ω différentes.

La TFCT comme la TO continue sont déjà des représentations extrêmement redondantes. Une analyse temps-fréquence multi-résolution fournit donc une quantité d'information beaucoup trop importante. Il est nécessaire d'organiser cette information. Le Matching Pursuit [MZ89, Gri99] permet d'effectuer cette organisation des données, en utilisant un algorithme itératif qui ne conserve que les atomes temps-fréquence dont l'énergie dans le signal original est significative.

Mathématiquement, le principe est de décomposer le signal original sur une famille d'atomes temps-fréquences donnés par :

$$g_{a,b,f}(t) = \frac{1}{\sqrt{a}} h\left(\frac{t-b}{a}\right) e^{j2\pi f(t-b)}$$

En pratique, les atomes sont issus d'un dictionnaire fini \mathcal{D} donné par :

$$\mathcal{D} = \{g_{a,b,f} \text{ tels que } a = 2^j, b = n2^j, f = k2^j\}$$

L'algorithme itératif pour effectuer cette décomposition est le suivant :

1. Calculer $|\langle s, g \rangle|^2, g \in \mathcal{D}$
2. Sélectionner g_i tel que $g_i = \operatorname{argmax} |\langle s, g \rangle|^2$
3. Calculer le résidu $R_s = s - \langle s, g_i \rangle g_i$
4. Réitérer en remplaçant s par R_s

Après M itérations, le signal original s'exprime alors sous la forme :

$$s = \sum_{m=1}^M \langle R_s^{m-1} | g_m \rangle g_m + R_s^M$$

Le signal étant décomposé en atomes g_m , la transposition en fréquence est réalisée simplement en modifiant les fréquences de chacun des atomes :

$$g'_{a,b,f} = g_{a,b,\alpha f}$$

2.4.6 Méthodes recourant à des décompositions préalables

Les décompositions préalables permettent d'appliquer différentes méthodes de dilatation-p aux différents signaux. Il s'agit en général d'extraire les transitoires et les conserver intacts à travers un décalage de leurs position selon la fonction de dilatation.

Décomposition sinus/transitoire/bruit

Hamdy *et al.* [HTCT97] réalisent une décomposition du type "sinus/transitoire/bruit" afin d'appliquer des méthodes de dilatation-p adaptées aux différents signaux avant de les sommer.

La partie "sinus" est extraite par une technique développée par Thomson [Tho82]. La dilatation-p est réalisée en interpolant les valeurs des signaux analytiques de chacun des partiels $s_k(\tau)$ démodulés, avant de les remoduler :

$$s'_k(\tau) = \left[s_k(\tau/\alpha) e^{-j\omega_k(\tau/\alpha)} \right] e^{j\omega_k\tau}$$

Les signaux analytiques sont obtenus grâce à des filtres dont les largeurs de bande sont adaptées à la discrimination fréquentielle de l'oreille. Cette technique est en fait similaire à celles employées dans les méthodes temps-fréquence. Un premier résidu est issu de cette étape, correspondant au signal original auquel a été enlevé la partie sinusoïdale.

La partie transitoire est extraite de ce premier résidu en utilisant une décomposition en paquets d'ondelettes dont les coefficients les plus élevés (haute énergie) sont considérés comme constitutifs d'un transitoire. Les positions de ces coefficients sont représentés par une forme d'onde carrée (1 pour un transitoire présent, 0 sinon) et leurs origines sont déplacés selon la fonction de dilatation D . Cela revient finalement à utiliser une méthode temporelle de dilatation-p. Un second résidu est issu de cette étape, correspondant au premier résidu auquel a été enlevé la partie transitoire.

La partie bruit, qui est le résidu final, est dilaté par une simple interpolation des coefficients d'ondelettes.

Verma et Meng [VM98] suggèrent le même type de décomposition dans laquelle les transitoires sont extraits et modélisés à partir d'une DCT (Transformée en Cosinus Discrète). Cette représentation possède l'avantage d'interpréter un transitoire comme une sinusoïde dont la fréquence est d'autant plus grande que sa position est éloignée du début de la fenêtre d'analyse. La dilatation-p d'un transitoire consiste donc à modifier les fréquences des sinusoïdes dans la représentation en DCT.

Levine [Lev98] modélise également les transitoires grâce à la DCT, mais la manière de les dilater, par décalage de leur position, est également à rapprocher des méthodes temporelles. Cette décomposition ressemble donc plus à une segmentation du signal, pour laquelle la partie sinus est dilatée grâce à des méthodes fréquentielles (dont l'implantation peut être faite par banc de filtres ou par transformée de Fourier) et la partie transitoire trouve naturellement sa place grâce à une méthode temporelle en utilisant la fonction de dilatation.

2.5 Conclusions sur la classification

On peut conclure de cette classification que :

1. Les méthodes temporelles sont des méthodes de dilatation-p, qui peuvent généralement s'écrire sous la forme suivante :

$$Dp[s](t) = \sum_i h_i(t - E_i)s(t - E_i + L_i)$$

avec h_i les fenêtres de pondération, L_i et E_i les marques de lecture et d'écriture. Elles diffèrent par les critères qui fournissent les marques de lecture et d'écriture, ainsi que par la taille et le type des fenêtres de pondération. La transposition n'est pas possible avec de telles méthodes.

2. Les méthodes fréquentielles sont basées sur le concept en 3 étapes "Analyse/Transformation/Synthèse" pour lesquelles :

L'**Analyse** est réalisée à bandes de fréquences constantes, et donnée par :

$$S(\tau, \Omega) = \int_{-\infty}^{+\infty} s(t)h(t - \tau)e^{-j\Omega(t-\tau)}$$

Cette analyse peut s'interpréter en banc de filtres, en Transformée de Fourier, ou encore en décomposition atomique, et mener ainsi à différentes implantations.

La **Transformation** de la représentation pour la **dilatation-p** est globalement donnée par :

$$S'_{Dp}(\tau, \Omega) = S\left(\frac{\tau}{\alpha}, \Omega\right)e^{-j\Omega(\frac{\tau}{\alpha} - \tau)}$$

Dans l'interprétation en banc de filtres, on peut voir cette transformation comme une modification de l'évolution temporelle (à travers le module M) de la sinusoïde de fréquence Ω avec modification adéquate des phases.

Dans l'interprétation en transformée de Fourier, on peut voir cette transformation comme une interpolation temporelle des spectres d'amplitude successifs avec modification adéquate des phases.

La **Transformation** de la représentation pour la **transposition-p** est globalement donnée par :

$$S'_{Tp}(\tau, \Omega) = S\left(\tau, \frac{\Omega}{\alpha}\right)e^{-j\Omega(\frac{\tau}{\alpha} - \tau)}$$

Dans l'interprétation en banc de filtres, on peut voir cette transformation comme l'application de l'évolution temporelle de la sinusoïde de fréquence $\frac{\Omega}{\alpha}$ à la sinusoïde de fréquence Ω avec modification adéquate des phases.

Dans l'interprétation en transformée de Fourier, on peut voir cette transformation comme une dilatation locale du spectre d'amplitude avec modification adéquate des phases.

Les différentes interprétations d'une même transformation-p sont seulement des points de vue différents d'une même formule mathématique, il ne faut pas oublier qu'elles sont équivalentes entre elles.

La transformation fait passer dans tous les cas d'une TFCT discrète $S(p, \Omega_k)$ à une TFCT discrète modifiée $S'(p, \Omega_k)$.

La **Synthèse** est réalisée en pratique soit par sommation des signaux sinusoïdaux temporels (interprétation en banc de filtres), soit par addition-recouvrement des grains temporels obtenus par FFT inverse (interprétation en transformée de Fourier).

3. Les méthodes temps-fréquence diffèrent des méthodes fréquentielles par le type d'analyse effectué (la largeur de bande n'est plus constante). Il s'agit en pratique, lors de l'analyse, de changer la taille de la fenêtre selon la fréquence analysée.

Aucune de ces méthodes ne semble donner entière satisfaction sur tous les types de sons : schématiquement, les méthodes temporelles souffrent de redoublements d'attaques et de discontinuités de partiels sur des signaux inharmoniques et basses fréquences; les méthodes fréquentielles souffrent d'étalement des transitoires et de coloration sur des sons complexes. Nous développons donc dans la suite de ce document plusieurs méthodes, qui nous ont été inspirées par cette classification, afin de les comparer pour sélectionner celle qui donne les meilleurs résultats en vue d'une implantation temps-réel sur l'HARMO.

Chapitre 3

Innovations algorithmiques et évaluations

Je présente dans ce chapitre les contributions algorithmiques que j'ai apporté dans le domaine des transformation-p.

Tout d'abord, j'expose une étude réalisée sur l'anisochronie émanant des méthodes temporelles, dont les conclusions sont utiles pour améliorer ces méthodes, mais peuvent également servir dans le cadre des méthodes temps-fréquence.

Ensuite, je présente différentes méthodes temps-fréquence dont la représentation est mieux adaptée à l'oreille que ne le sont les méthodes à base de TFCT

Puis je développe différentes méthodes "couplées", qui découlent naturellement de la classification que j'ai réalisée.

Enfin, je propose une amélioration des méthode temporelles, qui répond aux exigences fixées par les contraintes technologiques et de qualité sonore.

3.1 Etude de l'anisochronie émanant des méthodes temporelles

En se basant sur le diagramme d'entrée/sortie, la dilatation temporelle consiste à approximer au mieux la "droite idéale" (objectif de traitement pour lequel la fonction de dilatation $D(n)$ est respectée) par une succession de segments de droites, alternativement parallèles à la "droite originale" (correspondant aux segments non modifiés), et parallèles à l'axe des ordonnées (correspondant aux segments insérés).

La régularité perceptive d'une séquence rythmique sera conservée si les pulsations se trouvent proches de la "droite idéale". Ceci est réalisé lorsque les segments insérés sont courts et dépourvus de pulsations. Cependant, l'insertion de segments courts n'est pas toujours possible. En effet, dans le cas de signaux harmoniques, il faut insérer au minimum une période fondamentale du signal pour éviter à la fois une discontinuité de la composante basse fréquence et une modification de sa fréquence. Cette contrainte mène à une **anisochronie** (irrégularité rythmique) lorsque la distance d'une pulsation à la "droite idéale" dépasse un certain seuil. Cette rupture d'isochronie peut se manifester sous plusieurs formes, fonction du tempo et de la longueur des segments insérés.

Nous nous intéressons ici uniquement au cas où les segments insérés ne contiennent pas de pulsation (donc absence de redoublement). Cela implique que l' IOI ("Inter-Onset Interval", l'espacement entre les pulsations) doit être supérieur au double de la longueur K du segment inséré car le fondu-enchaîné est créé à partir de 2 segments successifs K_A et K_B de durée K .

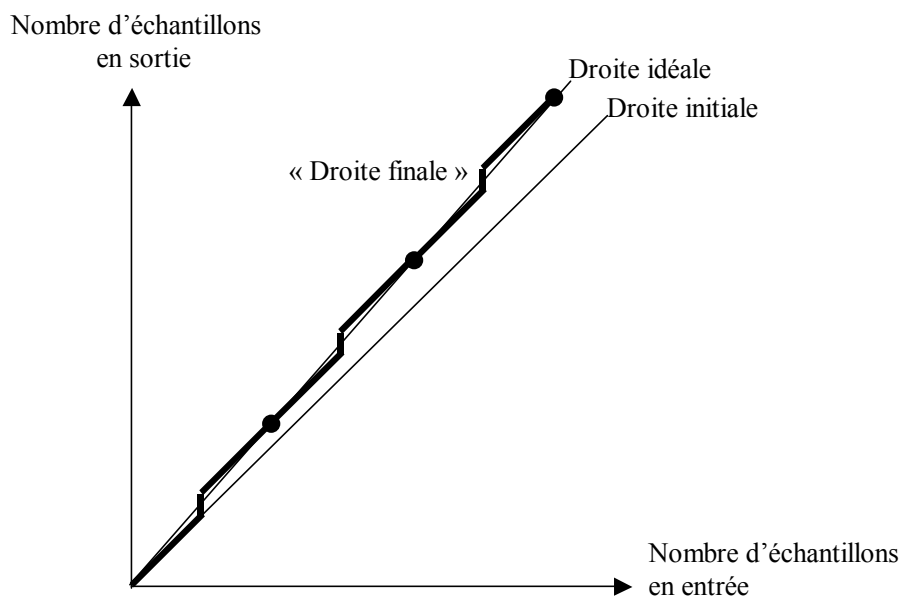


Figure 3.1 – *Droite finale approximant la droite idéale grâce à une succession de segments*

La figure 3.2 représente l'IOI d'une séquence rythmique régulière et indique la position des deux segments K_A et K_B donnant lieu à l'insertion du plus long segment K_M compatible avec notre contrainte. Un segment plus long entraînerait la duplication d'un transitoire.

Selon Woodrow [Woo51], l'intervalle temporel entre deux pulsations doit être supérieur à 100 ms de manière à entendre une succession de pulsations, et inférieur à 3 s de manière à les entendre comme un groupe de pulsations. Il en résulte que la contrainte précédente est assez faible puisqu'en pratique, les segments insérés sont inférieurs à 50 ms, soit la moitié de l'IOI minimal.

Lorsque la pulsation est peu marquée, comme par exemple dans les mélodies jouées legato, les déformations rythmiques sont analysées comme étant issues de l'interprétation musicale [DB93]. Puisque nous ne désirons pas modifier globalement l'interprétation, nous nous intéressons uniquement aux cas où les pulsations sont bien marquées et régulières, comme c'est généralement le cas dans les séquences rythmiques musicales traditionnelles. La définition du tempo utilisée ici diffère un peu de celle admise habituellement. Il s'agit ici de la plus petite subdivision du rythme.

3.1.1 Tempo "lent"

Pour un tempo suffisamment lent (les IOI sont grands par rapport aux "segments non modifiés du signal original"), il y a au maximum une pulsation par "itération". Si une seule de ces pulsations est présente immédiatement après un long segment inséré, on entendra un déplacement rythmique de cette pulsation, et de cette pulsation uniquement, mais le rythme final restera globalement proche du rythme idéal, comme le schématisent les figures 3.3 et 3.4.

Ce type d'anisochronie, nommé de "type 1" [FS95], a été également étudié dans [Sch78], [Hib83], [HD82], [HBG⁺94], [VVAf97].

Selon les expériences d'ajustement de Friberg et Sundberg [FS95], pour des IOI variant entre 100 et 240 ms, le JND ("Just Noticeable Difference" ou plus petite différence audible) est constant et d'environ 6 ms. Pour des IOI variant entre 240 et 1000 ms, le JND est plus élevé. Sa

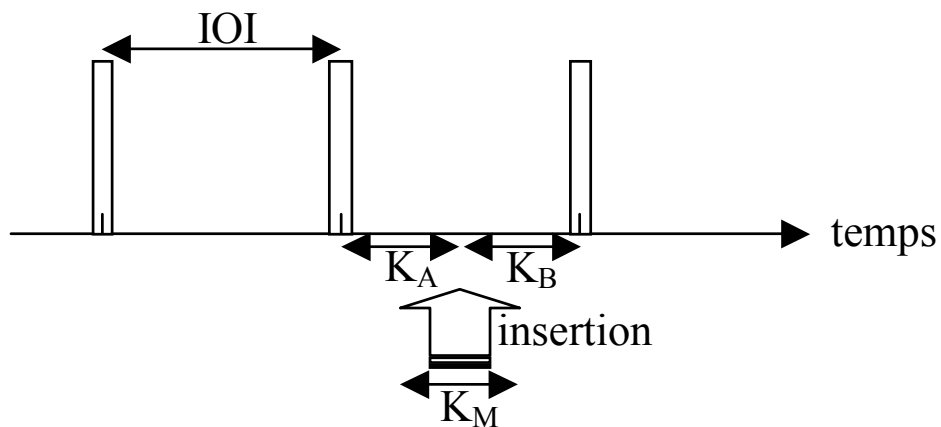


Figure 3.2 – Schéma d'une séquence rythmique dans laquelle est représenté le plus long segment inséré sans duplication du transitoire

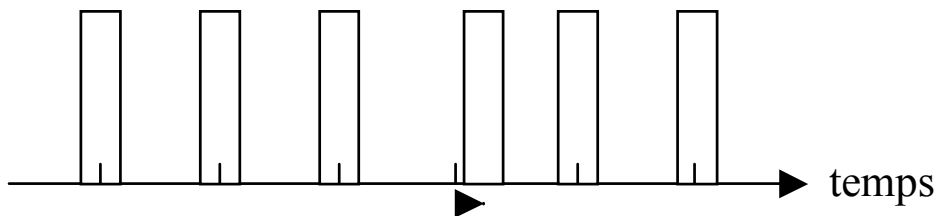


Figure 3.3 – Schéma du déplacement de la pulsation dans le cas d'un tempo "lent" (irrégularité rythmique de "type 1")

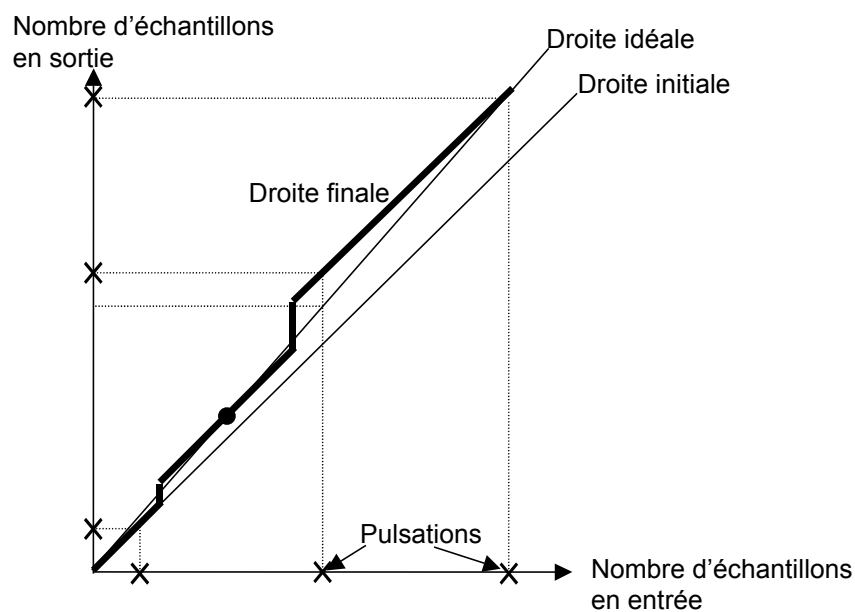


Figure 3.4 – Droite idéale bien approximée par des petits segments sauf pour une itération dans le cas d'un tempo lent (irrégularité rythmique de "type 1")

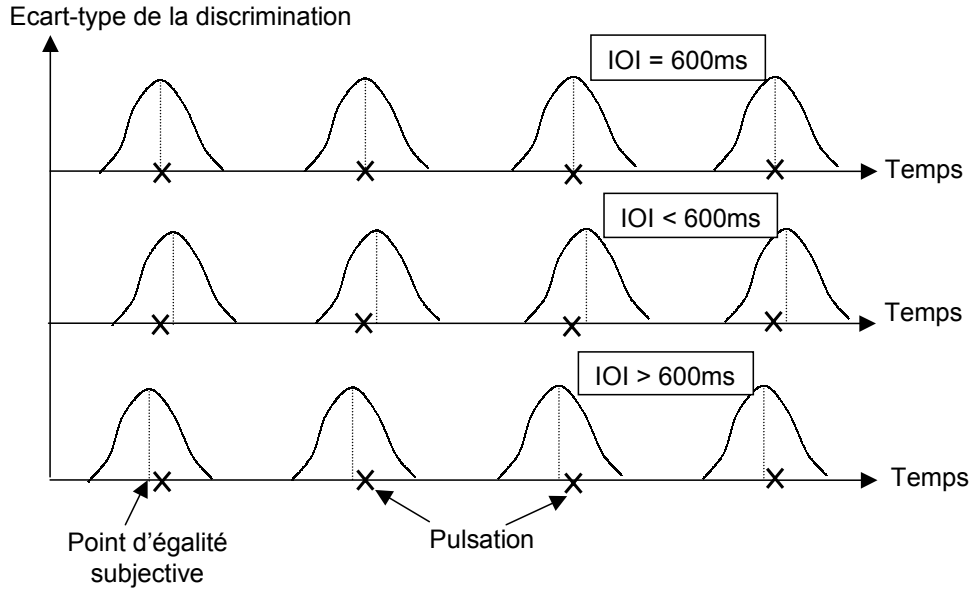


Figure 3.5 – Perception de la position des pulsations pour différents IOI

valeur relative est constante et d'environ 2,5%, c'est-à-dire qu'elle varie de 6 à 25 ms. D'autre part, le "point d'égalité subjective" (date à laquelle l'auditeur s'attend à entendre la pulsation) ne correspond au "point d'égalité objective", c'est-à-dire à la pulsation réelle, que pour un IOI d'environ 240 ms. A des tempi supérieurs ou inférieurs, la pulsation attendue se situe avant la pulsation réelle.

Selon les expériences à choix forcés de Halpern et Darwin [HD82], pour des IOI supérieurs à 400 ms, le JND relatif est d'environ 6%. Cependant, son expérience consiste à avancer ou retarder le quatrième "clic" d'une série de seulement quatre "clics", ce qui pourrait expliquer la diminution de sensibilité par rapport à [FS95]. Selon lui, le "point d'égalité subjective" correspond au "point d'égalité objective" pour un IOI d'environ 600 ms. Lorsque l'IOI est supérieur à cette valeur, le "point d'égalité subjective" est en avance par rapport à la pulsation réelle, et lorsque l'IOI est inférieur à cette valeur, le "point d'égalité subjective" est en retard par rapport à la pulsation réelle (ce qui pourrait expliquer que l'on a tendance à accélérer les tempi lents et à ralentir les tempi rapides), comme l'illustre la figure 3.5.

Puisque l'on désire au maximum une pulsation par "itération" pour respecter notre contrainte de "tempo lent", l'IOI doit avoir une valeur minimale, fixée par la valeur de la longueur du segment inséré. Ainsi, pour un facteur de dilatation α donné, les durées de K et de l'IOI sont liées par la formule suivante :

$$IOI = \frac{K}{\alpha}$$

Par exemple, pour $\alpha=4,2\%$, des segments insérés de 23 ms limitent le tempo à un IOI de 548 ms (109 battements par minute ou BPM). Si le tempo dépasse cette valeur, on se trouve dans le cas d'un "tempo rapide" qui est traité dans la suite. Le déplacement maximal d'une pulsation par rapport à la pulsation idéale correspond à la durée du segment inséré. Or un segment inférieur à 6 ms ne provoque pas d'anisochronie audible. Il en résulte que l'anisochronie de type 1 peut être audible uniquement pour des IOI supérieurs à $6/0,042=143$ ms (tempi inférieurs à 420 BPM). Les sons dont l'IOI est inférieur à cette valeur (battement de pales d'un hélicoptère par exemple) ne peuvent donc pas être affectés par ce type d'anisochronie.

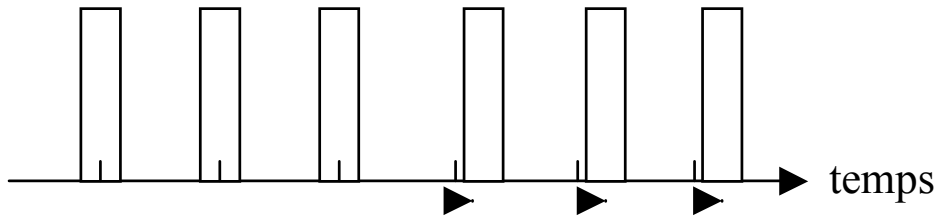


Figure 3.6 – Schéma du déplacement de la pulsation dans le cas d'un tempo "rapide" (irrégularité rythmique de "type 2")

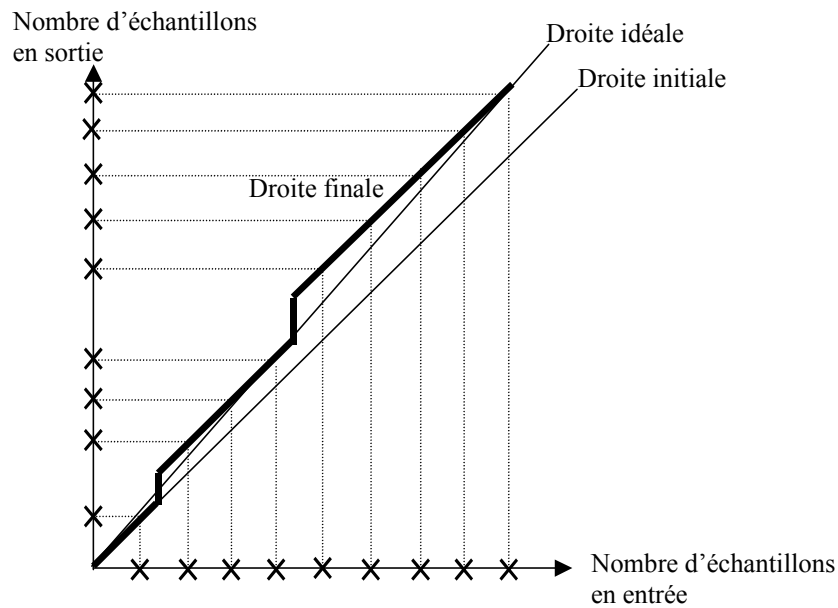


Figure 3.7 – Allongement ponctuel de l'IOI sur le tempo initial (irrégularité rythmique de "type 2")

Pour illustrer ce type de défaut, nous insérons dans un son original constitué de pulsations régulières ($IOI = 240$ ms, soit 250 BPM) (son [46]) un silence avant la 7^{ème} pulsation, de durée 6 ms (son [47]), 12 ms (son [48]) et 24 ms (son [49]).

3.1.2 Tempo "rapide"

Pour un tempo suffisamment rapide (les IOI sont petits par rapport aux "segments non modifiés du signal original"), il y a plus qu'une pulsation par "itération". Ces pulsations étant placées sur une droite parallèle à la "droite initiale", elles marquent alors le tempo initial. Lors de l'insertion d'un segment, on a un allongement local de l'IOI (irrégularité rythmique de "type 2"), schématisé en figures 3.6 et 3.7.

Selon Friberg et Sundberg [FS95], les valeurs de seuils sont sensiblement les mêmes que dans le cas précédent, c'est-à-dire 6 ms pour des IOI allant jusqu'à 240 ms, et 6% au-delà. Ce type d'irrégularité n'est pas rare, puisque l'insertion d'un segment de 45 ms est possible pour des tempi allant jusqu'à 666 BPM ($IOI > 2K = 90$ ms) sans présence de redoublement. Dans ce cas, on entendra une irrégularité de type 2 toutes les secondes, et selon le tempo, une seconde

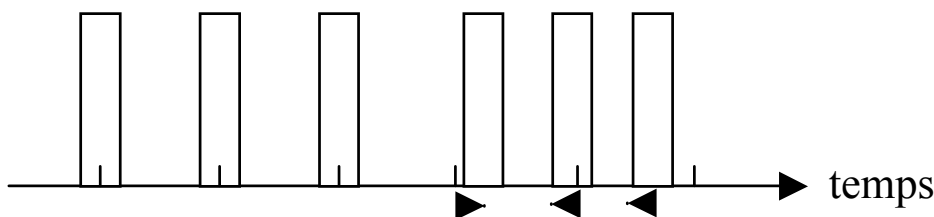


Figure 3.8 – Schéma du déplacement de la pulsation dans le cas d'un tempo "modéré" (irrégularité rythmique de "type 3")

peut contenir suffisamment de pulsations pour marquer un tempo.

Selon Hibi [Hib83], la sensibilité pour un IOI raccourci est légèrement plus grande que celle pour un IOI allongé pour des valeurs d'IOI comprises entre 143 et 240 ms, ce qui semble en accord avec [HD82]. Cependant, la sensibilité est plus grande pour l'allongement pour des IOI de 250 à 400 ms, et cette sensibilité s'inverse au-delà.

Pour illustrer ce type de défaut, nous insérons dans un son original constitué de pulsations régulières (IOI = 240 ms, soit 250 BPM) (son [46]) un silence avant la 7^{ème} pulsation qui décale toutes les pulsations suivantes, à l'opposé du cas précédent, de durée 6 ms (son [50]), 12 ms (son [51]) et 24 ms (son [52]).

3.1.3 Tempo "modéré"

Pour un tempo modéré, on peut se trouver dans le cas où la droite idéale est bien approximée (tempo idéal), et que l'insertion d'un long segment mène à la présence de plusieurs pulsations dans une itération (tempo initial). On est alors confronté à la fois à un décalage de la pulsation comme dans le cas du tempo lent (irrégularité de "type 1"), mais aussi à un changement de tempo (irrégularité de "type 3") comme le schématisent les figures 3.8 et 3.9.

Ce type d'anisochronie très spécifique n'est pas traité dans la littérature. Cependant, en ce qui concerne le changement de tempo, Michon [Mic64] trouve que la sensibilité est de 2% pour des IOI variant entre 300 et 1000 ms, donc potentiellement audible pour un facteur de dilatation de 4,2%.

Pour illustrer ce type de défaut, nous insérons dans un son original constitué de pulsations régulières (IOI = 240 ms, soit 250 BPM) (son [46]) un silence avant la 7^{ème} pulsation (et qui décale toutes les pulsations suivantes à l'opposé du cas précédent), de durée 6 ms (son [53]), 12 ms (son [54]) et 24 ms (son [55]), suivi d'une accélération de tempo.

3.1.4 Bilan sur l'anisochronie

Cette étude sur l'anisochronie nous amène à faire 3 remarques :

- Plus le tempo est lent, plus la sensibilité aux défauts rythmiques est faible.
- Les défauts rythmiques sont inaudibles si les segments insérés sont inférieurs à 6 ms.
- La perception d'un tempo est valable pour un IOI supérieur à 100 ms.

On peut former, grâce à ces remarques, la zone des défauts audibles indépendamment de leurs types (voir figure 3.10).

Si l'on se situe hors de cette zone, les défauts rythmiques sont inaudibles. C'est le cas pour des segments inférieurs à 6 ms quel que soit le tempo. Ainsi, pour la majorité des sons de

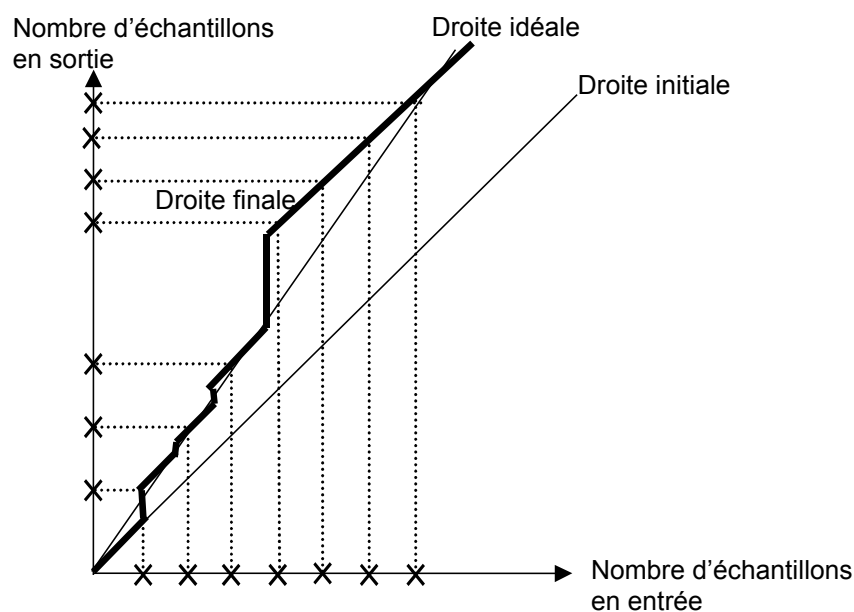


Figure 3.9 – Allongement de l'IOI et accélération de tempo (irrégularité rythmique de "type 3")

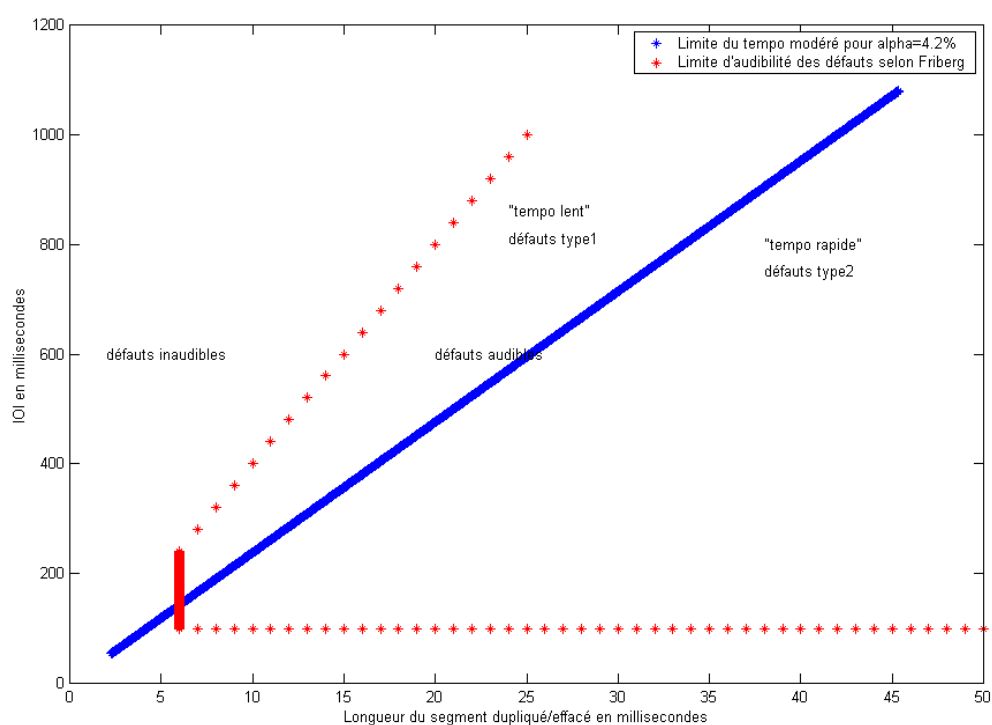


Figure 3.10 – Limites d'audibilité des défauts rythmiques

type transitoire, où les longs segments ne sont pas utiles, l'insertion de tels segments ne provoquera pas de déformation rythmique perceptivement audible. Malheureusement, pour des sons quasi-stationnaires harmoniques ayant des composantes basses fréquences inférieures à 167 Hz (correspondant à une période fondamentale de 6 ms), ou encore des sons inharmoniques dont le PPCM (plus petit commun multiple) des périodes fondamentales est supérieur à 6 ms, des segments aussi courts provoquent des discontinuités de synchronisation audibles.

D'autre part, la limite du tempo modéré marque la frontière entre "tempo rapide" et "tempo lent". Il en résulte que pour des IOI inférieurs à 143 ms (tempi supérieurs à 420 BPM), seuls les défauts de type 2 pourront être entendus. En revanche, pour des IOI supérieurs à 143 ms, les trois types d'irrégularité sont susceptibles de se produire, et même parfois de se combiner, selon les longueurs de segments insérés qui sont typiquement compris entre 10 et 40 ms. De plus, l'irrégularité rythmique se complique lorsqu'un segment inséré comporte une pulsation, qui donne alors naissance à un redoublement, ce qui est un cas que l'on n'a pas étudié ici.

Il semble que l'on ne peut pas tirer de conclusion en ce qui concerne une sensibilité rythmique plus ou moins accrue lors de la contraction ou de l'élongation temporelle car les résultats des expériences des différents auteurs (notamment sur le point d'égalité subjective) se contredisent.

Bien qu'on ne puisse pas conclure sur une longueur optimale de segment (celle-ci dépend des périodes fondamentales présentes dans le signal), on peut quand même tirer comme recommandation qu'en présence d'une séquence rythmique, il est préférable d'insérer des segments aussi courts possibles (idéalement inférieurs à 6 ms), à l'instar de ce que l'on ferait intuitivement pour approximer au mieux la droite idéale. La qualité du traitement de dilatation-p réside dans l'habilité à sélectionner les segments les plus courts possibles qui n'engendrent pas de discontinuité de synchronisation. On peut pour cela s'appuyer par exemple sur des considérations liées au masquage temporel.

Nous verrons que ces conclusions sont utilisées pour la mise au point de méthodes temps-fréquence et temporelles.

3.2 Méthodes temps-fréquence adaptées à l'audition

Le principe de ces méthodes consiste à utiliser une représentation temps-fréquence, mieux adaptée (en termes de résolution temporelle et fréquentielle) à l'audition que ne le sont les représentations à bandes de fréquences constantes (TFCT) et bandes de fréquences relatives constantes (TO). Il s'agit en fait d'une analyse tirant parti de ces deux types d'analyse, en se basant sur des considérations psychoacoustiques.

Principe

Une représentation temps-fréquence peut être obtenue grâce à la transformée suivante :

$$S(\tau, \Omega) = \int_{-\infty}^{+\infty} s(t) \bar{h}(t - \tau) e^{-j\Omega(t - \tau)} dt \quad (3.1)$$

Pour une fenêtre d'analyse h de support H constant, cette transformée correspond à la Transformée de Fourier à Court Terme (TFCT, voir section 2.3). Cette transformée possède une résolution temporelle, et donc fréquentielle, fixe. En effet, la résolution temporelle est donnée par la taille de la fenêtre temporelle d'analyse, soit H , et la résolution fréquentielle est donnée par la taille de la fenêtre fréquentielle d'analyse, soit $1/H$.

Pour une fenêtre d'analyse h_Ω qui dépend de la fréquence, de sorte que son support H_Ω soit inversement proportionnel à la fréquence Ω , cette transformée se comporte (à un terme de normalisation près) comme la Transformée en Ondelettes (TO, voir section 2.4.2), dans le cas où l'ondelette mère est une fenêtre modulée. Cette transformée possède une bonne résolution temporelle pour les hautes fréquences grâce à une petite fenêtre temporelle, et une bonne résolution fréquentielle pour les basses fréquences grâce à une grande fenêtre temporelle.

Par extension, nous remarquons qu'il est possible de jouer sur la taille des fenêtres d'analyse afin d'obtenir une résolution temporelle (et donc fréquentielle) adaptée à la fréquence d'analyse. Nous étudions donc dans la suite les mécanismes de l'audition afin de construire une représentation temps-fréquence qui soit mieux adaptée à l'oreille que ne le sont la TFCT et la TO. Le formalisme mathématique associé à ce type de représentation peut être trouvé par exemple dans [Tor91].

Filtrage auditif

Le système auditif réalise une analyse spectrale des stimuli audibles. La membrane basilaire contenue dans la cochlée (organe de l'audition qui convertit le stimulus acoustique en flux électrique) agit comme un banc de filtres dont les sorties sont ordonnées tonotopiquement, de sorte qu'il y a correspondance entre la position spatiale sur la membrane basilaire et la fréquence [Tra90].

Ce processus, couplé à un processus temporel effectif entre 0,5 et 5 kHz, permet de distinguer des fréquences très proches [Dem89]. La **discrimination fréquentielle**, définie comme l'aptitude qu'a l'oreille de distinguer deux fréquences présentées l'une après l'autre, semble être un bon point de départ pour établir les contraintes de notre représentation temps-fréquence.

Des études psychoacoustiques montrent que cette discrimination est fonction de la fréquence et du niveau [WJG77]. Le JND ("Just Noticeable Difference" ou plus petite différence audible) entre 2 sons purs est plus faible à basses fréquences, plus élevé à hautes fréquences, et semble être à peu près constant pour un niveau donné sur une échelle à bandes de fréquences relatives ($\Delta f/f$). Par exemple, à un niveau de 80 dB, on peut distinguer 2 sons purs autour de 1000 Hz

séparés de seulement 1 Hz, alors que de 2 sons purs autour de 4000 Hz doivent être séparés de 10 Hz pour être discriminés.

Il semble donc qu'une analyse en TO permet de refléter la discrimination fréquentielle réalisée par l'oreille.

Cependant, dans une telle représentation, une résolution fréquentielle acceptable (inférieure à 20 Hz) à 500 Hz induit une résolution temporelle inacceptable pour les basses fréquences (fenêtres temporelles de 500 ms à 50 Hz). Il est donc nécessaire de ne pas trop augmenter la résolution fréquentielle à basse fréquence de manière à conserver une résolution temporelle suffisante pour les transitoires basses fréquences.

De plus, en dessous de 500 Hz, les valeurs de discrimination fréquentielle diffèrent selon les auteurs [WJG77], et les variations inter-individus sont importantes [Can91].

Enfin, les tests de discrimination fréquentielle concernent des sons purs présentés successivement. Or, nous sommes attachés à différencier deux sons purs présentés simultanément : nous allons donc nous intéresser aux capacités de **résolution fréquentielle** de l'oreille plutôt qu'à ses capacités de discrimination fréquentielle.

La perception de deux sons purs de fréquences f_1 et f_2 présentés simultanément dépend principalement de la différence entre leurs fréquences ($\delta f = |f_1 - f_2|$) : si δf est grand, on les perçoit distinctement, mais si δf est petit, on les perçoit comme une seule composante de fréquence $\frac{f_1+f_2}{2}$ modulée en amplitude à la fréquence $\frac{|f_1-f_2|}{2}$. En d'autres termes, la perception auditive interprète la formule suivante selon le premier membre pour δf grand (2 sinusoïdes distinctes), et selon le second membre pour δf petit (1 sinusoïde modulée) :

$$\sin(f_1 t) + \sin(f_2 t) = \cos\left(\frac{|f_1 - f_2|}{2} t\right) \sin\left(\frac{f_1 + f_2}{2} t\right)$$

Entre ces deux modes extrêmes de perception, il existe une zone pour laquelle une seule composante est perçue, mais modulée tellement rapidement que le son devient rugueux. La valeur de l'écart fréquentiel, pour laquelle la rugosité disparaît et la consonnance apparaît, correspond à peu près à la largeur d'une bande critique [Pre98].

La notion de bande critique (ou bande de Bark) fut établie dans les années 1940 lors d'expériences sur le fonctionnement auditif comme un banc de filtres [Fle40]. Les bandes critiques, au nombre de 24, couvrent quasiment l'intégralité du spectre audible (20 Hz à 20 kHz). Elles jouent un rôle dans de nombreux aspects de la perception, dont celui de la résolution fréquentielle. Dans l'échelle associée, appelée échelle des Barks, la largeur des bandes critiques est constante. Cette échelle est linéaire en fréquences jusqu'à 500 Hz ($\Delta f = 100$ Hz), et proportionnelle à la fréquence au-delà ($\Delta f/f = 0,2$, soit approximativement "tiers d'octave"). Elle correspond donc au type d'analyse réalisé par la TFCT à basses fréquences, et au type d'analyse réalisé par la TO à hautes fréquences.

On en conclut qu'une analyse de type $\Delta f = A$ à basse fréquence et $\Delta f/f = B$ à haute fréquence semble être mieux adaptée qu'une analyse uniquement par TFCT ou par TO. Les constantes A et B doivent être ajustées pour effectuer le meilleur compromis possible entre résolution fréquentielle et résolution temporelle pour une fréquence donnée.

Analyse

Dans le type d'analyse retenu, la résolution (temporelle et fréquentielle) varie avec la fréquence. Il est donc nécessaire de faire varier la taille H_Ω de la fenêtre d'analyse $h_\Omega(t)$ en

fonction de la fréquence Ω . Pour cela, on remarque que la seule interprétation adaptée à un changement de fenêtre d'analyse selon la fréquence est la représentation en banc de filtres passe-bande (voir section 2.3.1), donnée par l'équation discrète suivante :

$$S_k(p) = (s * h_k)(p)$$

On interprète alors la représentation comme un filtrage du signal original $s(n)$ par une famille de filtres à valeurs complexes $h_k(n)$ avec $k \in [1, N_f]$.

Une fois la largeur de bande déterminée à basses et hautes fréquences, le nombre de filtres utilisés N_f dépend du recouvrement désiré entre les filtres. Plus le recouvrement est important, plus la redondance d'information est importante. À l'opposé, un recouvrement insuffisant provoque une perte d'information à certaines fréquences. On mesure cette perte d'information par l'amplitude de la modulation lors de la sommation des filtres en fréquence (une fois la normalisation effectuée¹, si cette somme est constante, toute l'information est conservée). Un bon compromis consiste à effectuer le recouvrement le plus faible possible tout en conservant la modulation inaudible, ce qui mène à un nombre minimal de filtres sans perte audible d'information. Typiquement, une excursion maximale de la modulation de 1 dB est tolérée et considérée comme inaudible.

L'utilisation d'un filtre analytique (dont la réponse en fréquence est nulle pour les fréquences négatives) permet d'exprimer le signal réel filtré $\tilde{s}_k(n)$ comme la partie réelle du signal complexe $s_k(n)$ exprimé sous la forme suivante :

$$s_k(n) = M_k(n)e^{j\varphi_k(n)}$$

où $M_k(n)$ et $\varphi_k(n)$ représentent le module et la phase du signal complexe de la k^{ieme} sous-bande. Le signal $s_k(n)$ est un signal analytique, c'est-à-dire que sa partie imaginaire correspond à la transformée de Hilbert de sa partie réelle.

Il existe plusieurs méthodes pour obtenir le signal analytique $s_k(n)$.

On peut convoluer directement dans le domaine temporel le signal original $s(n)$ par le filtre analytique (donc complexe) $h_k(n)$. On peut également effectuer ce filtrage dans le domaine fréquentiel, par la multiplication de la transformée de Fourier du signal original $\hat{s}(\omega)$ par la transformée de Fourier du filtre analytique $\hat{h}_k(\omega)$. Ces deux techniques de filtrage donnent théoriquement les mêmes résultats, et nous illustrons dans la suite chacune d'elles.

Transformation

Les transformations appliquées aux représentations temps-fréquence obtenues sont semblables à celles étudiées dans le chapitre précédent.

Pour la dilatation-p, il s'agit de modifier l'évolution temporelle (à travers le module $M_k(n)$ du signal analytique de la k^{ieme} sous-bande) des sinusoides susceptibles d'être présentes dans chacun des canaux, avec une modification adéquate des phases. Cette transformation peut s'écrire de la manière suivante pour chacune des sous-bandes :

$$s'_k(n) = M_k(n/\alpha)e^{j\alpha\varphi_k(n/\alpha)}$$

1. La normalisation consiste à multiplier chaque fenêtre à $\Delta f = C^{te}$ par $1/K_g = 1/\int |\hat{h}(\omega)|d\omega$, et chaque fenêtre à $\Delta f/f = C^{te}$ par $1/K_o = 1/\int |\hat{h}(\omega)|d\omega/\omega$

La multiplication de la phase instantanée par α est justifiée par le fait que le changement de variable n/α entraîne une modification de la fréquence instantanée, que l'on doit compenser pour conserver la fréquence originale.

D'autre part, travaillant sur des signaux échantillonnés, il paraît nécessaire de rééchantillonner convenablement les deux fonctions $M_k(n)$ et $\varphi_k(n)$ afin de conserver un taux de dilatation constant. Or, comme nous l'avons vu dans les méthodes temporelles, cette contrainte est à appliquer uniquement pour rester dans les limites de la régularité rythmique, c'est-à-dire qu'il semble possible d'interpoler ces fonctions "par morceaux", comme nous le verrons dans la suite.

Pour la transposition-p, il s'agit de multiplier la phase instantanée par α , ce qui revient à appliquer l'évolution temporelle de la sinusoïde du canal k à une sinusoïde dont la fréquence est α fois plus élevée. Cette transformation peut s'écrire de la manière suivante pour chacune des sous-bandes :

$$s'_k(n) = M_k(n)e^{j\alpha\varphi_k(n)}$$

La multiplication de la phase instantanée par α trouve sa justification dans le fait qu'on multiplie de la sorte la fréquence instantanée par α . Des précautions sont à prendre car la phase n'est obtenue que modulo 2π .

Le k^{ieme} canal devient le canal dans lequel est présent une sinusoïde proche de $\alpha\Omega_k$.

Nous illustrons dans la suite chacune de ces transformation-p.

Synthèse

Le signal modifié $s'(n)$ peut être obtenu par une sommation simple des signaux analytiques $s'_k(n)$ issus de chaque sous-bande :

$$s'(n) = \sum_{k=1}^{N_f} s'_k(n)$$

Cependant, cette méthode est valable uniquement lorsque la transformation n'introduit pas de fréquence située en dehors du filtre d'analyse.

Dans le cas contraire, il est nécessaire de filtrer chacun des signaux analytiques temporels par le même filtre que celui utilisé lors de l'analyse, afin d'éliminer toute fréquence indésirable :

$$s'(n) = \sum_{k=1}^{N_f} (s'_k * h_k)(n)$$

Ceci revient à prendre en compte l'existence d'un noyau reproduisant.

Nous illustrons dans la suite ces deux méthodes de synthèse.

3.2.1 Transposition-p par une méthode temps-fréquence

Le principe de cette méthode consiste à utiliser la représentation temps-fréquence adaptée à l'audition telle que nous l'avons définie ci-dessus. Cette représentation nous fournit une famille de signaux temporels analytiques correspondant aux sous-bandes fréquentielles. Les phases instantanées déroulées de chacune de ces sous-bandes sont multipliées par le taux de dilatation α , revenant ainsi à multiplier la fréquence instantanée, et les parties réelles de ces signaux modifiés sont simplement sommées pour obtenir le signal transposé [PBD⁺99].

Nous détaillons dans la suite chacune des étapes de la méthode.

Analyse

Nous effectuons ici la décomposition en sous-bandes dans le domaine fréquentiel, en appliquant des fenêtres fréquentielles sur la transformée de Fourier du signal temporel analytique. Nous aurions pu obtenir le signal analytique par l'utilisation d'une convolution par tranche, mais nous préférons dans un premier temps simplifier l'implantation en réalisant une seule transformée de Fourier sur la globalité du signal. Cette décision nous limite cependant à des durées de sons de l'ordre de la dizaine de secondes, à la fois pour des raisons d'espace mémoire et de durée des calculs. De plus, nous négligeons par cette technique les défauts engendrés par la convolution circulaire au début et à la fin du signal.

Construction des filtres

L'axe fréquentiel est tout d'abord divisé en 2 parties, hautes et basses fréquences, comme le montre le schéma 3.11).

La première s'étend de $f_{min} = 20$ Hz à $f_{seuil} = 500$ Hz et correspond à la portion où l'analyse est effectuée à bande de fréquence constante (type TFCT).

La seconde s'étend de $f_{seuil} = 500$ Hz à $f_{max} = \frac{\beta F_e/2}{\alpha}$ et correspond à la portion où l'analyse est effectuée à bande de fréquence relative constante (type TO). Le terme $\beta = 0,9$ permet d'éviter que l'ondelette la plus haute fréquence ne dépasse trop largement la fréquence d'échantillonnage et ne provoque de repliement spectral.

Ensuite, il est nécessaire de fixer les largeurs de bande de fréquence en-dessous et au-dessus de f_{seuil} .

Une résolution fréquentielle de 20 Hz semble être suffisante pour les sons complexes à basses fréquences comme on a pu le voir avec la TFCT (les harmoniques d'un son monophonique de fréquence fondamentale audible sont séparées de plus de 20 Hz). Nous retenons donc $\Delta f = 20$ Hz comme largeur de bande à basses fréquences.

En supposant que le filtre centré sur 500 Hz soit à la fois considéré comme faisant partie de la TFCT ($\Delta f = 20$ Hz) et de la TO, il en résulte que la largeur des bandes relatives pour la TO est de valeur $\Delta f/f = 20/500 = 0,04$ Hz (valeur proche de l'analyse en tiers d'octave). Nous prenons donc $\Delta f/f = 0,04$ comme largeur de bande relative à hautes fréquences.

La figure 3.11 indique les largeurs des bandes fréquentielles pour l'échelle des Barks et pour les filtres proposés.

Une fois la fréquence de seuil (entre répartition type TFCT et type TO) et les largeurs de bande fixées, il est nécessaire de déterminer le nombre de filtres à répartir sur la bande passante. Ce nombre minimum nous est dicté par l'excursion maximale de la modulation de la somme des fenêtres en fréquence, valeur fixée à 1 dB.

En choisissant d'utiliser des fenêtres gaussiennes, le critère précédent nous amène à considérer un banc de $N_f = 96$ filtres : 24 filtres $\hat{h}(\omega)$ de largeur de bande 20 Hz entre 20 et 500 Hz, et 72 filtres de largeur de bande $0,04f$ Hz au-delà de 500 Hz.

Les valeurs numériques proposées ($f_{seuil} = 500$ Hz, $\Delta f = 20$ Hz, $\Delta f/f = 0,04$, $N_f = 96$) sont des points de départ pour un ajustement plus précis qui doit être réalisé à l'oreille.

Filtrage

Tous ces filtres sont appliqués successivement à la transformée de Fourier $\hat{s}(\omega)$ du signal original par une simple multiplication des deux spectres. Afin d'obtenir un signal analytique, nous ne prenons pas en compte les fréquences négatives, qu'on suppose alors nulles. Le filtrage

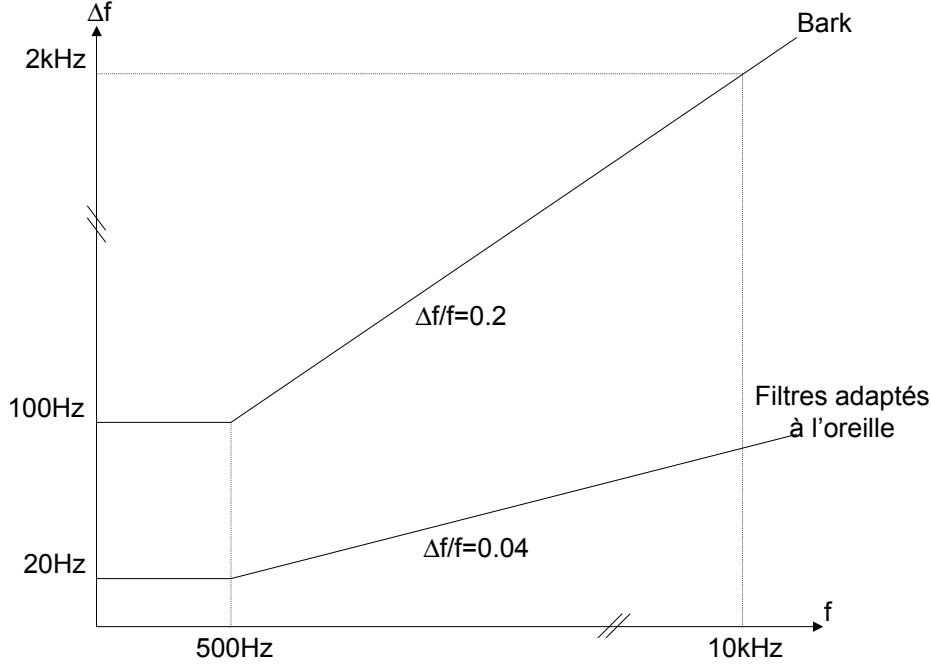


Figure 3.11 – Largeur des bandes fréquentielles en fonction de la fréquence basée sur l'échelle des Barks

est donc donné par l'équation suivante :

$$\begin{aligned}\hat{s}_k(\omega) &= \hat{s}(\omega)\hat{h}_k(\omega) & \forall \omega \geq 0 \\ \hat{s}_k(\omega) &= 0 & \forall \omega < 0\end{aligned}$$

La transformée de Fourier inverse fournit le signal complexe analytique $s_k(n)$ qui est le signal temporel issu du filtrage par la k^{ieme} sous-bande. Son écriture sous la forme $s_k(n) = M_k(n)e^{j\varphi(n)}$ met en évidence l'amplitude et la phase instantanée de ce signal.

Transformation

La transformation consiste simplement à multiplier la phase instantanée $\varphi(n)$ de chacun des signaux $s_k(n)$ par le facteur de transposition α :

$$s'_k(n) = M_k(n)e^{j\alpha\varphi_k(n)}$$

Cependant, puisque la phase est définie modulo 2π , il est nécessaire d'appliquer auparavant un algorithme de déroulement de phase², qui permet de rendre continue l'évolution de phase.

On peut également interpréter le déroulement de phase comme un moyen d'estimer sans discontinuité la fréquence instantanée $\nu_k(n)$, qui est la dérivée de la phase instantanée :

$$\nu_k(n) = \frac{1}{2\pi} \frac{\partial \varphi_k}{\partial n}$$

2. L'algorithme de déroulement de phase consiste à additionner $m2\pi$ ($m \in \mathbb{N}$) à la valeur de la phase à l'instant n lorsque la différence de phase avec l'instant $n - 1$ est supérieure à π . On choisit m de sorte que la variation de phase au cours du temps n'excède jamais π .

Synthèse

Les signaux $s'_k(n)$ étant analytiques, leurs parties réelles sont les signaux réels qui nous intéressent. Le signal transposé $s'(n)$ est donc simplement la somme des parties réelles des signaux modifiés des sous-bandes fréquentielles :

$$s'(n) = \sum_{k=1}^{N_f} \text{Re}[s'_k(n)]$$

où Re indique la partie réelle.

On remarque que des défauts peuvent surgir si la transformation crée au sein d'une sous-bande des fréquences qui ne sont pas sensées apparaître dans cette sous-bande.

Des conclusions sur cette méthode sont données en section 3.2.3.

3.2.2 Dilatation-p par une méthode temps-fréquence

Le principe de cette méthode consiste, comme précédemment, à utiliser la représentation temps-fréquence adaptée à l'audition telle que nous l'avons définie auparavant. Cette représentation nous fournit une famille de signaux temporels analytiques représentant les sous-bandes fréquentielles. Les amplitudes et fréquences instantanées de ces signaux sont des fonctions du temps qu'il s'agit de dilater. Cette dilatation peut être uniforme (rééchantillonnage des fonctions), mais nous préférons interpoler ces fonctions de synthèse (module et phase) selon le contenu du signal. Le signal dilaté est finalement obtenu par sommation des signaux analytiques, mais cette fois en filtrant chacun d'eux après la transformation, ce qui satisfait la contrainte du noyau reproduisant.

Nous détaillons dans la suite chacune des étapes de la méthode.

Analyse

Dans cette méthode, le filtrage est réalisé par la convolution dans le domaine temporel du signal original $s(n)$ par la famille des filtres complexes analytiques $h_k(n)$.

Ces filtres peuvent être définis dans le domaine fréquentiel, donc donnés par leur réponse fréquentielle $\hat{h}_k(\omega)$. On en déduit alors leur réponse impulsionnelle $h_k(n)$ par une simple transformée de Fourier inverse.

Nous choisissons cette fois de les définir directement dans le domaine temporel.

Construction des filtres

L'ajustement des paramètres des filtres qui donne de bons résultats auditifs mène à des valeurs différentes de la méthode précédente :

$N_f = 48$ filtres : 12 filtres $\hat{h}(\omega)$ de largeur de bande 40 Hz entre 20 et 500 Hz, et 36 filtres de largeur de bande $0,1f$ Hz au-delà de 500 Hz.

Filtrage

Le filtrage du signal original par la famille de filtres analytique fournit les signaux :

$$s_k(n) = M_k(n)e^{j\varphi_k(n)}$$

La fréquence instantanée, plus souple pour la manipulation, est déduite de la phase instantanée par dérivation de celle-ci :

$$\nu_k(n) = \frac{1}{2\pi} \frac{\partial \varphi_k}{\partial n}$$

Il est nécessaire d'appliquer un algorithme de déroulement de phase avant de calculer la fréquence instantanée.

En pratique, le filtrage est réalisé avec la fonction Matlab [Mat03] "filtfilt" dont la caractéristique est d'agir uniquement sur le spectre d'amplitude du signal, et aucunement sur le spectre de phase.

Transformation

Le principe de la transformation consiste à allonger ou contracter les fonctions de synthèse $M_k(n)$ et $\nu_k(n)$. Pour cela, nous pouvons utiliser le rééchantillonnage (dilatation homogène le long du signal) mais nous préférons cependant réaliser une dilatation variable, afin de sélectionner les parties à dilater en vue d'éviter de modifier les transitoires.

La technique utilisée consiste à repérer à l'intérieur d'un bloc de durée $B = 5$ ms, les $N_e = B(1 - \alpha)$ échantillons pour lesquels le module et la phase varient le moins. Un critère simple est utilisé pour cela. Il s'agit d'une minimisation du produit des valeurs de la variation du module instantané et de la variation de la fréquence instantanée. Les indexes des N_e plus petites valeurs définissent les candidats aux insertions, et correspondent a priori aux instants les plus stationnaires du bloc.

On insère ensuite, après chacun des candidats, un nouveau point dans le signal analytique, dont les valeurs du module et de la fréquence instantanée sont obtenues par interpolation linéaire des valeurs de ses voisins.

On effectue ainsi une sorte de rééchantillonnage variable des fonctions $M_k(n)$ et $\nu_k(n)$ menant aux fonctions modifiées $M'_k(n)$ et $\nu'_k(n)$ dont le support temporel est multiplié par α . La phase instantanée modifiée est obtenue par intégration temporelle des valeurs de fréquence instantanée :

$$\varphi'_k(n) = \sum_{i=0}^n \nu'_k(i)$$

Le signal de la k^{ieme} sous-bande est donc donné par :

$$s'_k(n) = M'_k(n) e^{j\varphi'_k(n)}$$

Synthèse

La synthèse du signal modifié est obtenue par sommation des différentes sous-bandes. Par contre, nous améliorons le résultat sonore par l'application d'un filtrage équivalent à celui de l'analyse, afin de supprimer toutes les fréquences qui auraient pu apparaître lors de la modification du signal analytique :

$$s'(n) = \sum_{k=1}^{N_f} Re[(s'_k * h_k)(n)]$$

où Re indique la partie réelle.

Des conclusions sur cette méthode sont données en section 3.2.3.

3.2.3 Discussion sur les méthodes adaptées à l'audition

Les méthodes de transposition-p et de dilatation-p présentées offrent un compromis entre résolution temporelle et résolution fréquentielle qui est adapté à l'oreille.

La résolution fréquentielle est élevée et constante à basses fréquences, au détriment d'une bonne précision temporelle toutefois superflue dans cette gamme de fréquences. Elle permet ainsi de modifier des signaux complexes car deux sinusoides proches peuvent être résolues et traitées correctement.

La résolution temporelle est élevée à hautes fréquences, au détriment d'une bonne résolution fréquentielle. Elle permet ainsi de modifier des signaux transitoires car même si les relations de phase entre canaux sont perdues, l'énergie d'un canal ne se répartit pas plus que sur la durée d'une fenêtre d'analyse.

Pour illustrer ces méthodes, nous proposons d'écouter le son complexe d'orchestre traité par la méthode de transposition-p puis rééchantillonné aboutissant ainsi à un son dilaté (son [56]), ainsi que le son traité par la méthode de dilatation-p (son [57]) et de les comparer au son original (son [88]) et au son obtenu par une méthode fréquentielle aveugle (son [44]).

De même, nous proposons d'écouter le son impulsif de castagnettes traité par la méthode de transposition-p puis rééchantillonné (son [58]), ainsi que le son traité par la méthode de dilatation-p (son [59]) et de les comparer au son original (son [11]) et au son obtenu par une méthode fréquentielle aveugle (son [45]).

La méthode de transposition-p nous montre la nette amélioration des résultats sonores à la fois sur les sons complexes et sur les sons impulsifs, et ce, uniquement grâce au type de représentation temps-fréquence utilisé. Le reste de la méthode reste en effet identique à la méthode aveugle par TFCT. On en déduit donc que d'autres progrès sont possibles, inspirées des améliorations effectuées sur le vocodeur de phase (notamment la technique de verrouillage de phase).

Le perfectionnement proposé dans la méthode de dilatation-p est un équivalent dans l'interprétation en banc de filtres des méthodes à taux de dilatation variable discutées en section 2.3.4. Elle est cependant ici présentée dans un contexte de représentation temps-fréquence adaptée à l'audition.

Les résultats obtenus peuvent sans doute être améliorés par l'ajustement des paramètres, notamment le nombre de filtres N_f , la fréquence seuil f_{seuil} , et les constantes de largeur de bande basse fréquence Δf et haute fréquence $\frac{\Delta f}{f}$. Cependant, les tests comparatifs auditifs, réalisés au moment du choix de l'algorithme à implanter, favorisaient la méthode temporelle pour $\alpha = 1,04\%$.

3.3 Dilatation-p par méthodes couplées

Les méthodes proposées ici reposent sur le principe suivant : on décompose le signal original en un certain nombre de signaux temporels, et l'on traite chacun de ces composants par la méthode la plus adaptée.

Nous présentons deux exemples de décomposition (décomposition en 2 sous-bandes fréquentielles et décomposition hybride) dont les signaux résultant sont modifiés soit par une méthode temporelle, soit par une méthode fréquentielle.

3.3.1 Décomposition en sous bandes

L'idée de cette méthode repose sur les remarques suivantes :

- Les basses fréquences souffrent généralement d'artefacts lorsqu'elles sont modifiées par des méthodes temporelles. Dans ce cas, il est possible d'augmenter la durée des segments insérés mais des problèmes d'anisochronie peuvent apparaître lorsque ces segments dépassent 6 ms. Les méthodes fréquentielles sont mieux adaptées pour de basses fréquences.
- Les hautes fréquences contenant des transitoires souffrent d'étalement (d'autant plus audible que les fenêtres d'analyse sont longues) lors de la modification par méthodes fréquentielles. Les méthodes temporelles avec de courts segments insérés sont mieux adaptées pour de hautes fréquences.

Il semble naturel dès lors de séparer hautes et basses fréquences et d'adapter le traitement à chacun de ces signaux : méthode temporelle pour la partie hautes fréquences, et méthode fréquentielle pour la partie basses fréquences, tel qu'illustré en figure 3.12.

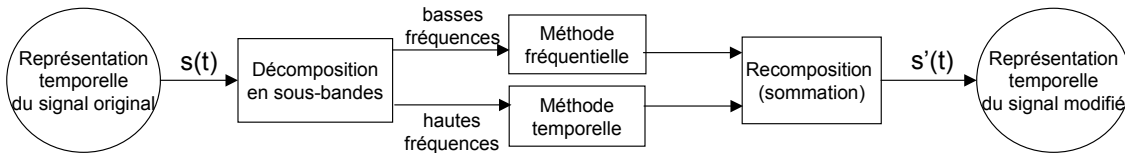


Figure 3.12 – Illustration de la méthode couplée par décomposition en sous-bandes

Décomposition temporelle

Pour cela, nous réalisons en premier lieu un filtrage du signal original. Nous utilisons 2 filtres complémentaires avec une fréquence de coupure notée f_{seuil} . Ce filtrage peut être effectué dans le domaine temporel par une simple convolution, mais il nous est plus direct d'utiliser le vocodeur de phase et le détourner de sa fonction de dilatation-p pour en faire un filtre, même si nous savons que ce n'est pas l'outil optimal pour cette application.

L'analyse est réalisée avec des fenêtres de Hanning de 2048 points ($F_e = 44100$ Hz), et les pas d'analyse et de synthèse sont fixés à 512 points. La fenêtre de filtrage passe-bas $\hat{W}_{PB}(\omega)$ est construite de sorte qu'elle vaut 1 pour les fréquences jusqu'à la fréquence f_{seuil} et 0 au-delà. La fenêtre de filtrage passe-haut est déduite de la précédente : $\hat{W}_{PH}(\omega) = 1 - \hat{W}_{PB}(\omega)$. De cette manière, on est assuré que la somme des deux signaux filtrés reconstruisent exactement le signal original.

Les signaux filtrés $s_{PB}(n)$ et $s_{PH}(n)$ sont synthétisés à partir des spectres à court terme multipliés par ces gabarits de filtres définis en fréquence.

Les sous-bandes basses et hautes fréquences sont obtenues ici par un filtrage pour lequel $f_{seuil} = 500$ Hz.

Transformation des signaux temporels

La dilatation-p de la sous-bande basses fréquences $s_{PB}(n)$ est réalisée par une méthode fréquentielle avec une fenêtre d'analyse $N_{FFT} = 2048$ points et un pas d'analyse $I_a = 512$ points. Nous utilisons ici le verrouillage de phase strict ("Rigid phase locking") proposé par Laroche et Dolson [LD99a].

La dilatation-p de la sous-bande hautes fréquences est réalisée par une méthode temporelle pour laquelle la durée des segments insérés est limitée à $1/f_{seuil} = 2$ ms. Cette valeur a été retenue car elle correspond à la fréquence 500 Hz, qui est censée être la fréquence la plus basse présente dans cette sous-bande. De plus, puisque nous désirons éviter les défauts audibles d'anisochronie, nous devons insérer des segments inférieurs à 6 ms.

La dilatation-p globale du signal est obtenue par simple sommation des sous-bandes modifiées.

Résultats de la méthode

Nous illustrons cette méthode par une dilatation-p de facteur $\alpha = 4,2\%$ d'un son contenant basse, batterie et guitare. Nous entendons successivement dans la plage [60] le signal original, la sous bande basses-fréquences suivie de sa version dilatée, la sous bande hautes-fréquences suivie de sa version dilatée, et enfin la somme des deux signaux dilatés.

Le résultat sonore de la dilatation-p semble tout à fait correct pour la sous-bande basses fréquences, par contre on détecte de la rugosité dans la sous-bande hautes fréquences. En effet, bien que le filtrage soit abrupte, il subsiste quand même des fréquences inférieures à 500 Hz à faible niveau, et les défauts de désynchronisation résultants deviennent audibles.

Les discontinuités de désynchronisation de la sous-bande haute fréquence peuvent être atténuées en permettant l'insertion de segments plus longs que la période fondamentale correspondant à la fréquence de coupure. Ainsi, on gagne en qualité pour les sons complexes au voisinage de la fréquence de coupure, mais par contre, des problèmes d'anisochronie risquent de surgir. De plus, le recouvrement inévitable entre les sous-bandes fréquentielles est à l'origine de défauts de phase, comme on peut l'entendre dans l'exemple sonore [61] où les segments insérés peuvent être jusqu'à 5 fois plus long qu'auparavant.

En effet, pour un signal quasi-stationnaire dont la fréquence varie lentement autour de la fréquence de coupure f_{seuil} , le déphasage entre les sous-bandes (dû à une estimation imprécise de la fréquence pour la méthode fréquentielle et dû à l'insertion de segments générant un taux de dilatation localement variable pour la méthode temporelle) entraîne des interférences entre les composantes de ces sous-bandes. Il en résulte une modulation d'amplitude parfois très importante.

Ce phénomène est observé pour l'exemple sonore [62], visualisé en figure 3.13.

Les sons impulsifs de type percussions donnent des résultats très satisfaisants comme le prouvent les exemples sonores [63] ($k_{max} = 2$ ms) et [64] ($k_{max} = 10$ ms).

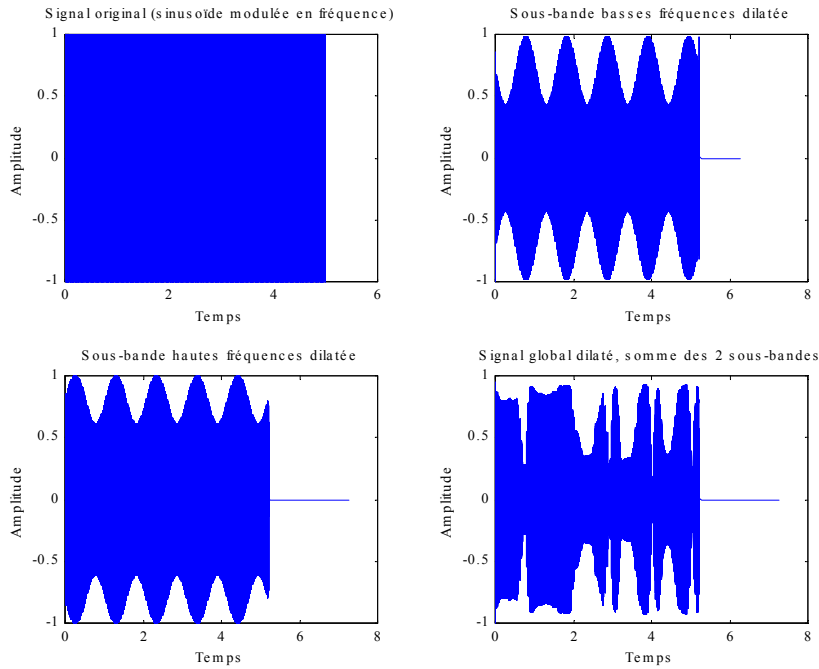


Figure 3.13 – *Dilatation d’une sinusoïde modulée en fréquence par une méthode couplée*

Cette technique de couplage des méthodes pourrait être étudiée plus en détail, notamment pour trouver un meilleur ajustement des paramètres tels que f_{seuil} , la largeur Δf des bandes de fréquence de la méthode fréquentielle et la durée maximale k_{max} du segment inséré de la méthode temporelle. Cependant, les relations de phases autour de la fréquence de coupure restent des problèmes provoquant des artefacts audibles.

3.3.2 Décomposition hybride

L’idée de cette méthode repose sur les remarques suivantes :

- Les transitoires souffrent d’étalement par les méthodes fréquentielles. Il s’agit là du principal défaut de ces méthodes lorsque les fenêtres d’analyse sont longues. Ils sont cependant bien respectés par l’utilisation de méthodes temporelles lorsque les segments insérés sont assez courts (pas de redoublement).
- Le signal qui n’est pas transitoire peut être assimilé soit à des composantes sinusoïdales, soit à du bruit. Ces types de signaux se comportent relativement bien quand ils sont traités par des méthodes fréquentielles avec des fenêtres d’analyse assez longues.

Il semble dès lors naturel de séparer transitoire et résidu (ce qui n’est pas considéré comme transitoire) et d’adapter le traitement à chacun de ces signaux : méthode temporelle pour la partie transitoire, et méthode fréquentielle pour la partie résidu, tel qu’illustré en figure 3.14.

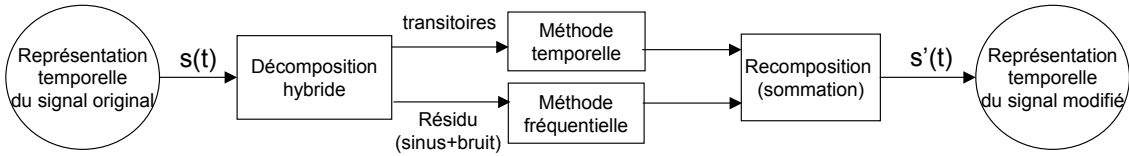


Figure 3.14 – Illustration de la méthode couplée par décomposition hybride

Décomposition temporelle

Pour cela, nous réalisons en premier lieu une décomposition temporelle hybride³ du signal original. De nombreux auteurs proposent différentes façons de détecter et d'extraire les transitoires d'un signal [Ser89, BN93, Mas96, VM98, Lev98, Gri99, Dau00, Ros00, DDS01]. Nous utilisons ici la technique proposée par Rodet et Jalliet [RJ01], basée sur la synchronisation temporelle d'énergie entre sous-bandes fréquentielles dans une représentation temps-fréquence obtenue par TFCT.

Le principe consiste à observer l'énergie de chacune des sous-bandes fréquentielles (à travers le module au carré de la TFCT) et y repérer une variation brutale, qui correspond alors à un candidat dans cette sous-bande. La synchronisation de ces candidats (avec un intervalle d'incertitude) à travers les différentes sous-bandes permet d'extraire un transitoire lorsque la somme des amplitudes des pics d'un même agrégat dépasse un seuil fixé.

On retient les valeurs complexes du plan temps-fréquence correspondant aux transitoires détectés et on annule toutes les autres valeurs. La méthode de reconstruction est celle proposée par Griffin et Lim [GL84]. On obtient de la sorte un signal temporel $s_t(n)$, composé uniquement de transitoires (ou plutôt ce que l'on a défini comme tels). Ce signal est ensuite soustrait à l'original pour donner le signal résiduel $s_r(n)$, a priori constitué des composantes stationnaires du signal (sinusoïdes et bruit).

Transformation des signaux temporels

La dilatation-p du signal transitoire $s_t(n)$ est réalisée par une méthode temporelle pour laquelle la durée des segments insérés est comprise entre 0,25 et 1 ms. Ces valeurs correspondent à une insertion correcte pour des fréquences minimales de 1000 Hz. Les fréquences inférieures peuvent souffrir d'artefacts, mais elles sont généralement masquées temporellement par l'énergie des transitoires.

La dilatation-p du signal résiduel $s_r(n)$ est réalisée par une méthode fréquentielle avec une fenêtre d'analyse $N_{FFT} = 2048$ points et un pas d'analyse $I_a = 512$ points. Nous utilisons ici également le verrouillage de phase strict ("Rigid phase locking") proposé par Laroche et Dolson [LD99a].

La dilatation-p globale du signal est obtenue par simple sommation des sous-bandes modifiées.

3. La bande passante des signaux issus d'une décomposition hybride n'est généralement pas limitée, contrairement aux signaux issus d'une décomposition en sous-bandes.

Résultats de la méthode

Nous illustrons cette méthode par une dilatation-p de facteur $\alpha = 4,2\%$ d'un son de percussions. Nous entendons successivement dans l'exemple sonore [65] le signal original, le signal transitoire suivi de sa version dilatée, le signal résiduel suivi de sa version dilatée, et enfin la somme des deux signaux dilatés.

Tout d'abord, nous remarquons que la décomposition fournit un signal transitoire très précis, mais que le signal résiduel possède peut-être encore trop d'attaque.

La dilatation-p du signal transitoire est d'extrêmement bonne qualité. Par contre la dilatation-p du signal résiduel souffre d'un étalement d'attaque. Ce "bavement" est largement atténué après sommation des deux signaux, car l'attaque (perceptivement) originale est déplacée à la position attendue, et masque en partie l'étalement d'énergie.

Cette méthode aurait besoin, d'un ajustement plus précis des paramètres. Par exemple, pour le son étudié, il aurait été nécessaire d'abaisser le seuil de détection du transitoire lors de la décomposition.

3.3.3 Discussion sur les méthodes couplées

Les méthodes présentées permettent d'éviter dans une certaine mesure le compromis inévitable des méthodes purement temporelles (défaut d'anisochronie à cause de segments insérés trop longs "versus" défauts sur les basses fréquences à cause de segments insérés trop courts) et le compromis inévitable des méthodes purement fréquentielles (étalement des attaques à cause des fenêtres temporelles trop longues "versus" coloration des sons complexes à cause des fenêtres temporelles trop courtes).

Les résultats sonores obtenus sont corrects mais encore insuffisants vis-à-vis des contraintes de qualité sonore imposées. Des études plus poussées avec un ajustement plus précis des paramètres n'excluent pas la possibilité d'obtenir des résultats très convaincants, surtout pour des taux de dilatation élevés.

Nous avons présenté deux exemples de méthodes couplées. Il est bien sûr possible d'en créer de nombreuses autres en changeant le type de décomposition utilisé. Par exemple, il est possible d'effectuer une décomposition transitoire-tonal-stochastique [DT02], et de traiter indépendamment ces trois signaux, et même de décomposer à nouveau le signal tonal en sous-bandes fréquentielles. De plus, nous pouvons utiliser en guise de méthodes fréquentielles, les méthodes temps-fréquence adaptées à l'audition proposées précédemment.

3.4 Algorithme de dilatation-p HARMO

Dans la suite de cet exposé, nous nous plaçons dans le cas qui s'est révélé être le plus défavorable, en prenant $\alpha > 1$. En effet, tous les tests informels, ainsi que les remarques des utilisateurs collectées, s'accordent pour conclure que l'élongation est généralement de moins bonne qualité que la contraction dans le cadre des méthodes temporelles. Nous pensons qu'une des raisons est que l'élongation induit parfois une duplication d'un transitoire, extrêmement audible lorsque la durée séparant ces deux événements devient grande. Lors de la contraction, le segment contenant le transitoire est supprimé, ce qui modifie généralement le timbre, mais il subsiste souvent une partie de transitoire (dû au fondu-enchaîné) sauvegardant ainsi le caractère impulsif. D'autres explications peuvent venir étayer cette opinion, notamment psychoacoustiques (effets de masquage) ou linguistiques (à un débit plus élevé, on attache peut-être moins d'importance aux détails des signaux vocaux).

3.4.1 Principe de la méthode

Le principe général de la méthode HARMO est dérivé des méthodes WSOLA ou SOLAFS, elles-mêmes issues de la méthode SOLA (voir sections 2.2.3), et développées par Dattorro [Dat87] et Laroche [Lar93, Lar00].

Il s'agit donc d'une méthode strictement temporelle dont le but est d'insérer en des points d'insertion I des segments de durée K donnée par une mesure de similarité. L'éventuelle discontinuité produite est atténuée grâce à un fondu-enchaîné de durée FE .

Le principe est illustré en figure 3.15 avec les définitions données ci-après.

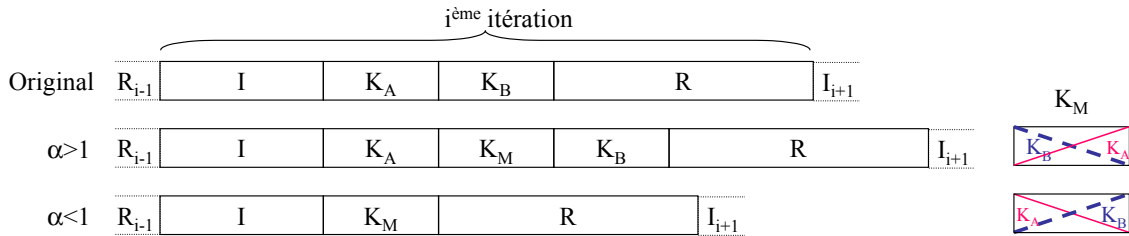


Figure 3.15 – Principe de la méthode HARMO

Définitions

Itération : Ensemble des opérations effectuées entre 2 insertions. Une itération commence par le calcul du point d'insertion I , se poursuit par le calcul de la durée du segment inséré K , la formation du segment mixé K_M , et se termine par le calcul de la durée R du résidu. Chaque début (et donc fin) d'itération se situe sur la droite idéale du diagramme d'entrée/sortie.

K : Durée du segment inséré. Cette valeur est donnée par une mesure de similarité.

K_A, K_B, K_M : Segments originaux (A et B) et issu du mixage (M), tous de durée $k_{min} < K < k_{max}$.

FE : Durée du Fondu-enchaîné. Nous supposons ici que $FE = K$.

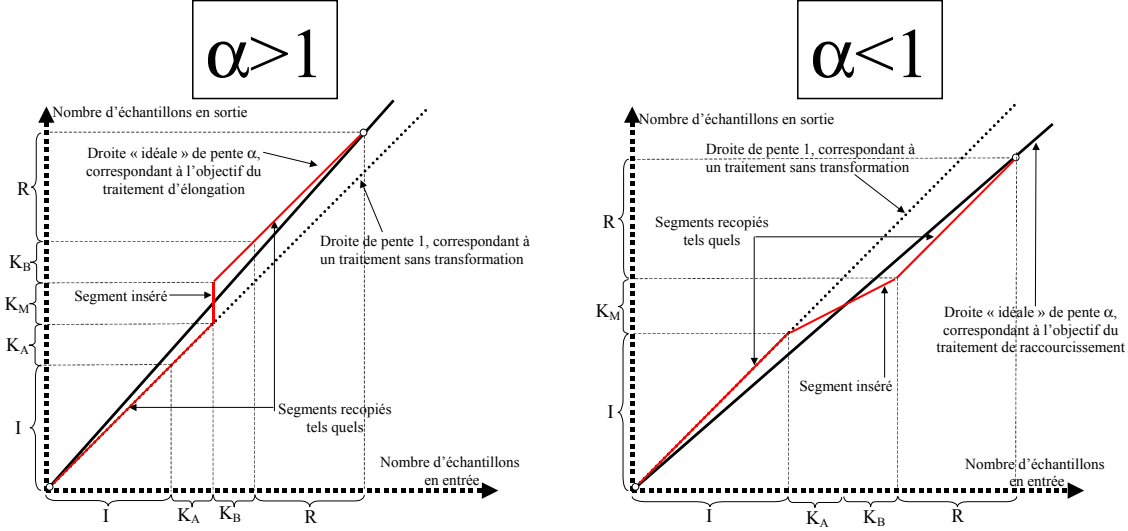
I : Point d'Insertion : Temps écoulé avant l'insertion des segments successifs $[K_A K_M K_B]$ pour $\alpha > 1$ ou du segment $[K_M]$ pour $\alpha < 1$.

R : Résidu : Temps écoulé après l'insertion pour respecter la fonction de dilatation. Cela correspond au nombre d'échantillons à recopier tels quels afin de se replacer sur la droite idéale du diagramme d'entrée/sortie.

Insertion : Opération consistant à insérer le segment K_M entre les segments K_A et K_B pour $\alpha > 1$, ou à la place des segments K_A et K_B pour $\alpha < 1$.

Bloc : Ensemble des segments correspondant à une itération. Ainsi le bloc original correspond à la concaténation des segments $[I K_A K_B R]$, et le bloc dilaté correspond à la concaténation des segments $[I K_A K_M K_B R]$ pour $\alpha > 1$ et $[I K_M R]$ pour $\alpha < 1$.

La figure 3.16 montre le type de diagramme d'entrée/sortie obtenu avec notre méthode.



Le problème peut sembler simple en apparence, mais en pratique le choix des segments K_A et K_B s'avère difficile et crucial. Il s'agit en effet d'adapter leur durée K aux fréquences présentes dans le signal pour éviter les discontinuités de désynchronisation dans le cas de signaux stationnaires, et de les rendre assez courts afin que la duplication d'un événement énergétique localisé soit inaudible dans le cas de signaux transitoires.

Discussion sur la durée R du résidu

Chaque itération respecte le taux de dilatation désiré, c'est-à-dire que le rapport de durée entre le bloc dilaté et le bloc original correspond au taux de dilatation. On en déduit ainsi R à partir de α , K et I :

$$\begin{aligned}
 \alpha(\text{Bloc original}) &= (\text{Bloc dilaté}) \\
 \Leftrightarrow \alpha(I + K_A + K_B + R) &= \begin{cases} I + K_A + K_M + K_B + R & \text{pour } \alpha > 1 \\ I + K_M + R & \text{pour } \alpha < 1 \end{cases} \\
 \Leftrightarrow \alpha(I + 2K + R) &= (I + rK + R) \begin{cases} r = 3 & \text{pour } \alpha > 1 \\ r = 1 & \text{pour } \alpha < 1 \end{cases} \\
 \Leftrightarrow R &= -I + K \frac{r - 2\alpha}{\alpha - 1} \begin{cases} r = 3 & \text{pour } \alpha > 1 \\ r = 1 & \text{pour } \alpha < 1 \end{cases}
 \end{aligned}$$

Nous nous imposons la contrainte que la durée R du résidu soit toujours positive ($R \geq 0$). Pour $\alpha < 1$, cette contrainte nous permet de réaliser tous les calculs nécessaires à une itération

à l'aide du seul signal original et non à l'aide du signal déjà traité, ce qui simplifie les flux de signaux. Bien qu'il eût été possible de fixer cette contrainte à $R \geq -K$ pour $\alpha > 1$ pour la même simplification, nous préférons conserver le même mode de fonctionnement pour tout α .

Cette contrainte, que nous pourrions facilement lever au détriment d'une légère simplification, mène à la limitation suivante pour $I = 0$:

$$\frac{1}{2} < \alpha < \frac{3}{2}$$

Dans notre application, cette contrainte ne nous est pas vraiment restrictive étant donné les taux de dilatation employés ($-20\% < \alpha < +20\%$). Elle pourrait cependant le devenir si l'on décidait de faire varier largement I , mais cela mènerait alors à des problèmes évidents d'anisochronie.

Conclusion sur le principe de la méthode

La qualité de la méthode utilisée repose principalement sur la manière d'extraire les trois paramètres K , I et FE . Il est indispensable de comprendre le rôle de chacun d'eux, puis de mettre au point des critères qui permettent de les sélectionner convenablement.

3.4.2 Développement de l'outil HARMOLAB

Avant d'améliorer l'algorithme temporel de type WSOLA, il est nécessaire de concevoir des critères de choix des paramètres K , I et FE qui mènent à un traitement sans défaut audible. Pour cela, il semble important, d'une part de comprendre les artefacts produits par certains types de sons, et d'autre part de tester s'il est possible d'éliminer ces imperfections.

Nous avons pour cela développé un outil spécifique d'analyse et de tests, appelé HARMOLAB et codé en Matlab [Mat03], permettant de chercher "manuellement" s'il existe des paramètres pour lesquels on n'entend plus l'anomalie. Cette étape nous permet de tirer des conclusions en vue de l'élaboration de critères autorisant une extraction automatique des paramètres. Il s'agit en quelque sorte d'effectuer un ajustement manuel des paramètres, tel que Scott [Sco67] le pratique pour déterminer la position des "pulses" afin de minimiser les discontinuités.

Les principales caractéristiques de cet outil, dont on peut voir une capture d'écran en figure 3.17 sur un son de castagnettes⁴, peuvent se décomposer en deux parties, symbolisées par les moitiés supérieures et inférieures de l'interface du logiciel.

La partie supérieure de l'interface permet une observation d'un signal original (à gauche) et du signal synchronisé correspondant (à droite) traité au préalable par un algorithme codé en Matlab, pour lequel les paramètres (I , K , FE) ont été stockés en mémoire. Cet aspect du logiciel permet de visualiser les défauts entendus dans le signal traité, et de comprendre pourquoi et comment ils sont apparus. En effet, des marques verticales donnent des indications sur le début et la fin des itérations, ainsi que sur les segments K_A , K_B et K_M utilisés. Pour le signal original, on peut voir les 2 segments successifs K_A et K_B qui ont été utilisés pour construire K_M , et sur le signal traité, on trouve K_M entre les segments précédents. Dans ce cas, on peut voir qu'une différence d'énergie entre les segments K_A et K_B donne naissance à un sursaut d'énergie entendu comme un redoublement.

La version d'analyse automatique des paramètres utilisés est affichée, ainsi que les paramètres eux-même dans le cadre "Résultats de l'analyse" (I , K , CF sont donnés en nombre

4. Les deux impulsions rapprochées ne sont pas perçues distinctement, mais donnent sa caractéristique au timbre.

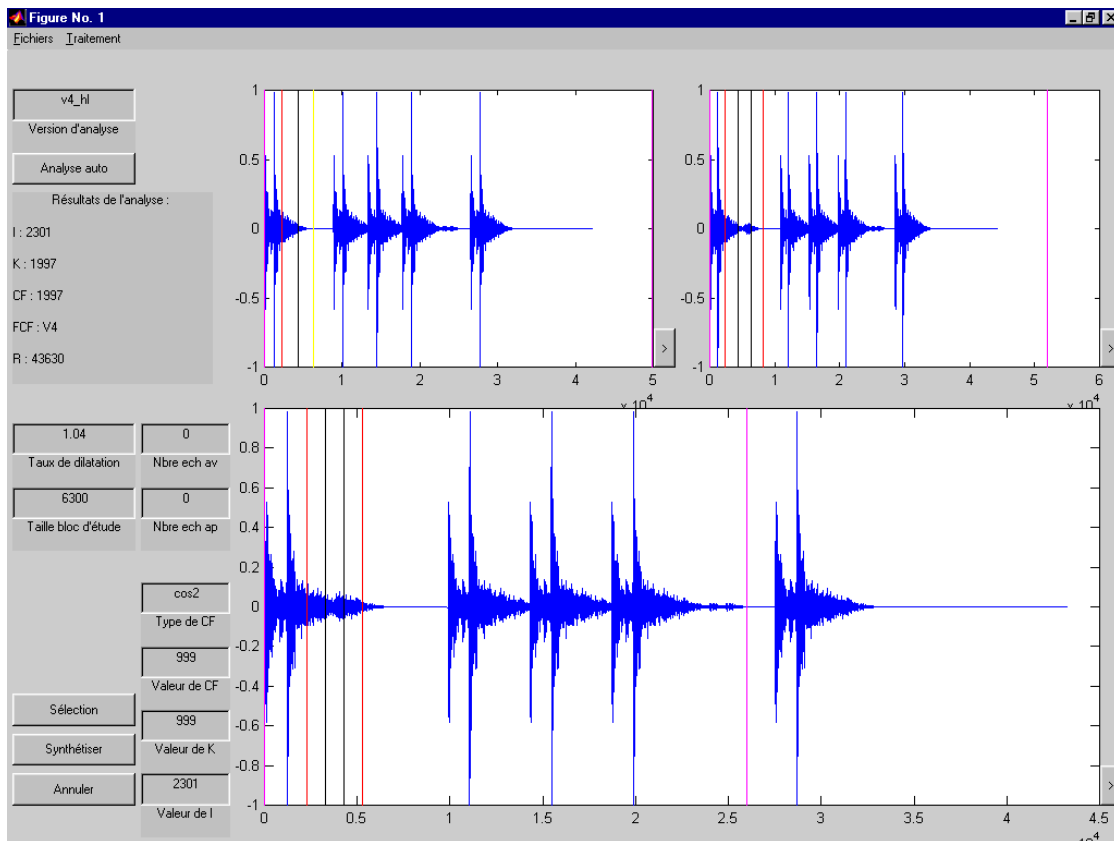


Figure 3.17 – Capture d'écran de l'interface HARMOLAB

d'échantillons, CF correspondant à FE , et FCF fournit un code relatif à la forme du fondu-enchaîné).

De cette manière, il est possible de visualiser les défauts audibles, mais il n'est pas possible de les modifier : c'est dans cette optique que la partie inférieure du logiciel a été conçue.

La grande fenêtre de visualisation est appelée "fenêtre de travail", car on peut lui appliquer tous les traitements manuels possibles et écouter le résultat de chaque modification. On a en effet la possibilité d'écouter les différents sons (original, modifié au préalable, modifié manuellement), avec un réglage des durées du son entendu avant et après le point d'insertion. Les paramètres manuels peuvent être tapés au clavier, mais il existe également la possibilité de sélectionner à la souris les segments K_A et K_B . La fenêtre de travail correspond donc au son original jusqu'à ce que la touche "synthétiser" soit activée, auquel cas le segment K_M calculé apparaît entre K_A et K_B .

On peut voir que pour ce son de castagnettes, la sélection d'un segment deux fois moins long ($K = 999$ au lieu de 1997) fait disparaître le redoublement.

Grâce à HARMOLAB, nous pouvons donc observer les défauts afin de comprendre les processus qui les ont induits, et ainsi de concevoir des critères permettant de les éviter.

3.4.3 Critère de sélection de K

Le premier des trois paramètres que nous étudions est la durée K du segment inséré. C'est celui qui permet d'améliorer radicalement la qualité d'un algorithme aveugle. C'est également celui dont les critères à adopter sont les plus difficiles à trouver.

Lorsque le signal original est périodique, une durée de segment inséré égale à un multiple de la période fondamentale ne provoque aucun défaut⁵.

Malheureusement, ce cas idéal se produit rarement car le signal n'est généralement pas exactement périodique. Dans de telles circonstances, il est nécessaire que les segments K_A et K_B , qui contribuent à la création du segment inséré K_M , soient le plus similaire possible afin que leur mixage produise un signal cohérent avec son voisinage. Il est donc indispensable de mettre en place une mesure de similarité afin d'extraire la durée K pour laquelle les deux segments successifs se ressemblent le plus.

Nous étudions ici différentes mesures de similarité, exprimées sous l'hypothèse que la marque de lecture à l'itération i est située à l'origine ($L_i = 0$), pour éviter de surcharger inutilement les formules. Nous voyons ensuite sous quelles contraintes les paramètres doivent être ajustés. Puis nous examinons ces paramètres dans le cas de sons quasi-stationnaires.

Mesures de similarités

Fonction de différence d'amplitude moyenne

La fonction de différence d'amplitude moyenne ou AMDF (pour "Average Magnitude Difference Function") peut être utilisée pour fournir une mesure de similarité entre deux formes d'onde [HMC92, VR93, Lar98]. Dans notre cas, s'agissant de comparer deux segments issus du même signal, on définit l'AMDF de la manière suivante :

$$AMDF(k) = \sum_{n=0}^{N_c-1} |s(n) - s(n+k)| \quad k \in [k_{min}, k_{max}]$$

avec N_c la durée sur laquelle est estimée la similarité.

Le minimum de cette fonction, trouvé pour $k = K$, indique que les segments $s(n)$ pour $n \in [0, N_c - 1]$ et $s(m)$ pour $m \in [K, N_c - 1 + K]$ sont les plus similaires (au sens de cette fonction bien sûr).

Cette fonction requiert une faible puissance de calcul, mais elle est cependant sensible au bruit [Lar98] ainsi qu'à l'amplitude du signal [HMC92].

Fonction d'autocorrélation

Nous définissons la fonction d'autocorrélation de la manière suivante :

$$C(k) = \sum_{n=0}^{N_c-1} s(n)s(n+k) \quad k \in [k_{min}, k_{max}]$$

avec N_c la durée sur laquelle est estimée la similarité.

On peut interpréter cette fonction comme un produit scalaire entre deux segments de longueurs égales N_c , le premier restant fixe, le second glissant avec k . La figure 3.18 schématise cette interprétation en produit scalaire "glissant". On y représente le signal original sous forme

5. Pour un signal composé d'une sinusoïde pure, la détection de deux passages successifs par zéro indique la période fondamentale. Malheureusement, ce type d'algorithme extrêmement simple ne fonctionne généralement pas dans le cas de signaux plus complexes.

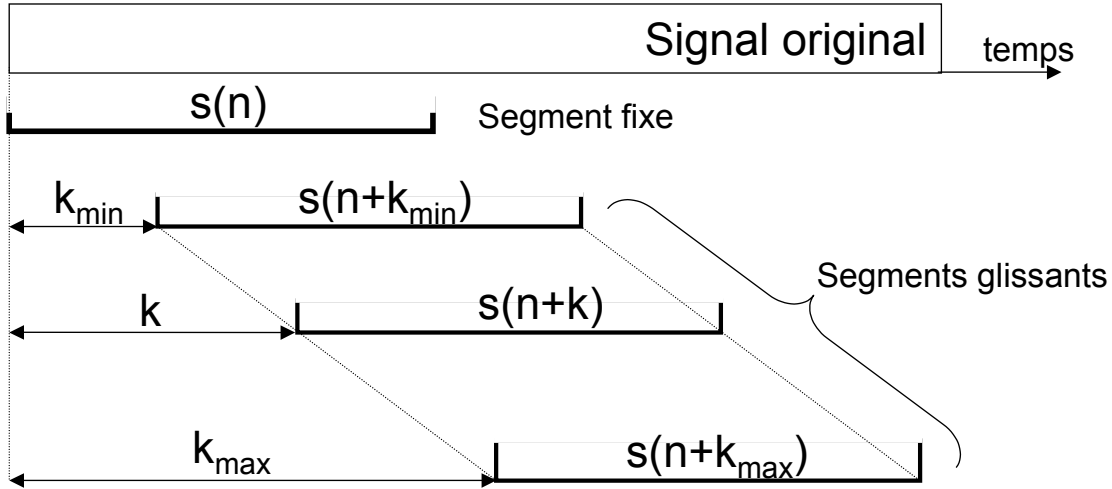


Figure 3.18 – *Interprétation de l'autocorrélation en tant que produit scalaire "glissant"*

d'une bande, et les segments du signal temporel $s(n)$ et $s(n+k)$ pour lesquels le produit scalaire est calculé.

Le maximum de cette fonction, donné pour $k = K$, indique la durée optimale du segment à insérer.

Les résultats de cette fonction sur un signal non-stationnaire ne sont cependant pas ceux attendus car plus le signal possède une énergie élevée, plus la fonction d'autocorrélation est élevée : le choix entre deux segments totalement semblables mais d'énergie faible et deux segments de forme semblable mais d'énergies différentes se porte parfois sur le deuxième cas ! C'est ce que révèle la figure 3.19, en montrant un signal sinusoïdal dont l'énergie varie brusquement, et sa fonction d'autocorrélation associée. Le maximum de cette fonction est donné pour $K = 200$, ce qui indique que les segments estimés les plus similaires sont composés de deux périodes identiques mais d'énergie extrêmement différente, ce qui mène à un signal dilaté défectueux.

Dans le cas d'un tel signal, il est nécessaire d'utiliser une mesure de similarité dont le maximum sera trouvé pour $K = 100$. Il est pour cela indispensable d'appliquer une pondération à la fonction d'autocorrélation.

Fonction d'autocorrélation normalisée

La variation de l'autocorrélation avec l'énergie du signal est un frein à une mesure de similarité lorsque l'on doit comparer la ressemblance des formes d'onde. Il semble dès lors évident que pour s'affranchir des dissemblances énergétiques, il est nécessaire de normaliser en énergie chacun des segments jouant un rôle dans le produit scalaire. On arrive alors à la formule de l'autocorrélation normalisée suivante :

$$CN(k) = \frac{\sum_{n=0}^{N_c-1} s(n)s(n+k)}{\sqrt{\sum_{n=0}^{N_c-1} s^2(n) \sum_{n=0}^{N_c-1} s^2(n+k)}} \quad k \in [k_{min}, k_{max}] \quad (3.2)$$

avec N_c la durée sur laquelle est estimée la similarité.

Comme dans le cas de la fonction d'autocorrélation, le maximum de cette fonction donné pour $k = K$ indique la durée optimale du segment à insérer. En effet, si l'on considère que

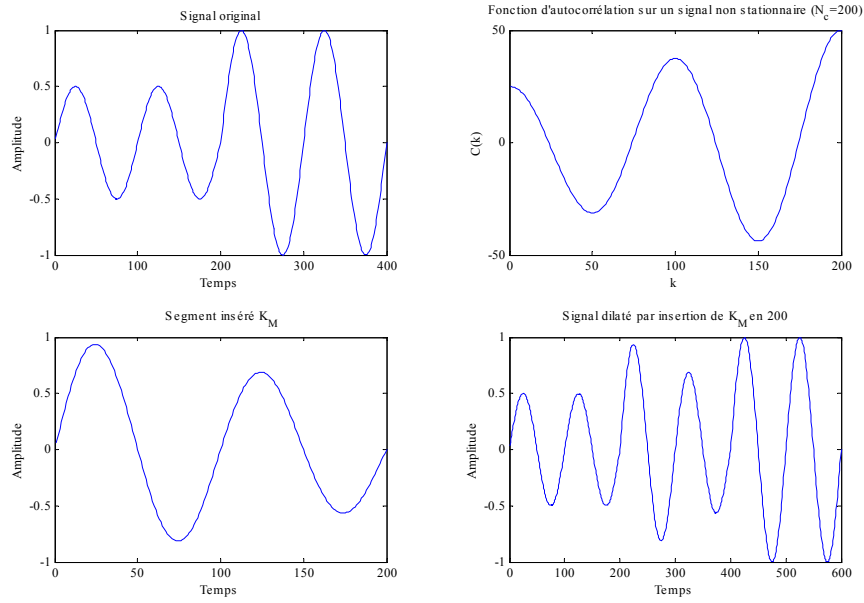


Figure 3.19 – Inconvénient de la fonction d'autocorrélation : choix du segment inséré provoquant un artefact (modulation d'amplitude)

le signal possède une période fondamentale T_0 , deux segments espacés d'une durée $K = T_0$ possèdent une corrélation maximum, égale à 1 pour un signal exactement périodique puisque la fonction est normalisée (ses valeurs varient entre -1 et $+1$). Il s'ensuit que les segments K_A ($s(n)$ pour $n \in [0, N_c - 1]$) et K_B ($s(m)$ pour $m \in [K, N_c - 1 + K]$) sont similaires.

Si $k_{min} = 0$, il est évident qu'un maximum de la fonction sera trouvé pour cette valeur puisqu'il ne peut exister un signal plus corrélé que lui-même. Or cette valeur ne nous intéresse pas car elle n'implique aucune insertion de segment ($K = 0$), c'est-à-dire qu'on ne dilate pas le signal.

Si le signal n'est plus exactement périodique, le maximum de la fonction nous permet d'extraire la durée K pour laquelle les deux segments successifs sont le plus similaires (au sens de la corrélation normalisée), même s'ils ne sont pas totalement équivalents. Plus la valeur de la corrélation est élevée, plus les segments sont similaires, et plus il est probable que l'insertion du segment K_M , mixage entre K_A et K_B , sera perceptivement inaudible.

Cette mesure de similarité règle le problème de comparaison de signaux d'énergie croissante ou décroissante. Dorénavant, les différences d'amplitude ne sont plus prises en compte, ce qui règle même les problèmes précédents par la possibilité de sélectionner des segments identiques, s'ils possèdent une faible énergie, comme on peut le voir dans la figure 3.20 (la fonction d'autocorrélation normalisée est maximale pour les multiples de $K = 100$).

Cependant, puisque l'on ne considère plus les amplitudes de la forme d'onde, il est indispensable de mettre au point un critère a posteriori capable de détecter lorsque les deux segments successifs K_A et K_B sont d'énergie très différente. En effet, l'insertion d'un segment K_M incorrect provoque un redoublement de la transition entre niveau faible et fort. Celui-ci peut alors être perçu distinctement s'il est trop long (à l'image de la figure 3.19). En revanche, il peut bénéficier du phénomène de masquage temporel [Can00] s'il est assez court.

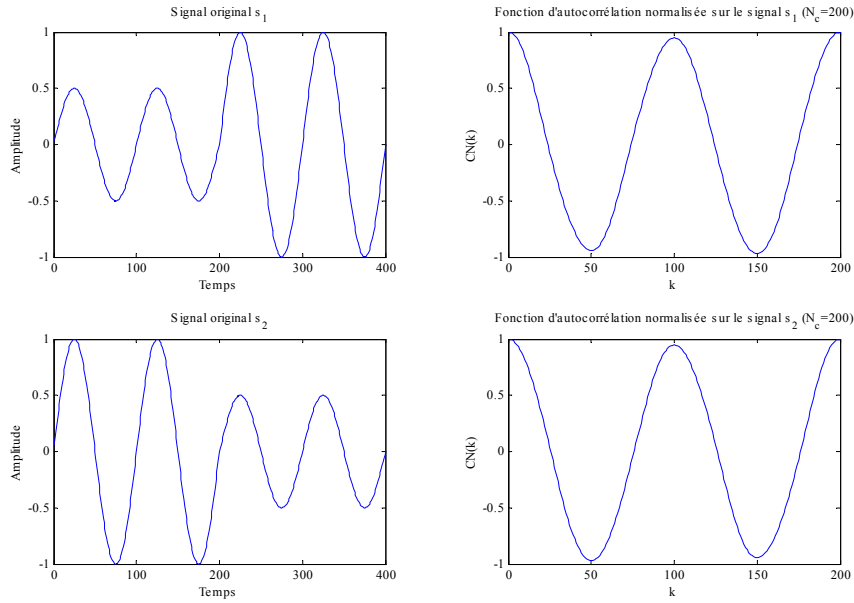


Figure 3.20 – Fonction d'autocorrélation normalisée sur un signal non stationnaire

Optimisations du calcul de l'autocorrélation

On remarque que puisque l'on maximise la fonction d'autocorrélation normalisée dans le but d'extraire K , seules les valeurs relatives nous intéressent. Le terme du dénominateur $\sum_{n=0}^{N_c-1} s^2(n)$ restant constant quel que soit k , il devient donc inutile de le calculer. Il en résulte que la maximisation de la fonction d'autocorrélation normalisée peut être réalisée à partir de la formule suivante :

$$CN(k) = \frac{\sum_{n=0}^{N_c-1} s(n)s(n+k)}{\sqrt{\sum_{n=0}^{N_c-1} s^2(n+k)}} \quad k \in [k_{min}, k_{max}]$$

Néanmoins, cette mesure n'assure plus que les valeurs sont comprises entre -1 et +1, ce qui peut être un inconvénient dans le cas d'une utilisation de l'autocorrélation normalisée en tant que mesure de stationnarité.

Laroche [Lar95] montre qu'un sous-échantillonnage de la fonction d'autocorrélation est possible. Cette technique permet de gagner un facteur 35 sur la puissance de calcul. On perd cependant en précision sur la durée K du segment. Par contre, en cumulant cette technique avec un procédé dichotomique (on précise la corrélation pour les valeurs de k voisines de celle pour laquelle a été trouvé le maximum), il est possible de conserver à la fois la précision et le gain en calcul. Cette technique n'assure plus de trouver le maximum absolu de la fonction d'autocorrélation normalisée, mais seulement un maximum local. Cependant cet inconvénient est généralement sans conséquence audible pour un taux de sous-échantillonnage raisonnable (inférieur à 10).

Laroche [Lar95] propose également d'utiliser l'algorithme rapide de FFT afin de calculer la séquence d'autocorrélation, puisqu'il est connu que sa transformée de Fourier correspond au

spectre de puissance du signal [OS75].

Ajustement des paramètres

La fonction d'autocorrélation possède 3 paramètres dont l'ajustement s'avère extrêmement critique pour la qualité de l'algorithme. Il s'agit de k_{min} , k_{max} et N_c . Chacune de ces valeurs est fixée en réalisant un compromis.

Ajustement de k_{min}

La valeur k_{min} doit être supérieure à une valeur pour laquelle l'autocorrélation normalisée est encore élevée du fait de sa proximité de la valeur maximale en $k = 0$ (on rappelle que $CN(0) = 1$). En effet, il n'est pas souhaitable de sélectionner un segment K plus court que la période fondamentale présente dans le signal, même si la fonction d'autocorrélation normalisée indique une forte ressemblance, car les fréquences inférieures à $1/K$ sont modifiées localement (si cette insertion n'intervient qu'une seule fois) ou globalement (pour des insertions périodiques). D'autre part, plus la valeur de k_{min} est faible, plus le coût en puissance de calcul est important. Si le coût en calcul n'est pas un problème, il peut être intéressant de calculer l'autocorrélation normalisée jusqu'à une valeur de $k = 0$, et d'adapter k_{min} selon la fonction obtenue (par exemple en fixant k_{min} au premier passage à zéro de la fonction d'autocorrélation normalisée).

La limite supérieure pour k_{min} est dictée par plusieurs facteurs : l'anisochronie susceptible de se produire (on doit être en mesure d'insérer un segment ne produisant pas de déformation rythmique), l'évolution du signal (plus le segment inséré est long, moins sa stationnarité est probable) et sa différence à k_{max} (discuté dans la suite).

Ajustement de k_{max}

La valeur k_{max} doit être supérieure à la période fondamentale susceptible d'être présente dans le signal. On doit en effet insérer un segment contenant au moins un multiple d'une période fondamentale, sans quoi ces basses fréquences sont victimes d'artefacts audibles (discontinuité des partiels et modification de la période fondamentale).

Inversement, une valeur trop élevée de k_{max} mène à la possibilité d'insérer un trop long segment, ce qui peut poser des problèmes d'anisochronie et de redoublements d'autant plus audibles que le segment est grand (le masquage temporel qui permet de fusionner deux transitoires très rapprochés ne joue plus, et ces deux transitoires sont entendus distinctement).

Ajustement de $(k_{min} - k_{max})$

La différence $(k_{min} - k_{max})$ joue également un rôle dans le choix de ces paramètres. En effet, les périodes fondamentales "couvertes" (ou détectées) par la mesure de similarité s'étendent de k_{max} à k_{min} . Les multiples de ces périodes fondamentales sont également couvertes (de $n.k_{max}$ à $n.k_{min}$ avec $n \in \mathbb{N}$). Il en résulte que si $k_{min} > k_{max}/2$, alors les fréquences comprises entre $1/k_{min}$ et $2/k_{max}$ ne sont pas détectées. Comme il est regrettable de ne pas assurer une continuité dans la détection des fréquences de $1/k_{max}$ à $F_e/2$, il est préférable de respecter la contrainte suivante :

$$k_{min} < \frac{k_{max}}{2}$$

Ajustement de N_c

La durée N_c sur laquelle est estimée la similarité fait l'objet également d'un compromis entre résolution temporelle et fréquentielle. D'une part, une durée courte permet de s'assurer d'une ressemblance entre les deux premiers segments successifs de longueur K (ce qui n'est pas

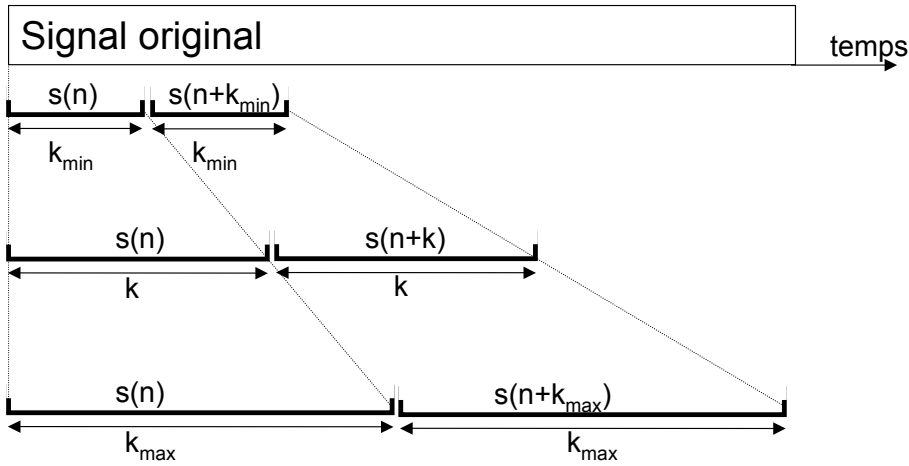


Figure 3.21 – Interprétation de l'autocorrélation en tant que produit scalaire "adaptatif"

forcément le cas pour un signal quasi-périodique avec N_c élevé, car la périodicité peut se révéler au milieu ou à la fin du segment, qui n'est pas pris en compte dans la construction de K_M . D'autre part, une durée longue permet de détecter des fréquences basses.

Ce compromis doit en outre être réalisé en tenant compte des paramètres k_{min} et k_{max} . Si $N_c < K$, le segment K_M est constitué des deux segments K_A et K_B dont une partie n'a pas fait l'objet de l'estimation de similarité. Si $N_c > K$, les échantillons situés au-delà de K ne sont pas utilisés pour construire K_M .

Cette dernière observation mène à l'élaboration d'une nouvelle mesure de similarité dite "adaptative" car la durée N_c sur laquelle est estimée le produit scalaire s'adapte à la durée k , comme le montre la figure 3.21 et conformément à l'équation suivante :

$$CN_a(k) = \frac{\sum_{n=0}^{k-1} s(n)s(n+k)}{\sqrt{\sum_{n=0}^{k-1} s^2(n) \sum_{n=0}^{k-1} s^2(n+k)}} \quad k \in [k_{min}, k_{max}]$$

Ainsi, pour une durée K donnée, la similarité est estimée sur cette durée précise, ce qui évite de tenir compte, dans l'estimation, d'une partie de signal qui n'est pas utilisé pour la création de K_M , ou au contraire de construire un segment K_M avec une partie de signal dont la similarité n'a pas été estimée.

L'inconvénient avec cette technique réside dans le fait que si l'on désire employer la fonction d'autocorrélation normalisée en tant que mesure de stationnarité du signal (plus la valeur est proche de 1, plus le signal est stationnaire et périodique), le résultat est biaisé : on peut en effet avoir une forte corrélation entre deux segments successifs très courts à l'intérieur d'un transitoire.

D'autre part, à cause du caractère quasi-stationnaire du signal (variation lente d'amplitude et/ou de fréquence), plus K est grand, moins les deux segments à comparer ont des chances de se ressembler.

Comportement des sons quasi-stationnaires

Nous définissons les sons quasi-stationnaires comme des sons variant suffisamment lentement dans le temps (en amplitude et en fréquence) pour que, localement, on puisse les considérer stationnaires. Il peut s'agir de sons harmoniques, ou bien inharmoniques.

Sons harmoniques

Un son est dit harmonique lorsque les fréquences de tous ses partiels (composantes sinusoïdales) sont des multiples entiers de la fréquence fondamentale (la composante sinusoïdale la plus grave). Le signal peut alors être considéré comme une somme de sinusoïdes $s_k(t)$ additionné d'un éventuel bruit $b(t)$, comme l'indique l'équation suivante :

$$s(t) = b(t) + \sum_{k=1}^{N_{\text{partiels}}} s_k(t) \quad (3.3)$$

$$= b(t) + \sum_{k=1}^{N_{\text{partiels}}} a_k \sin(2\pi k f_0 t + \varphi_k) \quad (3.4)$$

avec f_0 la fréquence fondamentale du signal, N_{partiels} le nombre de partiels, a_k et φ_k l'amplitude et la phase du $k^{\text{ième}}$ partiel.

Dans ce cas, les méthodes temporelles donnent d'excellents résultats tant qu'il est possible d'insérer un segment dont la durée est un multiple de la période fondamentale $T_0 = 1/f_0$. Si la durée du segment inséré est insuffisante (lorsque $k_{\text{max}} < T_0$), il se produit une discontinuité de la forme d'onde provoquant un "clic", atténué par l'utilisation d'un fondu-enchaîné mais généralement jamais totalement supprimé.

Il existe bien sûr une tolérance pour adapter la durée du segment inséré à la période fondamentale, que nous avons estimé empiriquement à environ 5% de la durée d'une période pour des sons purs allant de 20 Hz à 1 kHz.

Mais pour une valeur maximale du segment inséré de 25 ms [Lar93], tous les sons dont la fréquence fondamentale est inférieure à 40 Hz (à 5% près) ont des artefacts audibles, s'apparentant à des modulations basse fréquence localisées temporellement.

Or, il est possible de rencontrer des instruments capables de produire des fréquences plus basses que 40 Hz dans la musique (piano, harpe, saxophone contrebasse [Pie83]), et d'autre part, il est extrêmement courant d'en entendre dans les effets spéciaux liés au canal basse fréquence des films.

Ainsi, nous pensons qu'il est nécessaire d'augmenter la valeur de k_{max} de manière à éviter les artefacts dus à l'insertion d'un segment trop court par rapport à la période fondamentale.

Deux exemples sonores montrent les résultats d'un traitement sonore réalisé avec $k_{\text{max}} = 25$ ms [66] et $k_{\text{max}} = 40$ ms [67] sur un son synthétique d'une note de piano à 27,5 Hz ("A0", son totalement harmonique). On peut entendre des artefacts évidents dans le premier exemple, qui disparaissent dans le second.

Sons inharmoniques

Un son est dit inharmonique lorsque les fréquences de ses partiels ne sont plus uniquement des multiples entiers de la fréquence fondamentale. C'est le cas généralement pour des instruments ou des sources polyphoniques, mais également pour des notes uniques de certains instruments tels le piano (voir section 2.2.5) ou de manière plus évidente la cloche.

Dans ce cas, comme nous l'avons dit en section 2.2.5, il est difficile, voire impossible, de trouver un compromis permettant d'adapter la durée du segment inséré à toutes les périodes

présentes dans le signal. Cependant, en augmentant la valeur de k_{max} , on augmente les chances de s'approcher d'un PPCM (Plus Petit Commun Multiple) de la période de tous les partiels.

La tolérance de 5% sur l'estimation d'une période fondamentale couplée aux phénomènes de masquage intrinsèques dans des sons très complexes (rendant parfois les artefacts inaudibles) peuvent expliquer en partie les résultats sonores de haute qualité obtenus dans certains cas d'inharmonicité (sons [68, 69]; la figure 3.22 montre le spectre d'une des résonances inharmoniques) et de polyphonie (sons [8, 10]).

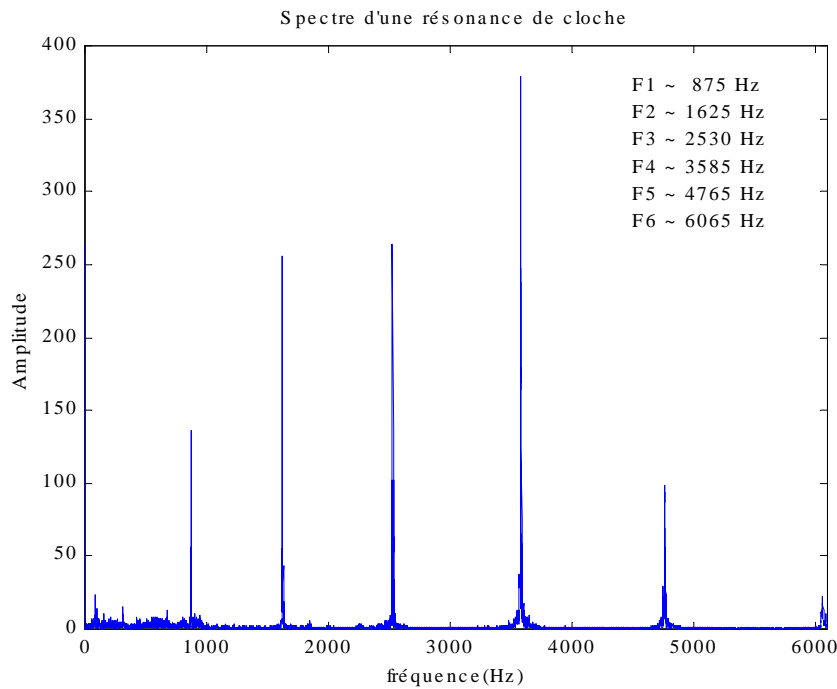


Figure 3.22 – Spectre d'une résonance inharmonique d'un son d'horloge

Conclusion sur le critère de sélection de K

Pour un point d'insertion I donné, la fonction d'autocorrélation normalisée permet d'extraire la durée K pour laquelle les deux segments successifs K_A et K_B sont le plus similaires. Ce calcul est très coûteux, mais il existe plusieurs manières de l'optimiser. L'ajustement des paramètres de cette fonction est critique, et nous montrons qu'une amélioration de la qualité sonore est possible en insérant un segment d'une durée supérieure à 40 ms afin de traiter correctement les sons à très basse fréquence (jusqu'à environ 20 Hz) et d'améliorer certains sons inharmoniques. Cependant, ce choix possède un double inconvénient : celui de rendre l'anisochronie encore plus audible dans le cas de séquences rythmiques, et de produire parfois des redoublements extrêmement perceptibles. Nous nous efforçons dans la section ?? de limiter ces effets indésirables.

3.4.4 Critère de sélection de I

Le point d'insertion I est le second paramètre des méthodes temporelles. Il offre la possibilité de se placer à un endroit du signal pour lequel l'insertion du segment K_M peut être réalisé sans artefact audible. Les critères assurant l'absence de défaut audible après insertion peuvent être différents : recherche d'une zone stationnaire et périodique afin d'insérer une période fondamentale, ou au contraire, recherche d'une zone transitoire afin d'insérer un segment comportant un

défaut devenu inaudible car bénéficiant du masquage temporel. Nous étudions ici des méthodes permettant de mettre en place de tels critères.

Il faut rappeler que le point d'insertion I est limité à la fois par les problèmes d'anisochronie et de la contrainte du "retour sur la droite idéale" ($R > 0$). On ne peut donc repousser indéfiniment la décision d'insérer un segment.

Détection de zone transitoire

De nombreuses méthodes de détection et de modélisation existent dans la littérature (voir section 3.3.2). Notre but n'est pas d'en développer une nouvelle, mais d'utiliser et de comprendre le fonctionnement de celle qui paraît la plus adaptée à notre problème.

Masri [Mas96] propose une comparaison de 3 méthodes.

La première méthode consiste à observer la distribution d'énergie dans le domaine fréquentiel. Le transitoire introduit généralement une discontinuité de phase, donc il y a une augmentation d'énergie aux hautes fréquences. Lorsque cette variation d'énergie dépasse un certain seuil, on estime qu'il y a un transitoire.

La seconde méthode est proche de la première, sauf qu'elle évalue les dissimilarités spectrales sur le spectre total, et pas uniquement à hautes fréquences.

La dernière méthode est basée sur un suiveur d'enveloppe temporelle⁶. Le principe est de créer une courbe $P(i)$ en sélectionnant les valeurs maximales $V(i)$ de chacun des blocs successifs de durée 1 ms extraits du signal redressé. La courbe obtenue est lissée par l'application de l'algorithme itératif suivant :

$$\begin{aligned} P(i) &= V(i) && \text{si } V(i) \geq P(i) \\ &= T_d \cdot P(i-1) && \text{sinon} \end{aligned}$$

Le temps de descente T_d n'est pas spécifié, mais nous l'estimons à environ 150 ms. Un transitoire est détecté lorsque le rapport $P(i)/P(i-1)$ dépasse un seuil généralement fixé à 1,7.

De ces trois méthodes, Masri affirme que pour la voix spécifiquement, même dans les autres cas, la troisième est la meilleure. De plus, elle bénéficie d'une meilleure résolution temporelle (2 à 4 fois meilleure que les autres). En outre, elle est extrêmement peu coûteuse en calculs. C'est donc la méthode que nous employons dorénavant pour détecter les transitoires.

Adaptation de la corrélation à la détection de transitoire

Nous proposons ici 3 mesures de similarités différentes, qui dépendent de la position des transitoires et de la stratégie adoptée : la première consiste à calculer la durée du segment inséré après un transitoire, la seconde avant un transitoire, et la troisième entre deux transitoires.

Autocorrélation normalisée "vers la droite"

La fonction d'autocorrélation normalisée définie en équation 3.2 compare les deux segments successifs qui glissent "vers la droite" (dans le sens d'un écoulement du temps positif), comme le montre le schéma 3.18.

Elle est adaptée à une mesure de similarité calculée après la fin d'un transitoire car on est assuré que le transitoire ne sera pas présent dans le segment inséré.

6. Des précisions sur les suiveurs d'enveloppe peuvent être trouvées dans [Zöl97].

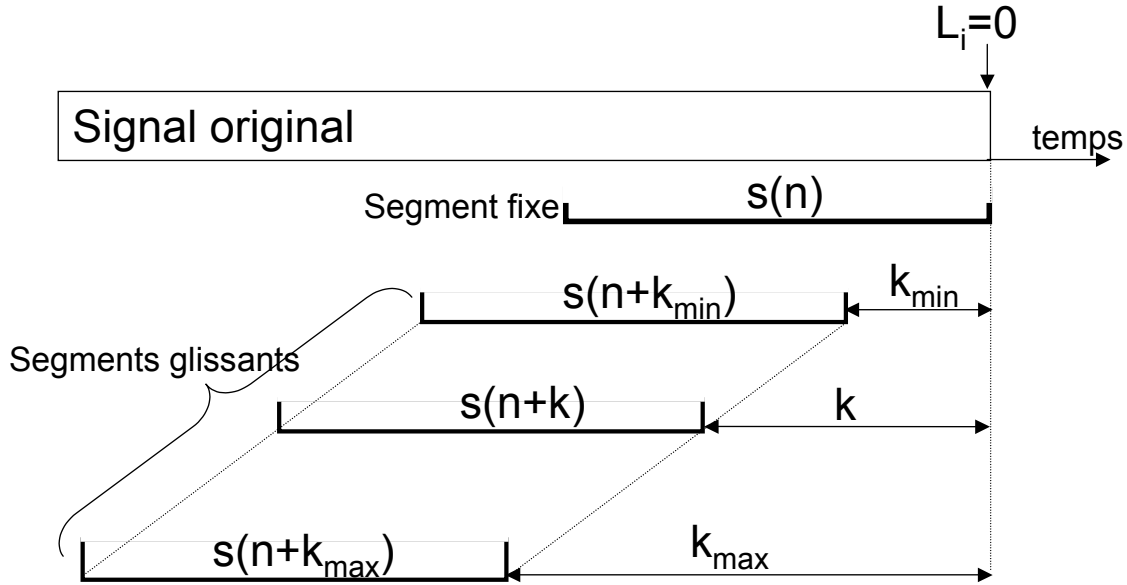


Figure 3.23 – Interprétation de l'autocorrélation en tant que produit scalaire glissant "vers la gauche"

Néanmoins, si un transitoire est présent dès les premiers échantillons suivant $L_i = 0$, celui-ci est inéluctablement dupliqué (même si la répétition peut être inaudible lorsque K est court).

Autocorrélation normalisée "vers la gauche"

En supposant qu'un critère de détection indique la présence d'un transitoire à partir de $L_i = 0$, il est préférable que le produit scalaire glisse "vers la gauche", à l'image du schéma 3.23 conformément à l'équation suivante :

$$CN_g(k) = \frac{\sum_{n=0}^{-N_c+1} s(n)s(n-k)}{\sqrt{\sum_{n=0}^{-N_c+1} s^2(n) \sum_{n=0}^{-N_c+1} s^2(n-k)}} \quad k \in [k_{min}, k_{max}]$$

Cette technique permet d'insérer un segment de quelque durée que ce soit, qui est toujours le plus proche possible du transitoire. Si un artefact dû au segment inséré est présent, il peut bénéficier du phénomène de masquage temporel et être rendu ainsi inaudible.

Autocorrélation normalisée "centrée"

On est parfois en présence d'une succession de transitoires. Afin d'éviter un redoublement, il est nécessaire d'insérer un segment issu d'une zone assez stable placée entre deux transitoires. Dans ce cas, une mesure de similarité symétrique de part et d'autre du point centré sur la zone stable, semble souhaitable. On aboutit alors à une mesure de similarité "centrée", définie par

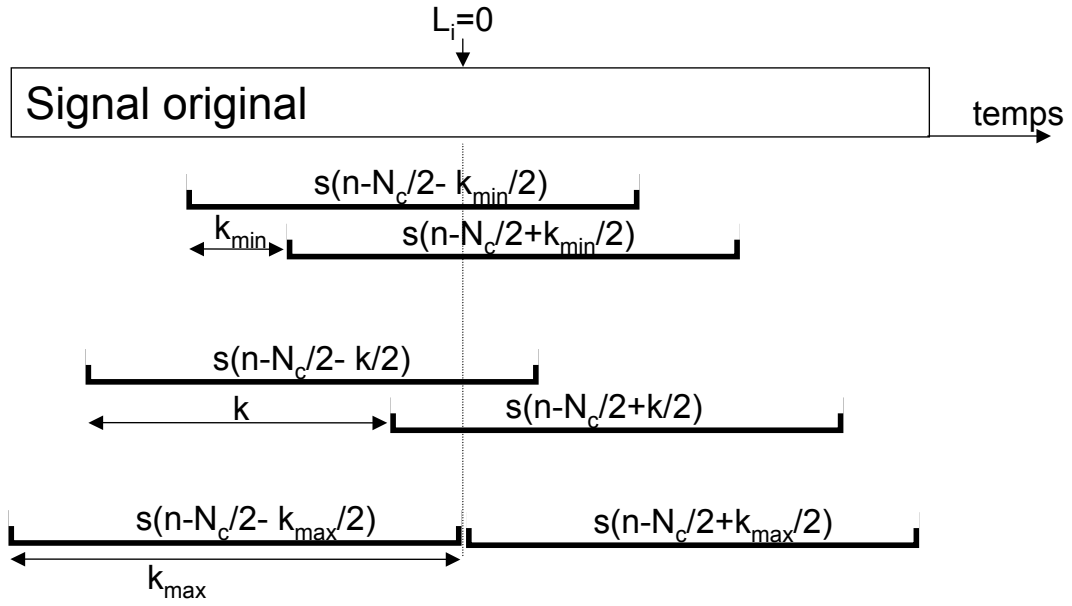


Figure 3.24 – Interprétation de l'autocorrélation en tant que produit scalaire "centré"

l'équation suivante et schématisée en figure 3.24 (avec $N_c = k_{max}$) :

$$CN_c(k) = \frac{\sum_{n=0}^{-N_c+1} s\left(n - \frac{N_c}{2} - \frac{k}{2}\right) s\left(n - \frac{N_c}{2} + \frac{k}{2}\right)}{\sqrt{\sum_{n=0}^{-N_c+1} s^2\left(n - \frac{N_c}{2} - \frac{k}{2}\right) \sum_{n=0}^{-N_c+1} s^2\left(n - \frac{N_c}{2} + \frac{k}{2}\right)}} \quad k \in [k_{min}, k_{max}]$$

Il est nécessaire d'adapter la formule précédente à l'utilisation avec un signal échantillonné en arrondissant à un entier les éventuelles fractions non entières. L'imprécision temporelle résultante (un demi-échantillon) n'a généralement aucune conséquence audible.

Ces différentes mesures de similarité, qui dépendent de la position du (ou des) transitoire(s), sont exprimées avec pour base l'autocorrélation normalisée donnée par l'équation 3.2, mais il est également possible d'utiliser toutes les autres mesures de similarité existantes.

Détection de zone stationnaire

Une zone stationnaire du signal se prête généralement bien à l'insertion d'un segment approprié. La détection d'une telle zone peut être déduite par complémentarité des zones transitoires, mais cette tactique possède des inconvénients⁷. Ici, nous proposons une méthode qui fournit directement une estimation relative de la stationnarité du signal en fonction du temps.

Jusqu'à présent, nous avons étudié la similarité entre deux segments pour un point I fixé. La fonction d'autocorrélation normalisée nous indique par son maximum la durée pour laquelle deux segments successifs sont similaires. De plus, cette valeur normalisée nous renseigne sur le degré

7. Un transitoire est un concept mal défini, qui s'étend (brièvement) dans le temps et dont la détection nécessite une intégration temporelle (il est donc préférable de parler de "zone transitoire") qui provoque une imprécision sur sa localisation temporelle. Il est donc parfois plus aisé de trouver des zones stationnaires que des zones transitoires.

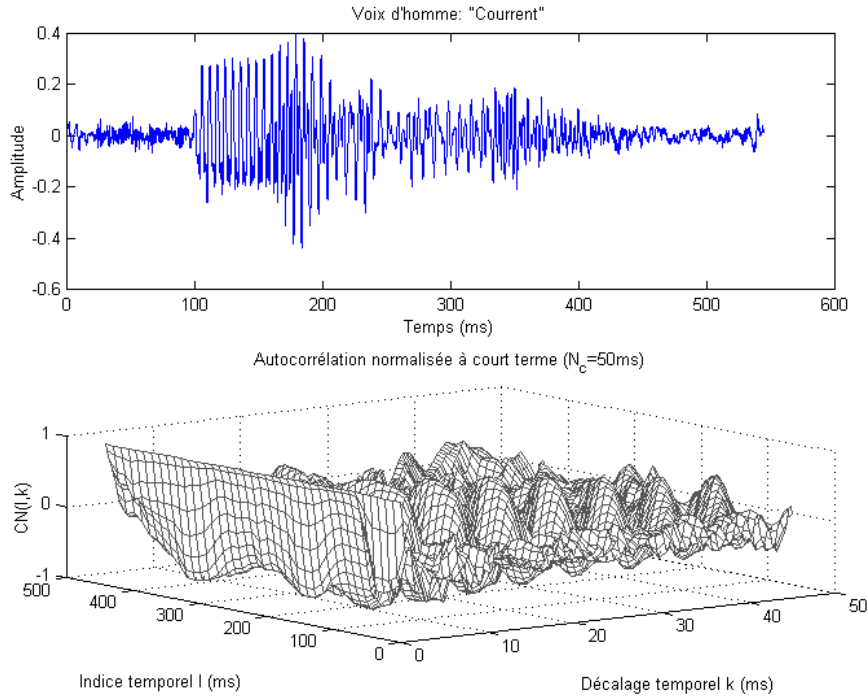


Figure 3.25 – Fonction d'autocorrélation normalisée à court terme sur un son de voix d'homme

de similarité. Il est donc naturel d'observer cette similarité au cours du temps, et ainsi déterminer simultanément la durée K et le point I pour lesquels la similarité est la plus importante. On aboutit alors à la "fonction d'autocorrélation normalisée glissante", dérivée de [Max86] et définie par :

$$CN(I, k) = \frac{\sum_{n=0}^{N_c-1} s(I+n)s(I+n+k)}{\sqrt{\sum_{n=0}^{N_c-1} s^2(I+n) \sum_{n=0}^{N_c-1} s^2(I+n+k)}}$$

On peut voir un exemple de cette fonction en figure 3.25 dans laquelle on remarque plusieurs choses :

- Pour $k = 0$, la fonction est toujours égale à 1 : nous avons déjà montré que cette valeur ne nous intéresse pas.
- Les valeurs pour $I < 100$ ms sont très faibles : le signal à ces positions correspond à un transitoire.
- Pour $100 < I < 200$ ms un pic revient périodiquement sur l'axe k : celui-ci indique la période d'environ 7 ms contenue dans le signal à ces instants.

Si l'on maximise cette fonction (en écartant les valeurs correspondant à $k < 5$ ms), on trouve donc une durée K qui correspond à la période fondamentale du signal dans la zone stationnaire. On extrait donc par cette méthode les deux segments qui sont le plus similaires dans les limites d'un signal donné.

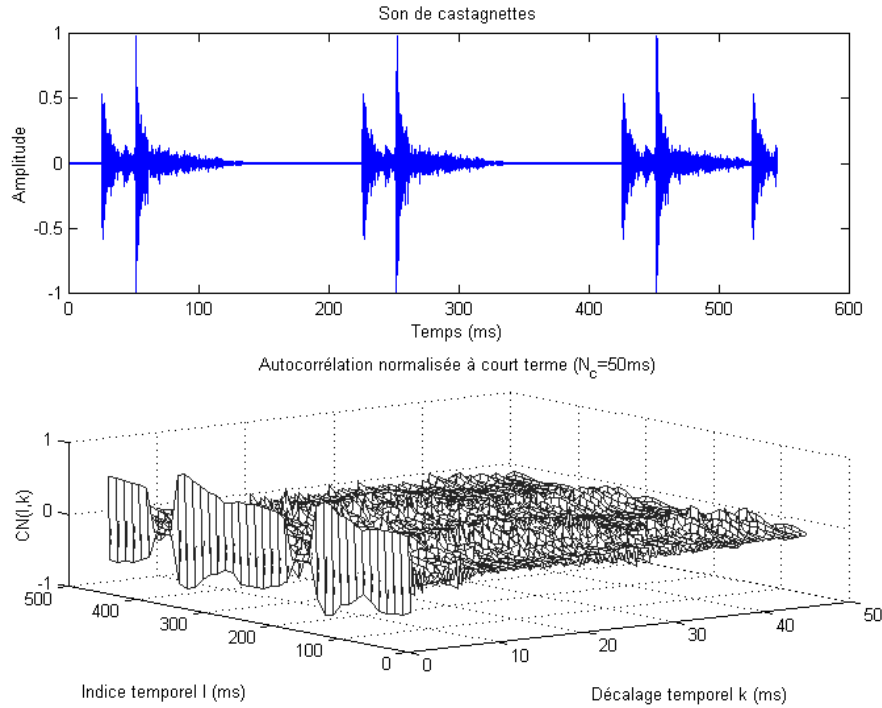


Figure 3.26 – Fonction d'autocorrélation normalisée à court terme sur un son de castagnettes

Cette mesure de similarité à court terme semble très efficace, mais elle est extrêmement coûteuse en espace mémoire et en puissance de calcul, bien que l'on puisse mettre au point un algorithme de calcul rapide pour cette fonction⁸.

La maximisation de cette fonction permet de fournir simultanément les paramètres I et K . Cependant, il peut être préférable dans certains cas de découpler la sélection de ces deux paramètres, car bien que les zones quasi-stationnaires soient bien repérées lorsqu'il y en a, la sélection d'un K long (pouvant mener à un redoublement) n'est pas impossible dans une zone transitoire où la corrélation semble aléatoire comme le montre la figure 3.26.

D'autre part, il n'est pas toujours souhaitable de dilater uniquement les parties stationnaires du signal. Par exemple, selon Covell, Withgott, Slaney [CWS97, CWS98] et d'après les travaux de van Santen [vS94], pour une application de contraction du signal vocal, les consonnes doivent en moyenne être plus compressées que les voyelles. En utilisant un point d'insertion I indépendant du signal, on réduit (ou on allonge) en moyenne aussi souvent les transitoires que les parties stationnaires.

Conclusion sur le critère de sélection de I

Le point d'insertion I peut être obtenu par l'application de différents critères liés à la présence ou l'absence de zones stationnaires ou transitoires. Cependant, la présence de ce degré de liberté ne fournit pas la garantie d'une modification sans artefact. En effet, même si l'on peut trouver le point d'insertion I pour lequel l'artefact devrait être le moins gênant, il n'est pas certain que le défaut soit inaudible. Il est donc nécessaire de se concentrer sur les problèmes (désynchronisation, redoublement, anisochronie) qui interviennent pour un point I arbitraire.

8. $CN(I+1, k)$ peut être déduit de $CN(I, k)$ en utilisant 2 MAC (multiplication-accumulation) pour la corrélation, et 2 MAC pour la normalisation de chacun des segments, soit au total 6 MAC.

3.4.5 Critère de sélection de FE

Le fondu-enchaîné utilisé pour atténuer les discontinuités de désynchronisation et d'amplitude est caractérisé à la fois par sa forme et par sa durée. Nous étudions chacune de ces caractéristiques et montrons que l'influence de la forme de la fenêtre sur le segment mixé est très faible comparée à l'influence de la taille de la fenêtre.

Forme du fondu-enchaîné

Problématique

Soit s_1 et s_2 deux signaux de durée identique K , devant être mixés afin d'obtenir un signal s' qui s'insère entre s_1 et s_2 dans le cas d'une élongation ($\alpha > 1$) ou qui remplacera s_1 et s_2 dans le cas d'une contraction ($\alpha < 1$).

Nous supposons que les deux signaux s_1 et s_2 se juxtaposent dans le signal original, et que le fondu-enchaîné est réalisé sur une durée K .

Soit w_1 et w_2 les fenêtres de pondération associées aux signaux s_1 et s_2 , jouant le rôle de fondu-enchaîné. Nous imposons à ces fonctions un certain nombre de contraintes :

- Tout d'abord, nous voulons éviter les discontinuités aux points de recollement. Pour cela, nous devons respecter les équations suivantes :

$$\left. \begin{array}{l} w_1(0) = 0 \quad ; \quad w_1(K) = 1 \\ w_2(0) = 1 \quad ; \quad w_2(K) = 0 \end{array} \right\} \alpha < 1$$

$$\left. \begin{array}{l} w_1(0) = 1 \quad ; \quad w_1(K) = 0 \\ w_2(0) = 0 \quad ; \quad w_2(K) = 1 \end{array} \right\} \alpha > 1$$
(3.5)

- Ensuite, nous souhaitons qu'elles soient monotones et régulières, afin de pouvoir prévoir facilement leur comportement.
- Enfin, nous leur imposons une contrainte de symétrie pour que la "mutation" d'un signal vers l'autre soit équilibrée (il n'y a aucune raison qu'un signal soit plus représenté qu'un autre) :

$$w_2(t) = w_1(K - t) \quad \forall t \in [0, K]$$
(3.6)

On peut remarquer que leur point d'intersection se situe en $t = K/2$.

Le signal s' obtenu par mixage est donné par l'équation suivante :

$$s'(t) = s_1(t)w_1(t) + s_2(t)w_2(t)$$
(3.7)

Optimisation sur des signaux corrélés

Supposons que les deux signaux que nous désirons mixer soient strictement identiques : $s_1 = s_2 \equiv s$.

Pour assurer un fondu-enchaîné perceptivement correct, on doit conserver l'amplitude entre le signal original s et le signal mixé s' (alors égal à s). Les fenêtres doivent alors vérifier l'équation suivante, obtenue à partir de l'équation 3.7 :

$$w_1(t) + w_2(t) = 1 \quad \forall t \in [0, K]$$
(3.8)

La contrainte de symétrie implique que $w_1(K/2) = w_2(K/2) = 1/2$.

En théorie, tous les types de fenêtres couramment utilisées dans l'analyse de signaux (Hanning, Hamming, Kaiser, Blackman-Harris...) peuvent être utilisées. En pratique, il est plus simple (en terme de conception, de puissance de calcul et d'espace mémoire) d'utiliser comme fenêtre de pondération la fonction linéaire décroissante w_{1c} (indice c pour "corrélé") de l'équation suivante, qui répond à toutes les contraintes pré-citées :

$$w_{1c}(t) = 1 - \frac{t}{K} \quad \forall t \in [0, K] \quad (3.9)$$

La fonction croissante w_{2c} qui lui est associée est donc donné par :

$$w_{2c}(t) = \frac{t}{K} \quad \forall t \in [0, K] \quad (3.10)$$

Il est trivial que si le signal mixé s' est identique aux signaux s_1 et s_2 , alors la puissance est également conservée.

Optimisation sur des signaux décorrélés

Nous sommes parfois confrontés, dans l'algorithme de dilatation-p, au cas où les signaux s_1 et s_2 à mixer sont décorrélés, par exemple dans un passage de bruit blanc. Pour étudier ce cas, nous supposons que s_1 et s_2 sont des signaux aléatoires dont les puissances instantanées notées $P_{s_1}(t)$ et $P_{s_2}(t)$ sont identiques :

$$P_{s_1}(t) = P_{s_2}(t) \equiv P_s(t)$$

La puissance instantanée est définie à l'aide de l'espérance mathématique E de la manière suivante :

$$P_s(t) = E[s^2(t)]$$

De plus, nous supposons que ces signaux sont centrés, ce qui se traduit par :

$$E[s_1(t)] = E[s_2(t)] = 0$$

Les signaux étant décorrélés, on a aussi :

$$E[s_1(t)s_2(t)] = 0$$

Le critère de conservation d'amplitude retenu pour des signaux corrélés n'a plus aucun sens dans le cas de signaux décorrélés. Le critère perceptif à retenir ici est la conservation de la puissance instantanée entre les signaux originaux et le signal mixé. On exprime donc la puissance instantanée $P_{s'}(t)$ du signal mixé en fonction des signaux $s_1(t)$ et $s_2(t)$:

$$\begin{aligned} P_{s'}(t) &= E[(s_1(t)w_1(t) + s_2(t)w_2(t))^2] \\ &= E[(s_1(t)w_1(t))^2] + E[(s_2(t)w_2(t))^2] + 2E[s_1(t)s_2(t)w_1(t)w_2(t)] \end{aligned}$$

Or, comme $w_1(t)$ et $w_2(t)$ sont des signaux déterministes (alors considérés comme des scalaires dans l'espérance mathématique), et que les signaux aléatoires $s_1(t)$ et $s_2(t)$ sont décorrélés, il en résulte :

$$\begin{aligned} P_{s'}(t) &= w_1^2(t)E[s_1^2(t)] + w_2^2(t)E[s_2^2(t)] \\ &= w_1^2(t)P_{s_1}(t) + w_2^2(t)P_{s_2}(t) \\ &= (w_1^2(t) + w_2^2(t))P_s(t) \end{aligned}$$

On en déduit que la conservation de la puissance instantanée est sous-tendue à l'équation suivante :

$$w_1^2(t) + w_2^2(t) = 1 \quad (3.11)$$

La contrainte de symétrie implique que $w_1(K/2) = w_2(K/2) = \sqrt{2}/2$.

En pratique, il est simple d'utiliser la fonction trigonométrique décroissante w_{1d} (indice d pour "décorrélé") de l'équation suivante, qui répond à toutes les contraintes précédentes :

$$w_{1d}(t) = \cos\left(\frac{\pi t}{2K}\right) \quad \forall t \in [0, K]$$

On en déduit la fonction croissante w_{2d} associée grâce à l'équation 3.11 :

$$w_{2d}(t) = \sin\left(\frac{\pi t}{2K}\right) \quad \forall t \in [0, K]$$

Fondu-enchaîné non adapté

L'utilisation de fenêtres de pondération adaptées aux signaux décorrélés (voir équation 3.11) ne convient pas pour des signaux corrélés ($s_1 = s_2 \equiv s$). La conservation de la puissance instantanée n'est en effet plus assurée à chaque instant. Par exemple, à l'instant $t = K/2$, la puissance instantanée du signal mixé est donnée par :

$$\begin{aligned} P_{s'}(K/2) &= E\left[\left(s(K/2)(w_1(K/2) + w_2(K/2))\right)^2\right] \\ &= E[2s^2(K/2)] \\ &= 2P_s(K/2) \end{aligned}$$

De même, l'utilisation de fenêtres de pondération adaptées aux signaux corrélés (voir équation 3.8) ne convient pas pour des signaux décorrélés. La conservation de la puissance n'est en effet plus assurée. Par exemple, à l'instant $t = K/2$, la puissance instantanée du signal mixé est donnée par :

$$\begin{aligned} P_{s'}(K/2) &= E\left[\left(s_1(K/2)w_1(K/2) + s_2(K/2)w_2(K/2)\right)^2\right] \\ &= E\left[\frac{s_1^2(K/2)}{2}\right] + E\left[\frac{s_2^2(K/2)}{2}\right] \\ &= \frac{P_s(K/2)}{2} \end{aligned}$$

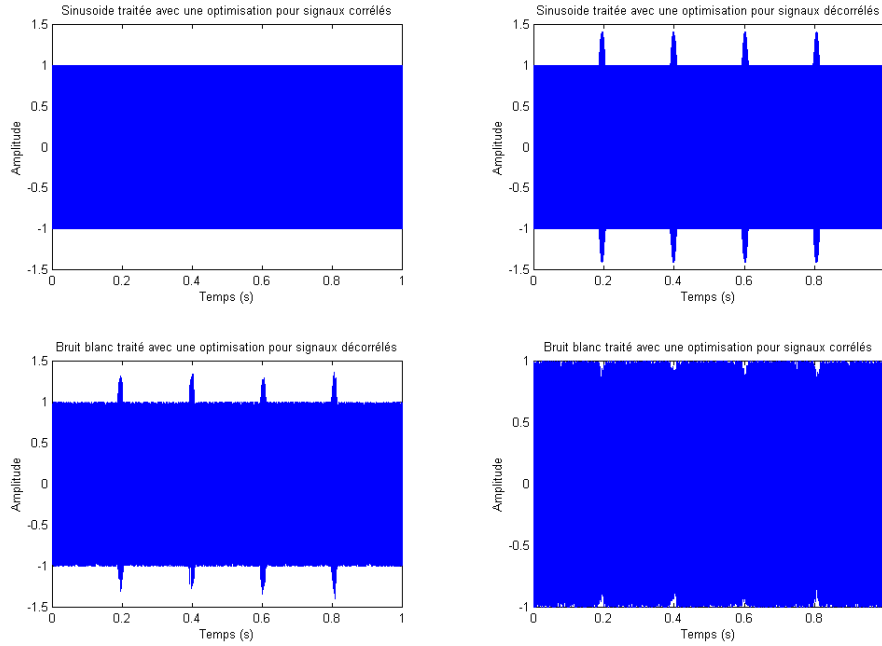


Figure 3.27 – Exemples de traitements optimisés pour des signaux corrélés et décorrés

La figure 3.27 et ainsi que les sons [70, 71, 72, 73] correspondants montrent l'exemple d'un signal sinusoïdal et d'un bruit blanc traités avec des fenêtres optimisées pour des signaux corrélés et décorrés.

La figure 3.28 montre l'enveloppe d'énergie de ces signaux.

Le traitement consiste à insérer en 4 endroits différents le signal résultant d'un mixage entre 2 signaux (sinusoïdaux ou bruités, de durée 45 ms) fenêtrés.

On peut voir que la sinusoïde traitée avec des fenêtres optimisées pour des signaux corrélés possède des enveloppes d'amplitude et d'énergie constante, ce signal ne peut donc pas être distingué de l'original (son [70]). Cependant, avec des fenêtres optimisées pour des signaux décorrés, on remarque visuellement et auditivement une modulation nette de l'amplitude et de l'énergie, (son [71]).

Pour le bruit blanc, on remarque visuellement, dans le traitement optimisé pour les signaux décorrés, une modulation d'amplitude qui n'est cependant pas audible (son [72]) car l'enveloppe d'énergie reste constante (le signal mixé possède en réalité la même énergie que tous les autres segments de même durée). Par contre, la faible modulation d'enveloppe observée dans le traitement optimisé pour les signaux corrélés se trouve être très audible (son [73]), ce qui est confirmé par l'observation de l'enveloppe d'énergie qui révèle une nette modulation.

Fondu-enchaîné adaptatif

Nous avons étudié deux cas particuliers et extrêmes de signaux exactement identiques ou totalement différents. Nous savons optimiser perceptivement la forme des fonctions de pondération pour chacun de ces deux types de signaux, mais nous avons aussi montré que ces deux optimisations ne pouvaient pas se substituer.

En pratique, les signaux que nous avons à traiter possèdent des caractéristiques qui se situent entre ces deux extrêmes.

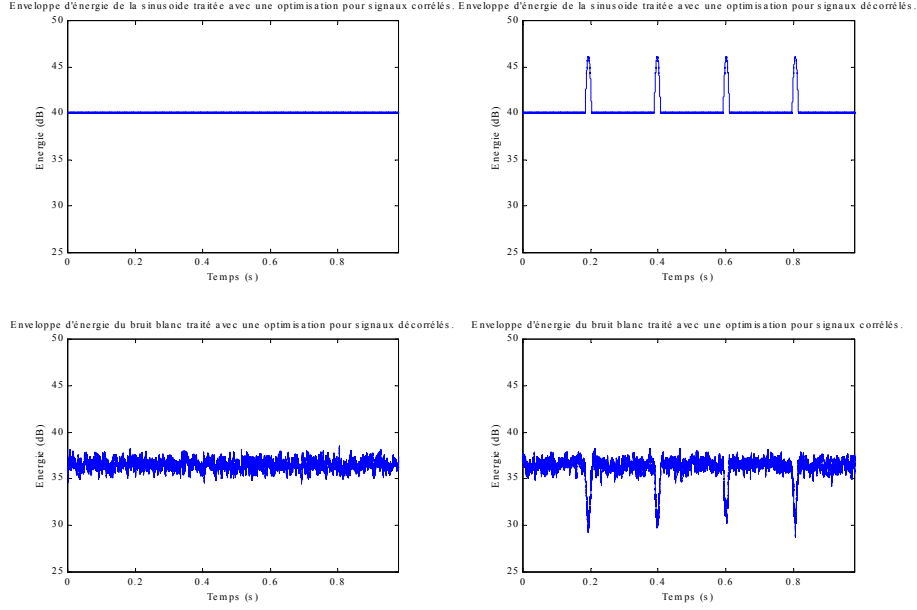


Figure 3.28 – *Enveloppe d'énergie des traitements optimisés pour des signaux corrélés et décorrés*

Nous avons constaté empiriquement qu'il est préférable d'optimiser la forme sur les signaux corrélés. En effet, une faible modulation d'amplitude (ou d'énergie) est plus perceptible sur des composantes stationnaires que sur du bruit. De plus, puisque l'on maximise la mesure de similarité (équivalente à une mesure de corrélation dans notre cas) pour trouver K , on se rapproche généralement du cas des signaux corrélés. Enfin, l'optimisation pour des signaux décorrés est valable uniquement pour du bruit totalement blanc, car pour du bruit filtré, la corrélation n'est pas nulle.

La valeur de la mesure de similarité utilisée pour extraire K nous donne une indication sur la corrélation des deux signaux. Il semble donc possible d'utiliser cette valeur afin d'engendrer une fonction de pondération mieux adaptée aux signaux à traiter.

Nous proposons pour cela une fonction de pondération $w_a(t)$ qui est une interpolation linéaire entre les fonctions optimales retenues, et qui s'exprime sous forme de l'équation suivante :

$$w_a(t) = C^\gamma w_c(t) + (1 - C^\gamma) w_d(t) \quad \forall t \in [0, K]$$

avec C la valeur de la mesure de similarité évaluée en K , et γ un coefficient à ajuster perceptivement. Pour $C = 1$ on retrouve l'optimisation pour des signaux corrélés, et pour $C = 0$ on retrouve l'optimisation pour des signaux décorrés.

La figure 3.29 montre des fenêtres de pondération optimisées pour des signaux corrélés (voir équation 3.8) et décorrés (voir équation 3.11) ainsi que deux exemples de fonctions adaptatives (voir équation 3.12) avec $C = 0,5$.

Le problème du choix de la forme semble cependant insoluble pour un signal composé simultanément de partiels et de bruit. En effet, plus K augmente (présence d'une très basse fréquence),

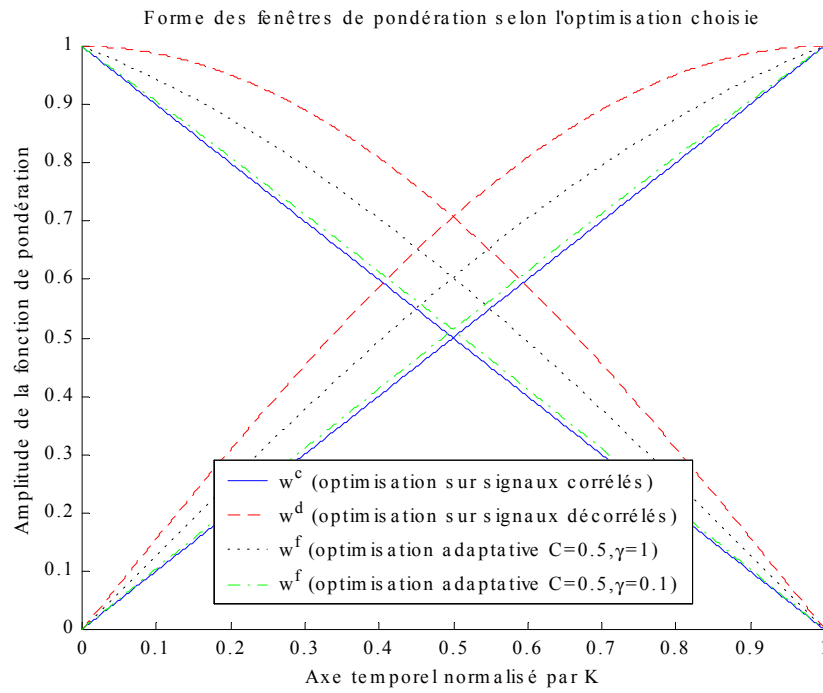


Figure 3.29 – Exemples de fenêtres de pondération

plus la modulation d'énergie (du bruit ou des partiels selon la fenêtre d'optimisation retenue) devient perceptible.

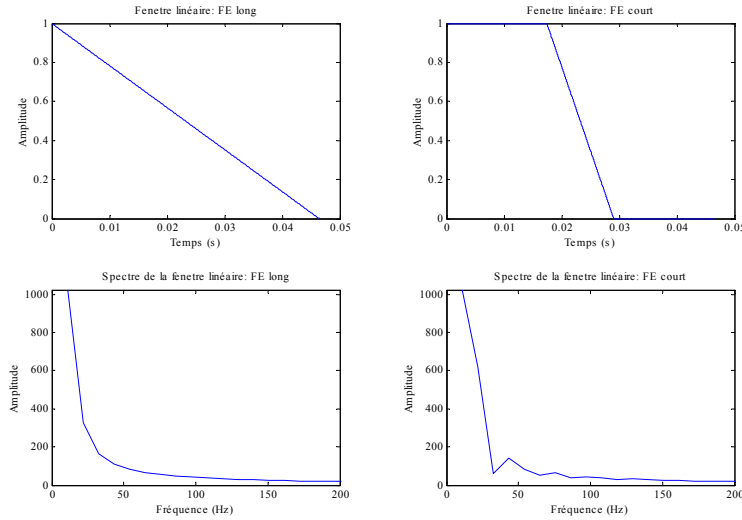


Figure 3.30 – Fenêtres de pondération de type linéaire

Durée du fondu-enchaîné

Nous définissons la durée du fondu-enchaîné FE comme étant la durée minimum pendant laquelle la fenêtre décroissante passe de la valeur 1 à la valeur 0. Comme nous imposons une contrainte de symétrie afin que la "mutation" d'un signal vers l'autre soit équilibrée, cette durée est identique pour la fenêtre croissante, et la valeur de ces deux fenêtres est identique en $FE/2$. A priori, la durée FE est indépendante de la durée K .

L'interprétation fréquentielle de la forme d'une fenêtre régulière dépend fortement de sa durée. Ainsi, comme on peut le voir dans les figures 3.30 et 3.31, pour des fenêtres de longueur totale similaire⁹, le spectre d'une fenêtre linéaire dont la décroissance FE est longue (3.30 à gauche) ressemble beaucoup plus au spectre d'une fenêtre de type Hanning à décroissance FE longue (3.31 à gauche) qu'au spectre d'une fenêtre linéaire dont la décroissance FE est rapide (3.30 à droite).

On en conclut donc que la forme du fondu-enchaîné (sa forme de décroissance ou croissance stricte) a peu d'importance comparé à sa durée, au moins d'un point de vue fréquentiel. C'est pourquoi nous nous intéressons plutôt à ce paramètre.

Dans le cas où $C(K) = 1$ ou $C(K) = 0$, nous sommes en présence d'un son exactement périodique ou d'un bruit totalement blanc. Dans ces deux cas extrêmes, si $FE = 0$ aucune discontinuité n'est mathématiquement présente ni audible. Par contre, pour $FE \neq 0$, nous avons montré précédemment que le résultat parfait dépendait de l'optimisation de la fenêtre de pondération retenue.

Dans tous les autres cas ($C(K) \neq 1$ et $C(K) \neq 0$), cette inégalité peut provenir de raisons différentes : son quasi-stationnaire, bruit coloré, transitoire, son inharmonique ou encore problème d'évaluation de la période fondamentale à cause de l'échantillonnage. Dans ces cas, l'utilisation de FE trop court produit généralement un "clic" (parfois masqué dans le cas du transitoire). D'un autre côté, l'utilisation de FE trop long mène parfois au redoublement d'un transitoire, ou la modulation d'un son quasi-stationnaire (de fréquence lentement variable) du

9. On complète la durée des fenêtres à décroissance rapide par des paliers constants égaux à 0 ou 1.

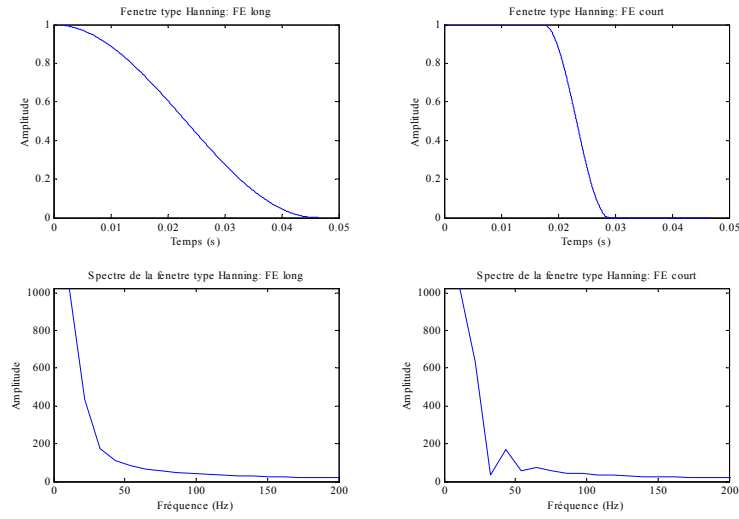


Figure 3.31 – Fenêtres de pondération de type Hanning

fait de l'interférence destructrice entre les deux ondes à mixer.

Nous pouvons utiliser la durée K sélectionnée en tant qu'indice sur les caractéristiques du signal :

- Si K est long, on suppose que les deux segments à mixer sont relativement similaires et qu'il n'existe pas de transitoire (les critères permettant de faire cette hypothèse sont évoqués à la prochaine section). Le son correspondant est généralement quasi-périodique et dans ce cas, l'utilisation de FE long donne généralement de bons résultats.

- Si K est court, on ne peut plus rien affirmer quant aux caractéristiques du signal. En effet, on peut être aussi bien en présence d'un son quasi-périodique de haute fréquence que d'un transitoire. Il est indispensable de maintenir le fondu-enchaîné sur une durée assez courte pour éviter de mixer deux événements potentiellement différents, mais assez longue quand même pour éviter la probable apparition d'un "clic".

Le choix d'une durée du fondu-enchaîné FE égale à la durée du segment inséré K semble être un bon compromis face aux nombreuses situation auxquelles nous devons faire face. Nous retiendrons donc cette égalité :

$$FE = K \quad (3.12)$$

Conclusion sur le critère de sélection de FE

L'influence de la forme des fenêtres de fondu-enchaîné sur le segment mixé est très faible comparée à l'influence de la durée de la fenêtre. Il en résulte que l'utilisation d'un fenêtrage simple permet d'optimiser les calculs sans dégrader la qualité.

Cependant, les relations énergétiques qu'elles entretiennent entre elles peuvent entraîner, selon le degré de corrélation du signal, des modulations d'amplitude d'autant plus audibles que la durée FE est grande. Pour cela, nous proposons de recourir à un fenêtrage adaptatif, bien qu'en première approximation, $FE = K$ mène généralement à des résultats satisfaisants grâce au critère de sélection de K .

3.4.6 Mesure de la distorsion

Cette section propose une évaluation objective et quantitative de la distorsion introduite par l'insertion d'un segment ne correspondant pas à un multiple de la période fondamentale du signal.

Définition

Nous nous inspirons de la mesure du taux de distorsion harmonique + bruit (TDH+N) pour définir un indice de distorsion due à l'insertion d'un segment de durée différente de la période fondamentale sur un son pur.

Pour cela, nous prenons un signal original $s_1(t) = \sin(2\pi f_0 t)$ normalisé en énergie et de durée T telle qu'une et une seule insertion de segment soit réalisée. Pour un taux de dilatation de +4,2%, nous avons donc :

$$T = \frac{K}{\alpha - 1} \simeq 24K.$$

Le signal dilaté $s_2(t)$, correspondant au signal dans lequel un segment de durée K a été inséré, est également normalisé en énergie.

Les spectres d'amplitude $S_1(f)$ et $S_2(f)$ correspondant aux signaux $s_1(t)$ et $s_2(t)$ sont ensuite calculés de manière à en extraire finement les valeurs M_1 et M_2 des amplitudes à la fréquence f_0 ¹⁰.

Le rapport de ces amplitudes élevées au carré, pour être homogène à une énergie, nous donne notre indice de distorsion, appelé ID. Exprimé en pourcentage, ID est donné par l'équation suivante :

$$ID = 100 \left(1 - \frac{M_2^2}{M_1^2} \right) \%$$

Observations

La figure 3.32 montre l'indice de distorsion calculé en fonction de la durée K du segment inséré pour un signal $f_0 = 220,5$ Hz. Le fondu-enchaîné est linéaire et de durée K . On y représente également la corrélation normalisée prise comme mesure de similarité, qui est une fonction du décalage temporel k .

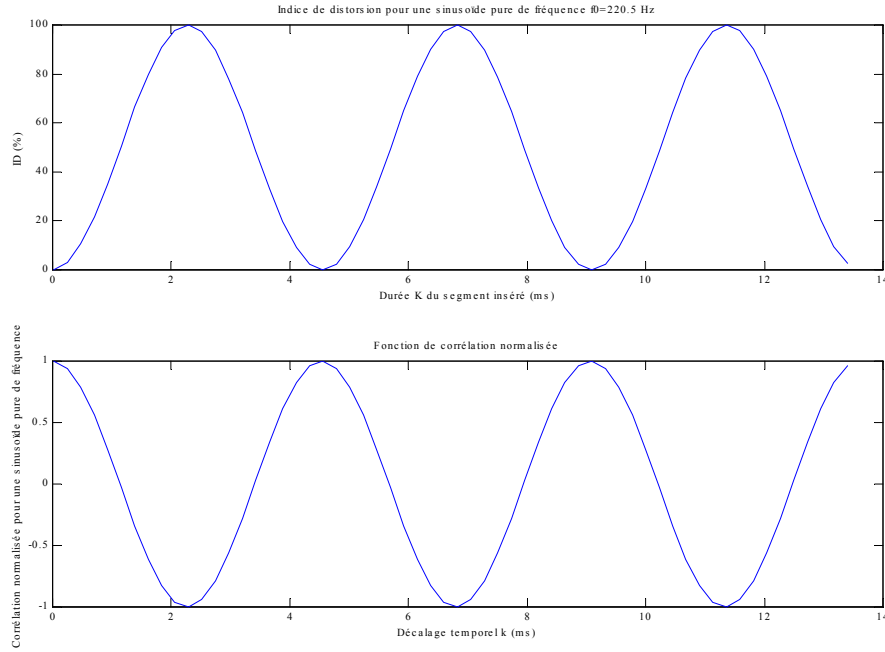
On remarque sans surprise que l'indice de distorsion est en adéquation avec la fonction de corrélation. Un indice de distorsion de 0% correspond bien à une corrélation maximale : on insère parfaitement un multiple d'une période fondamentale du signal. D'autre part, l'insertion d'un multiple impair de la moitié de la période fondamentale mène à une distorsion maximale du signal, dont le spectre est représenté en figure 3.33.

On remarque également que l'insertion d'un segment dont la durée est un multiple de la période fondamentale ne provoque aucune distorsion sur un signal stationnaire.

Pour un signal original de fréquence $f_0 = 220,5$ Hz, la période fondamentale correspond à un nombre entier d'échantillons pour une fréquence d'échantillonnage $F_e = 44100$ Hz. Dans ce cas, il est possible d'insérer un segment de durée correspondant exactement à une période fondamentale (200 échantillons dans notre cas).

La figure 3.34 montre un détail de la figure 3.32 pour une erreur de la durée K relative à la période fondamentale comprise entre -10 et +10 %, et le même détail pour une sinusoïde de

10. Un large remplissage avec des zéros ("zero-padding") est effectué avant la Transformée de Fourier Rapide pour augmenter la précision des valeurs M_1 et M_2 car f_0 ne correspond généralement pas à une fréquence discrète.

Figure 3.32 – *Indice de distorsion et fonction de corrélation : exemple général*

fréquence $f_0 = 219,95$ Hz. Cette dernière fréquence est telle que la période fondamentale correspondante possède un nombre non entier d'échantillons à notre fréquence d'échantillonnage (200,5 échantillons). On remarque que la tolérance accordée à l'estimation de la période fondamentale, évaluée aux alentours de 5%, correspond à une distorsion d'environ 2%.

On remarque donc que pour de telles fréquences, l'effet de l'échantillonnage du signal n'a pas d'incidence sur la distorsion engendrée. En d'autres termes, la période fondamentale est estimée de manière suffisamment satisfaisante malgré l'imprécision de $1/F_e \simeq 0,02$ ms.

Effet de l'échantillonnage

Cependant, pour des fréquences plus élevées, l'effet de l'échantillonnage peut introduire des défauts. La figure 3.35 montre cet effet en comparant deux signaux dont les périodes fondamentales correspondent à un nombre entier d'échantillons (10 échantillons, $f_0 = 4410$ Hz) et à un nombre non entier d'échantillons (10,5 échantillons, $f_0 = 4410$ Hz). Pour le deuxième signal, on remarque une distorsion minimale d'environ 2% pour $K = 10$ et $K = 11$, donc potentiellement audible.

Pour éviter cet effet de l'échantillonnage, la période fondamentale du signal doit être telle qu'une erreur d'un demi-échantillon reste inférieure à la tolérance relative estimée à 5%. On doit donc avoir :

$$\frac{K - T_0}{T_0} < 5\%$$

avec $K = T_0 + 0,5/F_e$, ce qui donne une fréquence limite donnée par :

$$f_0 < \frac{5}{100} \frac{F_e}{0,5} = 4410 \text{ Hz}$$

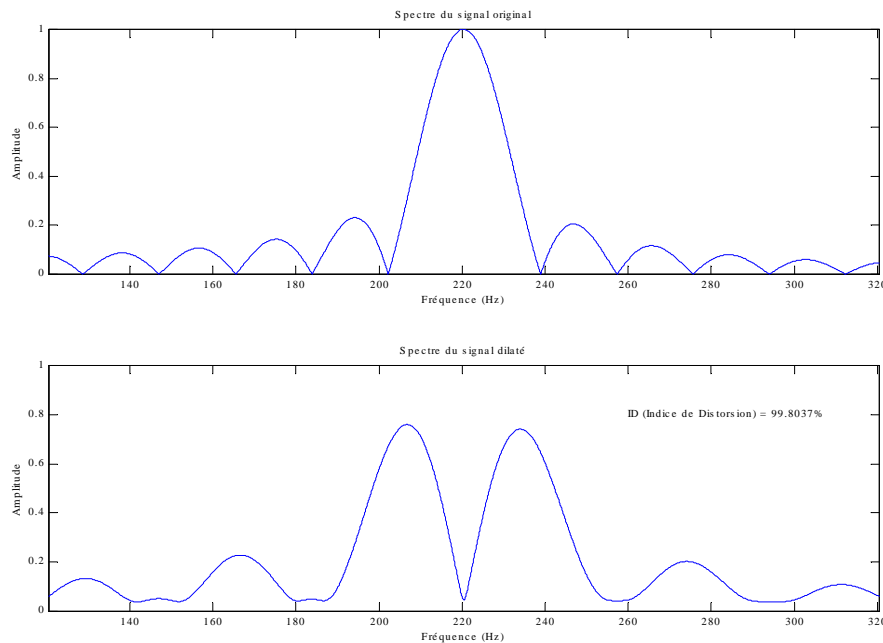


Figure 3.33 – Spectre d'un signal dont l'ID est de 100%

Il est cependant à noter qu'en pratique, des segments de durée inférieure à 0,22 ms (période correspondant à une fréquence de 4410 Hz) sont très rarement utilisés, car cela suppose qu'il n'existe aucune fréquence inférieure à 4410 Hz. En effet, si ce n'est pas le cas, les basses fréquences subiraient un changement de fréquence, car leur période se voit augmenter de la durée du segment inséré, ce qui n'est évidemment pas souhaité. De toute façon, face à une telle fréquence stationnaire, il est possible d'insérer un segment dont la longueur est un multiple de la période fondamentale sans introduire de distorsion, comme nous l'avons vu précédemment.

En conclusion, on peut affirmer que les défauts liés à l'échantillonnage dans l'estimation de la période fondamentale ne sont pas audibles tant que le segment inséré est supérieur à 0,22 ms, ce qui est généralement le cas pour tous les algorithmes temporels.

3.4.7 Traitement multicanal

Jusqu'à présent, nous nous sommes intéressés à la dilatation-p d'un signal unique. Or, à part dans le cas des films en mono, tous les films proposent une bande-son constituée de multiples canaux audio. Cette caractéristique est à prendre en compte dans le développement de l'algorithme.

Nécessité de conserver les relations de phases entre canaux

Le traitement de dilatation-p doit être effectué parallèlement sur les M_c canaux discrets (totalement indépendants des autres canaux, par opposition aux canaux dits "matricés" qui font l'objet d'une combinaison linéaire de signaux différents dans un but de réduction de données) constitutifs de la bande-son d'un film.

Les formats audio sont généralement représentés par deux nombres séparés par un point. Le premier représente la quantité de canaux discrets attribués aux signaux large bande et le second représente la quantité de canaux discrets attribués aux signaux très basse fréquence.

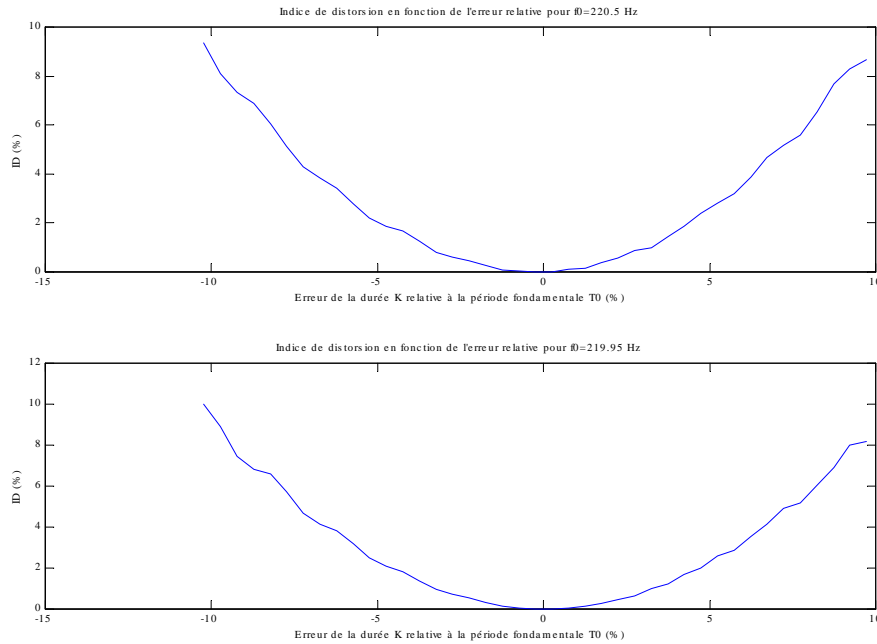


Figure 3.34 – *Indice de distorsion : détail pour $f_0=220,5$ Hz et $f_0=219,95$ Hz*

La quantité totale de canaux discrets à traiter est donc obtenu par la somme de ces nombres. Actuellement, le standard pour le DVD est de 6 (Dolby Digital 5.1 [Dol02a]), mais il possible d'en utiliser jusqu'à 8 (SDDS [SDD02]).

Signaux spatialisés

Les signaux issus de canaux discrets peuvent entretenir des rapports temporels appelés "relations de phase". Ces relations sont importantes car elles possèdent l'information nécessaire à l'oreille pour localiser des sons spatialisés lors du mixage (à l'aide de potentiomètres panoramiques stéréo ou 3D, ou bien par des effets de réverbération par exemple).

Modifier les relations de phase entre canaux revient donc à détruire tout le travail de mixage du film! Il est donc indispensable de conserver toutes ces relations de phase.

Signaux matricés

D'autre part, pour des signaux matricés (x signaux sont codés sur y canaux avec $x > y$), un décalage temporel d'un des signaux peut devenir catastrophique lors du dématricage. Par exemple, le format "LtRt" consiste à mixer les canaux gauche (L), droite (R), centre (C) et arrière (S) sur les deux canaux Lt ("Left total") et Rt ("Right total").

Schématiquement, le matricage utilisé peut être effectué de la manière suivante (des filtres, des réducteurs de bruit et une optimisation des coefficients sont utilisés en réalité) :

$$\begin{aligned} Lt &= L + \frac{1}{2}C + \frac{1}{2}S \\ Rt &= R + \frac{1}{2}C - \frac{1}{2}S \end{aligned}$$

Un dématricage dans ce cas de figure peut être le suivant (le dématricage réel est normalement aussi fonction de la corrélation existante entre Lt et Rt , nous montrons donc ici une simplification

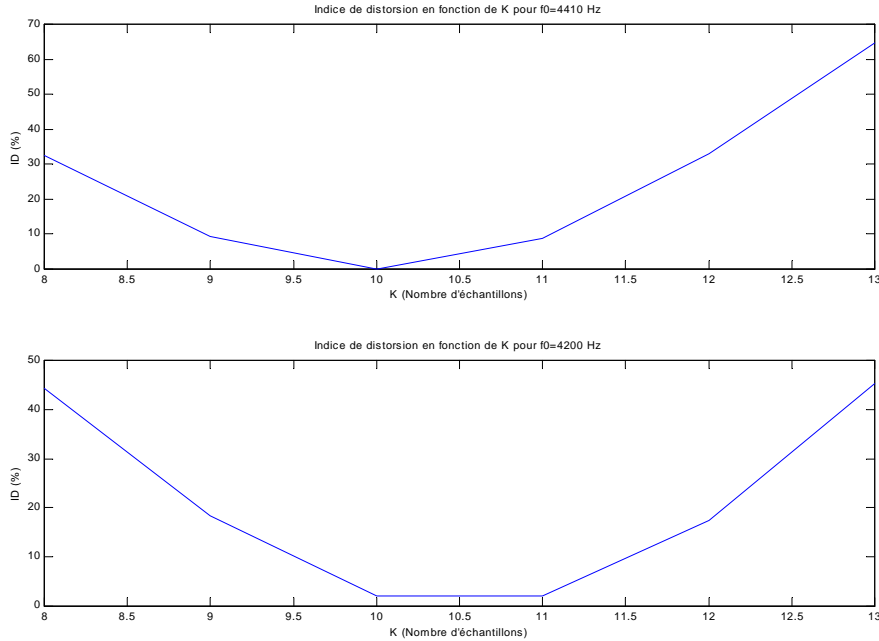


Figure 3.35 – *Indice de distorsion : détail pour $f_0=4410$ Hz et $f_0=4200$ Hz*

possible du décodage) :

$$\begin{aligned}
 L' &= Lt = L + \frac{1}{2}C + \frac{1}{2}S \\
 R' &= Rt = R + \frac{1}{2}C - \frac{1}{2}S \\
 C' &= Lt + Rt = C + L + R \\
 S' &= Lt - Rt = S + L - R
 \end{aligned}$$

On retrouve donc bien les canaux initiaux auxquels sont ajoutés une partie des signaux utilisés pour le matricage, alors considérés comme du bruit (grâce à des techniques évoluées, le rapport signal à bruit peut être supérieur à 30 dB).

Supposons maintenant qu'un décalage temporel soit introduit entre les signaux Lt et Rt . Dans ce cas, les signaux des canaux C et S , qui étaient répartis dans les canaux Lt et Rt lors du matricage, ne peuvent pas être reconstitués parfaitement lors du dématricage à cause des interférences destructive/constructives menant à un effet audible de filtrage en peigne. En effet, cela revient à mixer deux signaux identiques décalés dans le temps.

Le format "LtRt" est un exemple des nombreux systèmes de matricage utilisés dans les formats multicanaux (dont le "Dolby Digital EX" ou le "DTS ES"), mais il indique la nécessité absolue de conserver les relations de phase entre canaux.

Fonctionnement multicanal de l'HARMO

L'algorithme HARMO est basé sur une méthode temporelle. Le principe sous-jacent est donc de réaliser des décalages temporels sur le signal original. Puisque nous avons vu qu'il était indispensable de conserver les relations de phase entre canaux, il est donc essentiel que les

décalages temporels réalisés par l'algorithme, soient identiques sur tous les canaux. L'idée du fonctionnement multicanal de l'HARMO est d'utiliser un signal de référence unique, constitué par une combinaison des canaux discrets, et d'en extraire les paramètres de l'algorithme (I , K , FE) qui sont alors appliqués identiquement à tous les canaux constitutifs de la bande-son.

Le signal de référence $s_r(t)$ est donné par l'équation suivante :

$$s_r(t) = \sum_{k=1}^{M_c} P_k s_k(t)$$

où $s_k(t)$ représente le signal temporel du k^{ieme} canal et P_k le poids associé à ce canal.

Le choix des coefficients P_k est le fruit d'un judicieux compromis : prendre en compte un maximum de canaux dans le canal de référence permet d'optimiser l'extraction des paramètres pour la totalité de ces canaux, mais cela augmente également la complexité du signal à analyser dont les paramètres extraits sont valables pour le signal de référence global, mais pas obligatoirement pour chacun des signaux discrets.

Une alternative à cette technique retenue pour le fonctionnement multicanal, et exploitée dans le logiciel Pacemaker [Par02], consiste à extraire le paramètre K , non pas par maximisation de la fonction de corrélation sur un signal de référence unique $s_r(t)$ (obtenu par combinaison linéaire des signaux discrets) comme c'est schématiquement le cas avec l'HARMO, mais par une maximisation d'une combinaison linéaire des fonctions de corrélation des signaux discrets. Autrement dit, dans cette technique, il n'existe pas de signal de référence : la combinaison linéaire est effectuée après le calcul des fonctions de corrélation sur les différents canaux discrets.

Cette technique a été écartée d'une part parce qu'elle augmente beaucoup la puissance de calcul nécessaire (la fonction de corrélation, principale source de calculs, est multipliée par le nombre de canaux discrets à prendre en compte) et d'autre part parce qu'il semble préférable d'utiliser un canal de référence unique.

En effet, supposons que l'on veuille traiter un signal stéréo, et que l'on prenne comme mesure de similarité la fonction de corrélation définie par :

$$C_{xx}(\tau) = \sum_{t=0}^{N_c-1} x(t)x(t+\tau)$$

alors la durée K est donnée, dans la technique alternative, par maximisation de la somme $C_{s_1 s_1} + C_{s_2 s_2}$. Dans la technique de l'HARMO, il s'agit de maximiser C_{rr} . Or,

$$C_{rr} = C_{(s_1+s_2)} = C_{s_1 s_1} + C_{s_2 s_2} + C_{s_1 s_2} + C_{s_2 s_1}.$$

On s'aperçoit donc que les deux techniques sont équivalentes lorsque les canaux discrets s_1 et s_2 sont totalement décorrélés ($C_{s_1 s_2} = C_{s_2 s_1} = 0$), mais que dans le cas contraire la technique HARMO prend aussi en compte l'intercorrélation entre les canaux. Cela permet d'accentuer le poids des valeurs de corrélation pour lesquelles une fréquence fondamentale est à la fois présente dans les signaux s_1 et s_2 , ce qui semble être une décision raisonnable.

Des études psychoacoustiques, non réalisées dans le cadre de ce travail, pourraient être menées pour nous renseigner sur la validité des résultats théoriques attendus.

Des discussions avec des professionnels de la post-production audiovisuelle, qui sont au fait des habitudes de mixage des ingénieurs du son, nous ont mené à la conclusion qu'il était indispensable que le canal de référence prenne en compte au moins le canal central (C) dans lequel se situe tous les dialogues du film, ainsi que les canaux gauche et droite (L et R) dans lesquels se

situe la musique du film. Les canaux arrières (S) sont généralement de moindre importance, et le canal de basse fréquence peut éventuellement être traité de manière indépendante, car d'une part il est souvent totalement décorrélé des autres canaux, et d'autre part les fréquences utilisées nécessitent généralement l'insertion de segments beaucoup plus longs que les autres canaux.

3.5 Evaluations des méthodes

Une évaluation fiable des différentes méthodes de transformation-p ne peut être réalisée que par l'oreille humaine et non par une machine. D'une part il s'agit du maillon final de la chaîne sonore pour lequel le traitement est réalisé, d'autre part des mesures physiques sur le signal révèlent parfois des défauts qui sont en fait inaudibles, ou au contraire laissent passer des détails trop fins pour la mesure mais qui sont audibles.

Une évaluation idéale passe par une procédure de tests psychoacoustiques formels, mais un tel protocole est très lourd à mettre en place (nombre important de sujets, tests très longs) et ne peut être réalisé après chaque affinement d'une méthode.

Nous sommes donc contraints à évaluer les méthodes sur un corpus restreint de sons, avec un nombre faible d'auditeurs. Ces derniers sont l'auteur, ses collaborateurs, et parfois les utilisateurs. La banque de sons est discutée dans la section suivante.

Choix du taux de dilatation

Il est nécessaire de confronter des sons comparables issus des différentes méthodes à tester, ainsi nous apprécions soit la dilatation-p, soit la transposition-p, mais il est difficile de comparer un son dilaté et un son transposé. Nous décidons donc de comparer, entre eux et avec l'original, uniquement des sons dilatés-p (et non transposés-p), et ce pour plusieurs raisons :

- Tout d'abord parce que la comparaison d'un son original avec un son dilaté-p de facteur α est moins perturbante que la comparaison d'un son original avec un son transposé-p du même facteur α . Il semble en effet que l'on soit moins sensible à un changement de durée qu'à un changement de fréquence pour un même facteur de proportionnalité.
- D'autre part, la transposition-p qui nous intéresse modifie les formants, alors que la dilatation-p ne les modifie pas.
- Enfin parce qu'il s'agit du type de transformation réalisée sur la bande-son d'un film, qui correspond donc à notre problématique générale.

Nous utilisons généralement un facteur de dilatation qui correspond à une élongation ($\alpha > 1$) plutôt qu'à une contraction ($\alpha < 1$) pour les raisons que nous avons évoqué en début de section 3.4 (les défauts y sont plus audibles). Nous choisissons le plus souvent le facteur réellement utilisé pour cette application, soit +4,2%, mais les différences entre algorithmes étant parfois très subtiles, nous adoptons parfois un taux de +20% qui est le taux de traitement maximum de l'HARMO.

Construction de la banque de sons

Nous nous concentrons sur une population de sujets peu nombreuse, et sur un nombre faible d'exemples sonores. Il est donc indispensable de choisir des exemples précis et caractéristiques des sons rencontrés dans les bandes-sons de films. Nous remarquons que la plupart des sons posent peu de problèmes aux algorithmes de traitement, c'est pourquoi nous orientons nos choix vers des sons qu'empiriquement nous savons problématiques.

Ces sons proviennent de notre propre expérience par des tests préliminaires [Pal99], des sons de synthèse que l'on a construit en s'aidant de nos connaissances sur les défauts des méthodes, et également des sons issus de véritables bandes-sons de films, grâce à la collaboration des utilisateurs qui appliquent le traitement en "grandeur nature" (certains sons peuvent poser problème ponctuellement sur une bande-son commerciale).

Nous cherchons à construire une banque de sons représentative des éléments sonores présents dans une bande-son. Ces éléments sont généralement décrits en termes de composantes de base : parole, musique et bruitages [AOP02]. Cette classification donne une première idée des types de sons que l'on peut rencontrer dans un film, mais elle n'est pas assez précise pour notre besoin. Nous nous basons donc sur la typologie des sons de Schaeffer [Sch66a] que nous développons pour l'adapter aux termes de traitement du signal. Nous indiquons pour chaque type de sons les types d'artefacts éventuellement audibles en fonction des méthodes employées.

Résultats de l'évaluation

1. Tenue (quasi-stationnaire)

Un son tenu ou quasi-stationnaire possède la particularité d'évoluer lentement dans le temps.

(a) Tenue monophonique

Un son monophonique est constitué d'une seule source sonore, émettant un son soit harmonique (tous les partiels sont des multiples de la fréquence fondamentale), soit inharmonique.

i. Tenue monophonique harmonique

Sur ce type de son, où l'on distingue généralement voix parlée¹¹ (sons [21, 30]), chantée (son [74]), instruments solistes naturels (son [75]) et synthétiques (son [76]), tous les algorithmes adaptatifs donnent généralement de très bons résultats.

ii. Tenue monophonique inharmonique

Pour ce type de son, les algorithmes temporels introduisent des discontinuités sur certains partiels, légèrement audibles pour des sons naturels (son [18]), et mis en évidence dans des conditions extrêmes de son synthétique (son [32]).

(b) Tenue polyphonique

Dans un son polyphonique, plusieurs sources sonores (provenant de plusieurs instruments ou d'un seul instrument polyphonique) émettent en même temps.

Un son polyphonique pose le même type de problème qu'un son inharmonique pour les méthodes temporelles, bien que le phénomène de masquage semble atténuer les défauts aussi bien sur la voix (sons [15, 77, 78]) que sur les instruments (son [79]).

D'autre part, alors qu'un son monophonique possède rarement deux partiels très proches, ce n'est souvent plus le cas pour un son polyphonique, et cela provoque parfois des artefacts pour des méthodes fréquentielles.

2. Impulsion (transitoire)

Les impulsions ou transitoires ne trouvent nulle part dans la littérature une définition précise et tranchée. On peut dire globalement qu'il s'agit de signaux localisés précisément dans le temps et donc à spectre étendu (sons [11, 80, 81, 82]).

Dans les méthodes temporelles, le défaut principal associé à ce type de son est le redoublement lorsque le segment dupliqué est trop long. L'artefact devient inaudible lorsque le segment devient assez court.

Dans les méthodes fréquentielles, une impulsion est interprétée comme une somme de sinusoïdes synchronisées précisément. La modification de phase introduit une

11. D'un point de vue traitement du signal, la voix et certains instruments sont considérés comme quasi-stationnaire, à part pour les transitoires comme les "plosives" et les "glottales").

désynchronisation des composantes menant ainsi généralement à l'étalement audible du transitoire.

3. Itération

Un son itératif est un son constitué par un ensemble d'événements sonores. Nous interprétons ces événements comme constitutifs d'un rythme ou bien d'une modulation d'amplitude. En termes de signal, nous nous situons ici à un niveau de représentation plus élevé que les éléments de base que nous avons vu jusqu'ici.

(a) Rythme

Pour les méthodes temporelles, l'ensemble des sons constituant un rythme peut être à l'origine d'un défaut d'anisochronie (défaut de régularité rythmique) d'autant plus prononcé lorsque les segments insérés sont longs. Ainsi une séquence rythmique avec des composantes basses fréquences est un des sons les plus difficiles pour ce type de méthode (son [83]).

(b) Modulation d'amplitude

La modulation d'amplitude (son [84]) peut parfois être à l'origine de défauts aussi bien pour des méthodes temporelles que pour des méthodes fréquentielles, selon le taux de dilatation employé.

4. Sons-test

Nous intégrons à cette banque différents sons de test permettant de vérifier le fonctionnement des algorithmes.

(a) Signal constant :

Le signal constant (son [89]) permet de vérifier le fondu-enchaîné des algorithmes temporels, ainsi que les techniques OLA des méthodes fréquentielles. Pour ces dernières, une modulation d'amplitude peut se révéler en fonction du taux de dilatation.

(b) Signal monochromatique :

Le signal monochromatique (ou pur, ou sinusoïdal : son [85]) est couramment utilisé comme signal test puisque sur un tel type de son, aucune des méthodes adaptatives ne doit provoquer d'artefact audible pour des fréquences assez hautes (supérieures à environ 100 Hz). Pour des fréquences voisines de 20 Hz, les méthodes temporelles et fréquentielles peuvent introduire des défauts si elles ne sont pas bien adaptées.

(c) Bruit blanc :

Le bruit blanc (son [40]) est un signal-test parfois révélateur des méthodes fréquentielles, car un filtrage en peigne (coloration) évoluant parfois dans le temps (dynamique) apparaît dans les méthodes n'utilisant pas de verrouillage de phase. Pour les méthodes temporelles, l'utilisation d'un taux de dilatation très élevé induit la duplication répétée d'un segment bruité, générant ainsi une périodicité dans le signal perçue comme une coloration du son.

5. Sons mixtes

Tous les types de sons vus précédemment, ainsi que leurs types de défauts associés peuvent être cumulés par mixage entre ces sons, ou par l'utilisation de sons spécifiques.

(a) Inharmonique + impulsif :

Les sons de cloches (sons [68, 86, 87]) sont typiques des difficultés rencontrées pour tous les types de méthodes (présence simultanée de sons inharmoniques et d'impulsions).

(b) Polyphonique + impulsif :

Un son polyphonique complexe contenant des impulsions (sons [83, 88]) est généralement source de problèmes multiples pour les méthodes temporelles (anisochronie, discontinuité de partiels) mais aussi fréquentielles (étalement de transitoire, coloration).

Le tableau suivant récapitule les types de signaux, les sons et les artefacts souvent associés aux méthodes temporelles et fréquentielles.

Type de signal	Son	Méthodes temporelles	Méthodes fréquentielles
Harmonique	[21, 30, 76]	-	-
Inharmonique	[18, 32]	disc. partiel	-
Polyphonique	[15, 77]	disc. partiel, redoublement	coloration
Impulsif	[11]	redoublement	étalement
Itératif	[83]	anisochronie	-
Constant	[89]	-	modulation
Sinus	[85]	-	-
Bruit blanc	[40]	-	coloration

Le terme "disc. partiel" signifie une discontinuité sur un partiel, c'est-à-dire un défaut audible sur une composante d'un son harmonique ou inharmonique. Toutes les méthodes peuvent souffrir de discontinuité sur les très basses fréquences lorsqu'elles ne sont pas adaptées à des sons aussi graves.

Il est bien évident que les sons intéressants pour établir cette banque de sons-tests sont des signaux qui posent le plus de problèmes aux algorithmes. C'est pourquoi la majorité de ces sons provoquent des artefacts plus ou moins audibles selon les méthodes. Ils ont été choisis et même construits dans le but de repousser les algorithmes dans leurs derniers retranchements, ils doivent donc être considérés comme des cas extrêmes. La plupart des sons habituellement utilisés au cinéma ne sont pas aussi problématiques.

3.6 Conclusions sur les innovations algorithmiques

L'étude sur l'anisochronie nous renseigne sur les durées maximales des segments que l'on peut insérer dans un son sans que l'on détecte une irrégularité rythmique. Quel que soit le tempo, l'insertion d'un segment inférieur à 6 ms est inaudible, sous réserve qu'il ne contienne pas de transitoire audible (entendu alors comme un redoublement).

J'ai utilisé ce précieux renseignement lors de la construction de nouvelles méthodes conçues à partir de la classification proposée au chapitre 2.

Les deux méthodes temps-fréquence adaptées à l'audition (dilatation-p et transposition-p) reposent sur un type d'analyse dont le compromis entre résolution temporelle et fréquentielle reflète mieux les caractéristiques de l'oreille que ne le fait la TFCT. Il en résulte une amélioration simultanée de la transformation des sons complexes et transitoires. Pour les sons complexes, deux composantes sinusoïdales proches sont résolues par l'utilisation de filtres sélectifs à basses fréquences. Pour les transitoires, les fenêtres temporelles à hautes fréquences sont suffisamment étroites pour que l'éventuelle dispersion de l'énergie ne soit pas audible.

Les deux méthodes couplées reposent sur une décomposition temporelle du signal original en différents signaux temporels ayant leurs caractéristiques propres. J'adapte la transformation aux caractéristiques des signaux, à savoir une méthode temporelle avec insertion de segments courts pour les signaux hautes fréquences ou les signaux transitoires, et une méthode fréquentielle (ou bien temps-fréquence) pour les signaux basses fréquences ou reflétant le caractère quasi-stationnaire du signal.

Ces quatre méthodes ont montré une nette amélioration par rapport aux méthodes basiques, mais aucune d'entre elles n'a prouvé une supériorité qualitative suffisante pour satisfaire les contraintes de qualité sonore imposées.

Il est sans doute possible d'étudier plus en détails chacune de ces méthodes afin de les améliorer, mais les contraintes de délais de réalisation nécessitent de faire un choix à un moment donné sur la base des résultats acquis. Ce choix s'est porté en faveur de la méthode temporelle pour laquelle la qualité sonore semblait supérieure aux autres algorithmes disponibles.

Nous avons donc retenu l'algorithme temporel pour l'implanter sur la machine HARMO, auquel j'ai apporté des améliorations. Ces perfectionnements permettent d'obtenir le meilleur compromis entre la qualité des sons basses fréquences, l'anisochronie et la duplication des transitoires.

Chapitre 4

Conceptions matérielle et logicielle

4.1 Développement matériel

Le développement matériel a été principalement conçu et réalisé par M. Deschamps, ingénieur électronicien de GENESIS.

4.1.1 Enoncé du besoin

L'harmoniseur doit effectuer l'algorithme de transposition-p en temps réel sur un canal AES [AES92] stéréo, avec possibilité de synchroniser plusieurs machines pour pouvoir traiter les 4 canaux AES nécessaires au format SDDS. La machine doit être facile à utiliser et démarrer rapidement, l'ensemble des 4 machines synchronisées ne doit pas être trop volumineux, ce qui pousse à réaliser la machine unitaire la plus compacte possible.

L'ensemble de ces critères nous a fait écarter l'utilisation d'un ordinateur standard (PC) avec adjonction de cartes spécifiques, pour nous tourner vers une architecture spécifique à base de processeur de signal (DSP, de l'anglais "Digital Signal Processor").

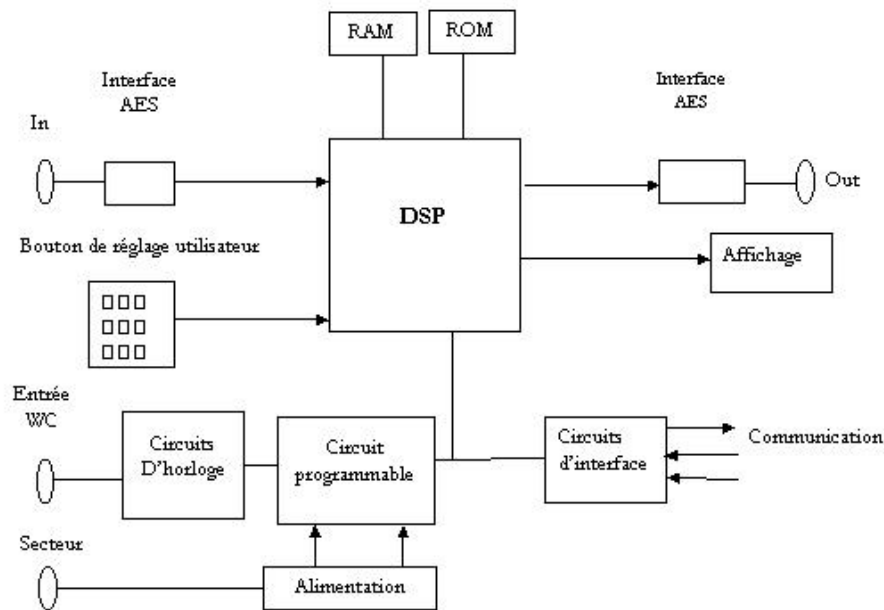
Le format mécanique usuel de ce type d'appareil est du type "rack 1U" c'est-à-dire un boîtier d'environ 43 x 31 x 4,5 cm pouvant se monter dans une baie de 19 pouces.

Les choix qui ont prévalu à l'élaboration de cette architecture reposent sur la conception d'une machine stéréo supportant une application de transposition fréquentielle (HARMO). Cependant, nous ne perdons pas de vue que cette machine pourrait être déclinée en un système 8 canaux autonome, qui accueillerait d'autres types d'applications temps-réel (algorithmes de spatialisation, de restauration...).

4.1.2 Choix de l'architecture et des composants

Le cœur du système est le processeur de signal qui permet d'effectuer les calculs. Il est associé à différents composants que sont :

- la mémoire vive et morte,
- le composant logique programmable (FPGA),
- les générateurs d'horloge,
- les circuits de communication avec l'extérieur,
- le circuit de conversion de fréquence d'échantillonnage (SRC),
- les circuits d'entrée/sortie AES,

Figure 4.1 – *Synoptique de l'architecture matérielle*

- les dispositifs d'affichage et de commande pour l'utilisateur,
- l'alimentation.

Le synoptique de l'architecture des composants est donné en figure 4.1. La figure 4.2 montre une photo de l'intérieur de la machine et la figure 4.3 celle d'une partie de la carte électronique avec ses principaux composants.

Nous détaillons dans la suite chacun de ces composants.

Processeur de signal

Selon la norme AES [AES92], les mots d'entrée et de sortie peuvent être codés sur 24 bits, donc les processeurs à mots de 16 bits sont écartés, bien qu'ils présentent un très bon rapport puissance de calcul/prix.

Motorola, avec sa série DSP 56xxx, est le seul à proposer des processeurs 24 bits à virgule fixe, qui sont très utilisés en audio. Cependant, le calcul est virgule flottante apporte un confort de développement et de mise au point des algorithmes, car on ne se soucie pas du cadrage des bits inévitable en virgule fixe. Nous préférons donc nous orienter vers des processeurs 32 bits à virgule flottante. Texas Instrument détient une très grosse part du marché des DSP avec des produits intéressants, mais dans le secteur de l'audio, c'est un processeur d'Analog Devices qui est le plus utilisé.

Notre choix se porte donc sur un processeur Analog Devices, le SHARC 21065L [Ana98]. Celui-ci fait partie de la famille SHARC 2106x. Son prix bas s'explique par une réduction de la taille de la mémoire interne. Il conserve cependant l'avantage d'une puissance performante, et la possibilité de calculer en entier ou en flottant 32 bits avec la même vitesse d'exécution.

La puissance de calcul d'un processeur est de 180 Mflops ("Mega floating point operations per second", soit le nombre de millions d'opérations à virgule flottante par seconde) en crête (sur

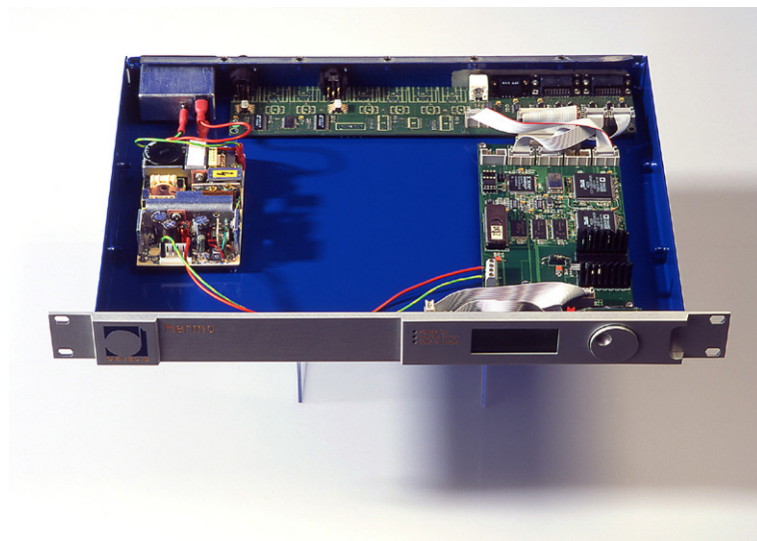


Figure 4.2 – Photo de l'intérieur de l'HARMO

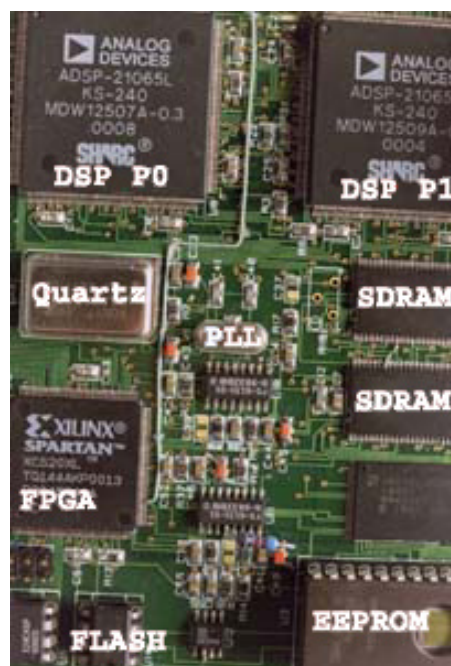


Figure 4.3 – Photo des composants principaux de l'HARMO

calcul de FFT). Les puissances réellement utilisables dépendent beaucoup de la façon d'écrire le code et du temps utilisé dans les transferts de données.

Ce processeur est conçu pour fonctionner très facilement en mode multiprocesseur, pour lequel les différents processeurs partagent les mêmes ressources externes. Le mode bi-processeur est en particulier très simple à implanter.

Ce processeur comporte quatre DMA (fonction permettant de réaliser des transferts de données sans interférer avec le processeur arithmétique) qui peuvent accéder à une mémoire externe.

L'inconvénient majeur de ce processeur réside dans sa faible taille en espace mémoire interne : si la taille réduite de la mémoire interne n'est pas très gênante pour les données (une extension est faite en mémoire externe), elle l'est fortement pour le code, car l'extension en mémoire externe n'est ni simple, ni efficace (bus 32 bits au lieu des 48 nécessaires au code).

A noter qu'Analog Devices corrige cet inconvénient avec sa version suivante du SHARC 21165, qui est plus puissante et permet facilement de stocker le code en mémoire externe (bus 48 bits). Ce processeur n'était pas disponible au moment de nos choix, et n'est pas compatible physiquement pour être monté à la place de l'autre.

Mémoires

Il est important de distinguer les types de mémoires présentes sur l'architecture électronique. Le DSP effectue ses opérations arithmétiques et logiques à partir d'emplacements mémoires appelés "registres", au nombre de 16 et numérotés de 0 à 15. Ce sont les mémoires de base de tous calculs, c'est-à-dire que les données utilisées pour toute opération sont stockées temporairement et inévitablement dans ces registres.

Parallèlement, le DSP possède une mémoire interne. Les accès à cette mémoire sont très rapides (un seul cycle processeur généralement pour accéder à une variable). Néanmoins, ce type de mémoire est relativement limitée compte tenu de notre application.

Ainsi, nous avons à disposition une mémoire vive externe de type SDRAM commune aux deux DSP, dont l'accès est moins rapide que celui de la mémoire interne, mais qui possède une capacité de 4 Mo (environ un million de mots de 32 bits).

Pour des applications temps-réel telle que la nôtre, la gestion de la mémoire est souvent critique : il est indispensable d'accéder rapidement à des données souvent utilisées. C'est pourquoi nous plaçons uniquement les volumineux buffers (ou mémoire tampon) des échantillons d'entrée/sortie en mémoire externe, et le code ainsi que les autres variables en mémoire interne.

Mémoires vives internes PM (Program Memory) et DM (Data Memory)

La mémoire interne du SHARC 21065L s'organise selon 2 blocs : le bloc nommé PM possède 36 Ko de mémoire et s'organise habituellement en trois colonnes profondes de 2048 mots de 48 bits (longueur des instructions processeur). Le bloc nommé DM possède 32 Ko de mémoire et s'organise généralement en quatre colonnes profondes de 2048 mots de 32 bits (longueur des données du processeur). La figure 4.4 schématise cette mémoire interne. Les bus pour accéder à DM et PM sont physiquement différenciés, ainsi on peut accéder simultanément à ces deux blocs-mémoire, ce qui permet d'exécuter du code sur des données en un seul cycle processeur.

Il est cependant possible de mettre des données 32 bits dans le bloc PM (ainsi que des instructions 48 bits dans le bloc DM), moyennant une réorganisation des blocs, et par conséquent, un peu de perte de mémoire. On peut de la sorte bénéficier de la rapidité de certaines instructions du processeur, notamment le MAC (Multiplication-Accumulation en un seul cycle-processeur)

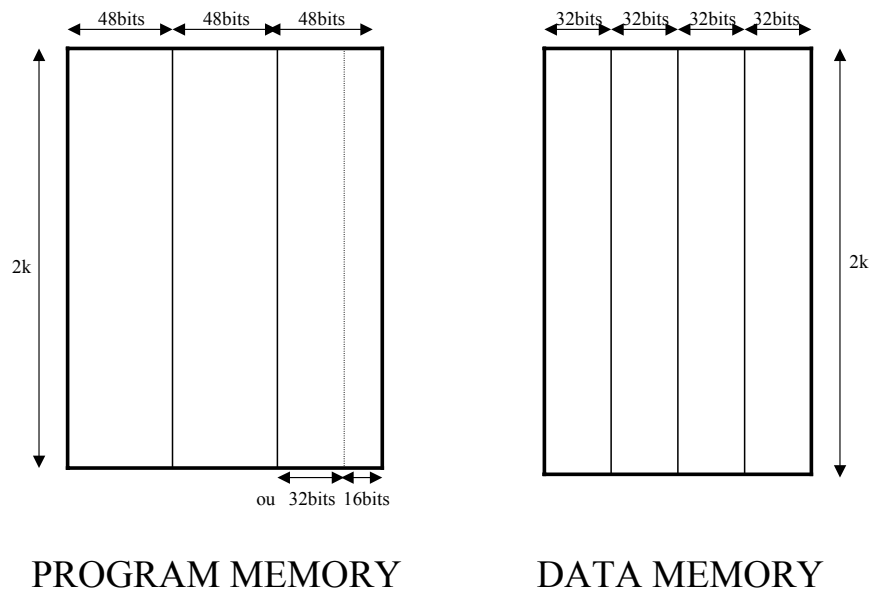


Figure 4.4 – Organisation de la mémoire interne

qui nécessite un accès simultané aux données à multiplier. C'est pourquoi il est indispensable dans notre application de réserver un espace mémoire de 2048 mots pour des données 32 bits en PM dans le processeur P0.

Mémoire vive externe SDRAM ("Synchronous Dynamic Random Access Memory")

Etant donné la taille limitée de la mémoire interne du SHARC, il est indispensable de lui adjoindre de la mémoire externe. On veut pouvoir stocker quelques secondes de son sur plusieurs canaux, soit environ 1 Méga-mot. Pour cette capacité, la SDRAM est la solution la plus économique et la plus performante, d'autant plus que le SHARC gère directement ce type de mémoire (c'est une gestion un peu compliquée que tous les processeurs ne savent pas faire). La capacité la plus standard est de 4 M x 16 bits (soit 8 Mo), deux boîtiers sont donc nécessaires pour le bus de 32 bits.

Nous justifions maintenant la nécessité d'un espace large en mémoire externe.

La méthode de dilatation-p retenue requiert un temps de latence relativement élevé : il est indispensable de connaître le futur proche du signal, donc retarder la sortie des échantillons. Ce temps de latence ne peut donc être inférieur à une centaine de millisecondes, auquel on doit ajouter le temps de calcul de la corrélation, ce qui rend le système inadapté à une utilisation musicale en directe.

Par contre, cela ne pose aucun problème aux utilisateurs des studios de post-production, pour lesquels le seul souci est de toujours conserver le synchronisme son/image (quelle que soit la fréquence d'échantillonnage ou le taux de transposition utilisé). Ils doivent donc introduire sur la chaîne visuelle une latence identique à celle de la chaîne sonore. Une latence d'exactly une seconde a été approuvée car simple à retenir.

Le temps de latence est directement proportionnel à la taille de mémoire nécessaire pour stocker les échantillons. Pour un fonctionnement à une fréquence d'échantillonnage de 48 kHz, une seconde de latence correspond à un buffer de $48000 \times 2 = 96000$ échantillons pour un signal stéréo.

Nous avons donc prévu un buffer circulaire d'entrée conséquent (100000 échantillons) et un

buffer circulaire de sortie de même taille. Ces choix confortables nous permettent d'avoir de la marge pour absorber les retards dus aux traitements, car la charge de calcul ne peut pas être répartie régulièrement dans le temps.

La mémoire SDRAM contient donc principalement les buffers d'entrée/sortie des échantillons.

Mémoire morte ou non-volatile

Le besoin en mémoire non volatile est de trois natures :

- pour stocker le "boot-strap", partie de code exécutée à la mise sous tension,
- pour stocker le programme de l'application. Cette application pouvant évoluer, l'idéal est d'avoir une mémoire programmable,
- pour mémoriser les paramètres en cours et les conserver durant la mise hors tension de la machine.

Deux types de mémoire ont été implantés :

- une UV Prom (Ultra-violet Programmable read-only memory) de capacité maximum 1 M x 8,
- une Flash de capacité 512 K x 16 bits (soit 1 Mo),

La Prom est montée sur support. Elle est effacée et programmée hors carte. Le système de développement SHARC génère par défaut un seul fichier avec le "boot-strap" et l'application. Ce fichier est transféré dans la Prom.

Il est possible de télécharger une application et de l'écrire en mémoire non volatile : c'est à cet effet qu'est implantée la mémoire Flash. Celle-ci est programmable sur la carte, et découpée en secteurs. Un secteur peut donc mémoriser le boot, si on est capable de l'écrire à la première mise en route de la machine, ou de souder la Flash déjà programmée. Dans ce cas, la Prom devient inutile.

Composant programmable

Ce composant a pour fonction principale d'effectuer les sélections d'horloge numérique, mais il a de nombreuses autres fonctions auxiliaires. Il est programmé à l'aide d'un langage spécifique, le VHDL, qui permet sur PC une simulation efficace du fonctionnement.

Le choix s'est porté sur un FPGA ("Field Programmable Gate Array"), le Xilinx XCS20XL [Xil02], qui offre une certaine souplesse et une grande capacité, mais nécessite l'ajout d'une Prom de configuration. La Prom est une Atmel 17C256 programmable hors carte.

Convertisseur de fréquence d'échantillonnage (SRC)

Pour simplifier et accélérer les développements logiciels, il est décidé d'opérer le changement de fréquence d'échantillonnage (ou SRC, de "Sampling Rate Converter") transformant la transposition-p en dilatation-p à l'aide d'un circuit spécialisé, le Crystal CS8420 [Cry98].

Ce composant semble être le seul composant du marché à pouvoir effectuer la conversion de fréquence d'échantillonnage avec une qualité professionnelle.

Celui-ci nous permet également de doter la machine d'entrées numériques asynchrones (la fréquence d'échantillonnage d'entrée peut être indépendante de celle de sortie) grâce à cette fonction SRC. Ainsi, le traitement interne est réalisé à une fréquence d'échantillonnage identique à celle de sortie, indépendamment de celle d'entrée.

Enfin, ce composant effectue l'interface entre la liaison AES (entrée et sortie du rack) et la liaison I2S (bus du SHARC). Il assure donc le rôle d'interface d'entrée/sortie au format AES.

Circuit programmable d'horloge

Pour effectuer un changement de fréquence d'échantillonnage de F_e à αF_e , le SRC a besoin de ces deux signaux d'horloge.

La valeur de la première fréquence est donnée par le signal entrant. La valeur de la seconde fréquence est déduite de la première grâce au calcul du générateur d'horloge programmable à base de "boucle à verrouillage de phase" (ou PLL, de l'anglais "Phase Locked Loop" [Bes93]). Pour cela, le facteur α doit être mis sous la forme $\frac{1000+r}{1000}$ où r est un entier compris entre -200 et +200.

Une autre fonction du circuit programmable d'horloge nécessaire est la multiplication par 256 de l'horloge d'entrée "Word Clock" cadencée à 48 kHz (destinée à synchroniser l'horloge à un signal numérique externe).

Le FS6131 d'AMI [AMI98] est un composant capable d'assumer ces deux fonctions.

Circuits de communication

Une liaison série RS232 est implantée. Elle peut servir à télécharger de nouvelles versions de logiciel à partir d'un PC. Un UART ("Universal Asynchronous Receiver/Transmitter", le contrôleur du port série) est intégré au FPGA, et un circuit d'interface RS232 est monté sur la carte électronique face-arrière.

Pour coupler les racks, on utilise les liaisons série du SHARC. Celles-ci permettent un échange relativement rapide (30 Mbit/s) ce qui est utile pour une synchronisation précise des racks.

Connectique

En interne, la place n'est pas spécialement réduite, et une connectique classique au pas de 2,54 sur nappe est adoptée.

En externe, les liaisons AES se font sur connecteurs de type XLR placés sur circuit imprimé. Pour permettre des extensions futures, il est prévu de pouvoir monter quatre entrées et quatre sorties, alors qu'une entrée et une sortie sont nécessaires dans cette version.

Les liaisons inter-racks se font sur des prises "sub-D" haute densité 26 points.



Figure 4.5 – Photo de la connectique arrière de l'HARMO

La figure 4.5 est une photo illustrant la connectique arrière, avec de gauche à droite :

- deux liens série identiques pour coupler les systèmes,
- une liaison RS232 pour la maintenance du système,
- une prise "Word Clock" pour la synchronisation d'horloge,
- une sortie au format AES,
- une entrée au format AES,
- une prise d'alimentation secteur.

Dispositif d'affichage et de commande

La figure 4.6 est une photo de la face avant de l'HARMO. On y distingue de droite à gauche :

- la commande manuelle (bouton rotatif),
- l'afficheur graphique,
- 3 LED ("Light Emitting Diod" ou Diode Electro-Luminescente).

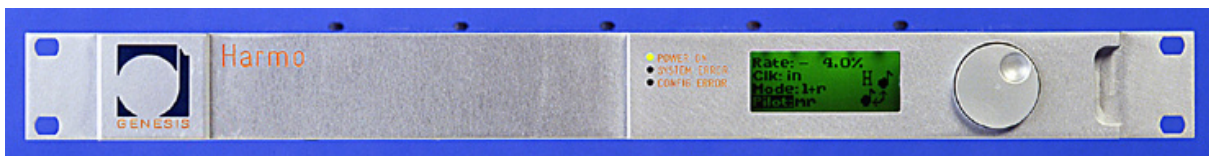


Figure 4.6 – Photo de la face avant de l'HARMO

Commande manuelle

Nous écartons l'utilisation de multiples touches dédiées à des fonctions spécifiques (utilisables uniquement pour l'harmoniseur). Des touches avec des flèches et un bouton "validation" seraient possibles, mais nous optons pour une molette unique pour plus d'originalité.

La molette est un encodeur optique rotatif, qui permet d'incrémenter ou de décrémenter un compteur, la fonction "validation" est effectuée en poussant la molette.

Afficheur

De même, nous écartons l'utilisation de multiples LED qui seraient spécifiques à l'harmoniseur pour choisir un afficheur graphique, qui permet de varier la taille et la disposition des caractères, et permet d'afficher un indicateur des bits actifs du signal.

Le choix de la hauteur du rack (36 mm utiles) est assez limitatif pour le choix de l'afficheur (afficheur graphique de 33*100 pixels). Nous avons dû développer nous-même le pilote de ce composant afin d'afficher les caractères nécessaires à notre application.

L'afficheur est translectif, c'est-à-dire qu'il est visible à la fois par son éclairage intégré et par la lumière incidente. L'affichage est complété par trois LED : une verte de bon fonctionnement, une rouge d'erreur de système et une orange d'erreur de configuration.

Alimentation

Le Crystal et la PLL s'alimentent en 5 V (Volts), le SHARC et le Xilinx s'alimentent en 3,3 V. Le Xilinx est compatible 5 V au niveau des broches, c'est-à-dire qu'un composant alimenté en 5 V peut lui être connecté. Le SHARC n'est pas compatible 5 V pour ses entrées, ce qui nécessite quelques précautions. L'alimentation secteur sort du 5 V, un régulateur à faible chute de tension (MIC 2300) fournit du 3,3 V. La SDRAM et la Flash sont alimentés en 3,3 V et se connectent directement au SHARC. La Prom et l'afficheur sont alimentés en 5 V et leur bus de données passe par le Xilinx, ce qui décharge un peu le bus du SHARC.

Un circuit de surveillance d'alimentation (MAX6303) a été choisi pour son entrée programmable. Le circuit génère un "reset" (ré-initialisation de tous les composants) si la tension passe en dessous de 3,13 V ou si le SHARC ne donne plus de signe de vie ("chien de garde").

Bilan de l'architecture

L'architecture matérielle choisie permet d'implanter à peu près toutes sortes d'applications de transformation de signaux audio stéréo, dans la limite de la puissance de calcul disponible.

Son entrée asynchrone permet de s'intégrer facilement à des environnements où plusieurs fréquences d'échantillonnage cohabitent. La qualité des composants lui assure un fonctionnement répondant aux normes de qualité exigées par les professionnels.

L'IHM et les entrées/sorties ne nécessitent pas beaucoup de puissance de calcul, mais nécessitent un espace mémoire non négligeable et mobilisent quand même le DSP (interruptions et sauvegarde de contexte), ralentissant ainsi fortement le moteur de calcul audio. Nous préférons donc découpler IHM et moteur audio afin de pouvoir dédier toute la puissance de calcul d'un DSP à l'algorithme. Pour cela, nous décidons de monter 2 processeurs SHARC sur la carte électronique, fonctionnant alors en mode bi-processeur.

4.2 Développement logiciel

Le développement logiciel a été principalement conçu par moi-même, et réalisé avec l'aide d'ingénieurs et stagiaires de GENESIS: M. Deschamps, B. Jacquier, F. Jaillet, M. Monteil et M. Adam.

Chaque système abritant 2 DSP, il est nécessaire de répartir les tâches entre ces 2 processeurs. Les fonctions spécifiques développées, écrites totalement en langage C, asm (assembleur), ou encore dans un compromis C/asm (routine d'une fonction écrite dans l'autre langage), sont les suivantes :

Processeur P1

- Automate d'état (C)
- Scrutation des ports de communication (C/asm)
- Scrutation du bouton (C)
- Programmation des PLL (asm)
- Programmation des Crystal (asm)
- Ecriture en mémoire flash (C/asm)

Processeur P0

- Automate d'état (C)
- Transmissions des flux d'échantillons (C)
- Conversions d'entiers en flottants (asm)
- Conversions de flottants en entiers (asm)
- Calcul de l'énergie (asm)
- Calcul de corrélation normalisée (asm)
- Synthèse du segment inséré (asm)
- Mise à jour des index (C)
- Tests de contraintes temps-réel (C)
- Génération de fonctions-tests (C/asm)

Certaines fonctions sont écrites entièrement ou partiellement en assembleur (asm), cela pour la simplicité de programmation, mais aussi pour optimiser à la fois le temps de calcul (énergie et corrélation) et la taille du code. D'autres fonctions sont écrites intégralement en langage C pour une meilleure lisibilité du programme ainsi qu'une rapidité d'écriture, au détriment parfois de l'optimisation en terme d'efficacité et de taille de code.

Distribution des rôles des processeurs

L'architecture électronique étant basée sur 2 DSP SHARC 21065L, il faut répartir les tâches selon un certain nombre de contraintes (nombre de DMA, de timers, liaison aux LEDS, accès aux liens inter-racks).

Le processeur nommé P1 gère l'interface homme-machine (IHM), la communication inter-racks et les tâches annexes de programmation de composants (mémoires, interface numérique, PLL).

Le processeur nommé P0 gère la gestion des flux d'entrée-sortie des échantillons et la tâche de traitement des échantillons, lourde en puissance de calcul. La communication entre les DSP s'effectue à travers des registres spéciaux (IOP Registers).

Un automate d'état dans chaque processeur permet une synchronisation des états (ARRET, INIT, MARCHE, ERREUR).

4.2.1 Processeur P1

Programmation de l'interface SDRAM

Pour pouvoir accéder à la mémoire externe SDRAM, il est nécessaire de programmer l'interface correspondante du DSP auparavant (seule manière de rendre visible la SDRAM par le processeur). Cette programmation est réalisée par une routine au début de l'application. Il en résulte qu'aucune initialisation d'une variable en mémoire externe ne doit être effectuée lors de la déclaration sous peine de ne pas être prise en compte puisque la programmation de l'interface SDRAM a lieu après. On est donc tenu d'initialiser les variables de la mémoire externe lors du déroulement du code (mise à zéro des échantillons correspondant au temps de latence par exemple pour éviter d'avoir du bruit au démarrage), ce qui surcharge malheureusement la taille du code en mémoire interne.

Programmation des interfaces numériques (Crystal CS8420)

Dans l'application de post-production audiovisuelle, il est nécessaire d'accepter en entrée un signal dont la fréquence d'échantillonnage est différente de celle à fournir en sortie (voir "Ajustement des fréquences d'échantillonnage" dans la section 4.2.1).

Nous avons donc opté pour des entrées numériques asynchrones (la fréquence d'échantillonnage en entrée est indépendante de celle en sortie) réalisé grâce à l'utilisation d'un SRC ("Sampling Rate Converter") en entrée. L'algorithme retenu étant basé sur une dilatation-p du signal suivi d'un rééchantillonnage, nous avons donc également besoin d'un second SRC pour effectuer cette fonction. Il en résulte l'intégration de deux interfaces numériques stéréo [Cry98].

La figure 4.7 schématise les flux d'échantillons sonores, d'horloge et indique les fréquences d'échantillonnage utilisées.

La programmation du SRC consiste à transmettre des fonctions simple de fonctionnement (initialisation, remise à zéro, mode silencieux, contrôle...). Pour cela, on établit un protocole de communication entre le DSP et le SRC, permettant la lecture et l'écriture de données.

Les taux de rééchantillonnage ne sont pas transmis explicitement aux SRC, mais implicitement à travers les signaux d'horloge qui leur sont fournis en entrée et en sortie. Ainsi, le taux de rééchantillonnage est "calculé" automatiquement par le SRC en fonction des signaux d'horloge fournis par les PLL. Ce sont donc elles qu'il faut programmer pour obtenir le taux désiré.

Programmation des PLL (AMI FS6131)

Comme nous l'avons vu au paragraphe précédent, ce sont les PLL qui déterminent les taux de rééchantillonnage.

La première PLL, située en entrée, est programmée de manière à ce qu'elle se synchronise avec le signal numérique entrant, si ce dernier possède une fréquence d'échantillonnage comprise entre 32 et 48 kHz. Elle permet de réduire les fluctuations de gigue (déviations de la fréquence) d'horloge.

La seconde PLL est celle qui impose la cadence au DSP et fixe la fréquence d'échantillonnage de sortie. Son signal provient soit de l'entrée AES lorsqu'entrée et sortie de l'HARMO sont

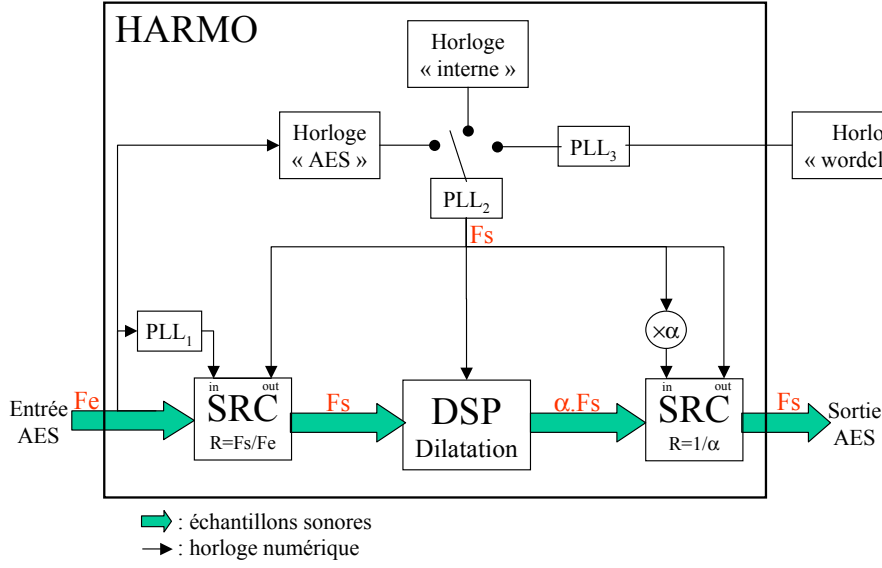


Figure 4.7 – Schéma des flux d'horloge et des fréquences d'échantillonnage

synchronisés, soit d'un quartz interne générant du 48 kHz, soit de l'entrée de synchronisation "Word Clock".

La troisième PLL est uniquement dédiée à accueillir l'horloge issue de l'entrée "Word Clock".

Ajustement des fréquences d'échantillonnage

Soit F_e la fréquence d'échantillonnage en entrée du système, et F_s la fréquence d'échantillonnage en sortie du système, nous étudions les transformations possibles en fonction de ces paramètres.

Le premier SRC est alimenté en entrée par la première PLL et en sortie par la seconde PLL. Le taux de rééchantillonnage est donné par $\frac{F_s}{F_e}$, et l'opérateur correspondant s'écrit donc :

$$R_e = R_{\frac{F_s}{F_e}}$$

Le second SRC est alimenté en entrée par la seconde PLL dont l'horloge est multipliée par α , et en sortie directement par la seconde PLL. Le taux de rééchantillonnage est donné par $\frac{F_s}{\alpha F_s} = \frac{1}{\alpha}$, et l'opérateur correspondant s'écrit donc :

$$R_s = R_{\frac{1}{\alpha}}$$

Sachant qu'il est effectué au sein du système une dilatation-p de facteur α , la transformation totale de la machine est donnée par l'opérateur suivant :

$$H_{HARMO} = R_s Dp_{\alpha} R_e = R_{\frac{1}{\alpha}} Dp_{\alpha} R_{\frac{F_s}{F_e}}$$

Premier exemple d'utilisation

D'une part, lorsque $F_s = F_e$, on a $H_{HARMO} = R_{\frac{1}{\alpha}} Dp_{\alpha} = Tp_{\alpha}$, c'est-à-dire que l'HARMO effectue une transposition-p du signal. Dans ce cas, le premier SRC n'est pas utilisé. Ce type de traitement est généralement envisagé lorsque l'utilisateur possède un convertisseur de fréquence



Figure 4.8 – Photo de l'affichage de l'HARMO avec ses 4 paramètres

d'échantillonnage qui lui permet de dilater "mathématiquement" (dilatation ET transposition) la bande originale ou la bande traitée : cette transformation convertit alors la transposition-p en une dilatation-p.

Second exemple d'utilisation

D'autre part, lorsque $F_s = \alpha F_e$, on a $H_{HARMO} = Dp_\alpha$, c'est-à-dire que l'HARMO effectue une dilatation-p du signal. Ce type de traitement ne nécessite pas de convertisseur de fréquence d'échantillonnage externe (on utilise en fait celui proposé par le SRC du circuit d'entrée). Il suffit en effet de lire la bande originale, dont la fréquence d'échantillonnage F_s est celle habituellement utilisée (soit 48 kHz dans le milieu professionnel), à la nouvelle fréquence $F_e = \frac{F_s}{\alpha}$ grâce à la fonction de "lecture à vitesse variable".

L'HARMO, dont le fonctionnement en interne est réalisé à la fréquence F_s , reçoit donc un signal de fréquence d'échantillonnage F_e grâce à ses entrées asynchrones et doit fournir en sortie un signal de fréquence d'échantillonnage F_s . La fréquence F_s est fournie à l'HARMO par l'intermédiaire de l'interface "Word Clock".

La programmation des PLL consiste donc à régler les différentes horloges, et à sélectionner la source pour l'horloge de sortie.

Programmation de l'Interface Homme-Machine (IHM)

L'IHM de l'HARMO est relativement simple puisqu'elle se résume à l'utilisation des 4 paramètres visibles sur la photo de la figure 4.8 :

Rate permet de sélectionner le taux de transposition en fréquence. Ce taux peut varier de -20,0% à +20,0% par pas de 0,1%.

Clk permet de choisir l'horloge sur laquelle se synchronisent tous les signaux AES de sortie de tous les racks. Cette horloge est choisie sur le rack maître (voir ci-dessous), et peut être :

- soit l'entrée "Word Clock" ("wc"),
- soit l'entrée AES ("aes"),
- soit l'horloge interne ("in") cadencée à 48 kHz.

Mode permet de choisir (sauf en pilote "esclave") la voie de référence : "left" (première voie du canal AES), "right" (seconde voie du canal AES) ou "left + right" (somme des deux voies du canal AES).

Pilot permet de sélectionner un rack en maître ("mr"), esclave ("sl") ou indépendant ("ind"). Le rack maître, unique, transmet à tous les autres son taux de transposition, son horloge de sortie, et uniquement aux esclaves ses paramètres algorithmiques (I et K).

L'IHM permet l'établissement d'un dialogue entre l'utilisateur et la machine. L'utilisateur donne des ordres à la machine par l'intermédiaire du bouton. Il faut donc que cette dernière soit à "l'écoute" de ces ordres par un processus de scrutation.

Une IT (interruption : processus qui interrompt ponctuellement le processeur dans sa tâche active pour effectuer une routine prioritaire) peut être utilisée, mais une réponse ultra-rapide de la part de la machine n'est pas nécessaire car démesurée par rapport à la réactivité de la perception humaine. De plus, une IT nécessite une sauvegarde du contexte (état des registres avant l'IT), ce qui requiert un nombre assez important de cycles du processeur et fait chuter les performances.

En retour, la machine émet ses messages grâce à l'afficheur graphique. La fonction de scrutation est appelée régulièrement dans le processeur P1. Elle vérifie si une action a été réalisée sur le bouton ("tourner à droite", "tourner à gauche" ou "pousser"). Si c'est le cas, elle détermine la routine à réaliser en fonction de l'état des paramètres à cet instant. Les multiples combinaisons possibles doivent être prises en compte.

La fonction d'affichage requiert des routines d'initialisation, d'affichage d'icônes (et donc création de ces icônes), de rafraîchissement et de vidéo-inverse.

4.2.2 Processeur P0

Programmation des interruptions

Une interruption matérielle (IT) consiste en un signal physique envoyé au DSP qui force l'exécution d'une partie de code spécifique à une fonction devant être réalisée de manière prioritaire. Avant d'effectuer ce code, une "sauvegarde de contexte" est réalisée de manière à restituer ce contexte après que le programme correspondant à l'IT soit terminé. Ainsi, on peut voir une IT comme l'exécution d'une partie de code en parallèle avec le code principal (en réalité, le traitement est séquentiel).

Une IT intervenant pendant un lourd calcul a tendance à ralentir fortement le processus, car les "sauvegardes de contexte" peuvent utiliser beaucoup de cycles processeur. C'est pourquoi on préfère réduire le nombre d'appels aux IT pour favoriser l'efficacité. Ces dernières sont cependant indispensables pour "charger" les registres des DMA (voir le paragraphe suivant).

Programmation des DMA (Direct Memory Access)

Un DMA est une fonction intégrée au DSP qui permet de réaliser des transferts de données numériques de manière simultanée avec le processeur arithmétique. Ainsi, à part l'étape de programmation du DMA, ce transfert ne monopolise aucune ressource de calcul.

Cette programmation consiste globalement à indiquer les adresses source et cible ainsi que la taille des blocs de données à transférer. Il y a possibilité de "chaîner" le transfert, c'est-à-dire que le DMA prend automatiquement les paramètres d'un nouveau transfert une fois qu'il a terminé le précédent. Ceci est réalisé en forçant le DMA à envoyer une IT lorsque son transfert en cours est terminé, et à prendre en compte les paramètres pour le nouveau transfert avant de relancer le DMA. Ainsi, une fois lancés, les DMA n'ont plus besoin d'intervention de la part du programme principal.

Programmation d'entrée/sortie des échantillons

Les échantillons stéréo émis par l'interface numérique d'entrée Crystal CS8420 arrivent sur le port série du DSP. Ils sont alors transférés par DMA dans un double-buffer situé en mémoire interne (DMA0 chaîné). La mémoire interne étant d'une taille restreinte comparée à nos besoins

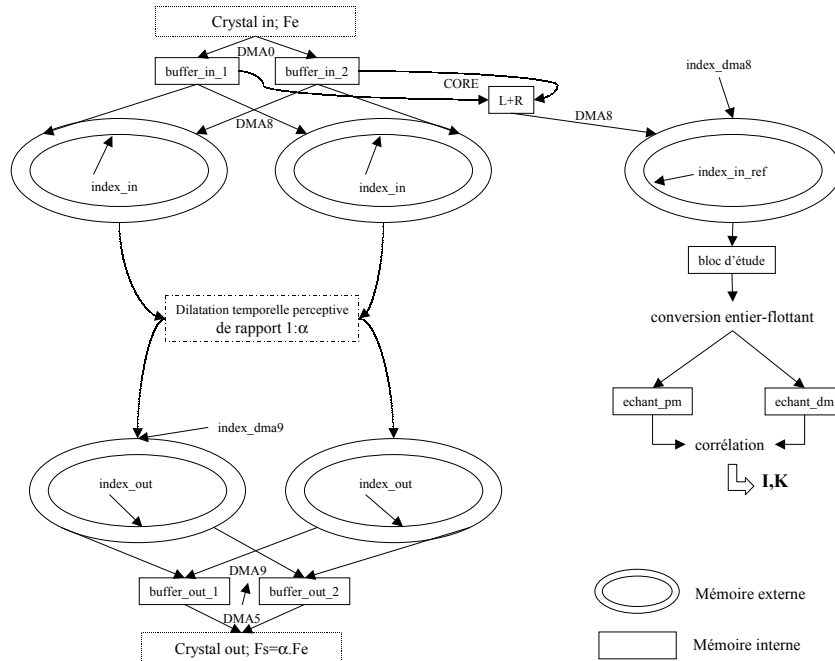


Figure 4.9 – Schéma des flux d'échantillons distinguant mémoire interne et externe

de stockage, nous sommes amenés à placer les échantillons entrant en mémoire externe. Lorsque le premier buffer interne est plein, une IT est envoyée au processeur, l'informant qu'une nouvelle série d'échantillons est disponible. Ce dernier sauvegarde son contexte avant d'effectuer la routine d'interruption adéquate :

- Reprogrammation et lancement de ce même DMA pour qu'il remplisse le second buffer.
- Incrémentation de l'index du buffer externe d'entrée.
- Désentrelacement des échantillons stéréo.
- Formatage des échantillons afin d'assurer la compatibilité des calculs (extension de signe).
- Création d'un troisième flux "left + right" en plus des 2 flux "left" et "right", qui sert de canal de référence.
- Reprogrammation et lancement du DMA externe (DMA8) pour qu'il transfère le buffer interne plein vers la mémoire externe.

Les trois flux ("left", "right" et "left + right") sont automatiquement dirigés en mémoire externe (SDRAM) vers leurs buffers circulaires respectifs, qui est le réservoir d'échantillons à partir duquel tous les calculs seront effectués. Ainsi, lorsqu'on a besoin d'échantillons, on les transfère de la mémoire externe vers la mémoire interne, afin de bénéficier de la rapidité des calculs.

Les échantillons issus de la fonction de dilatation-p sont remplacés de la même manière (par DMA) vers la mémoire externe où un processus identique à celui d'entrée est effectué, à savoir le transfert sur IT par DMA vers les double-buffers de la mémoire interne et émission via le port série vers les interfaces numériques CS8420 de sortie.

La figure 4.9 schématise les flux d'échantillons de l'algorithme de l'HARMO.

Etapes d'une itération de dilatation-p

Nous définissons une itération comme étant le processus qui, à partir d'un point origine du signal (situé sur la droite idéale), regroupe le calcul du paramètre K , le calcul du segment synthétisé, le transfert de ce segment et éventuellement (dans le cas de l'élongation) des segments K_A et K_B , le calcul de la durée R du résidu et le transfert des R échantillons résiduels vers la sortie. Ainsi, le traitement global n'est qu'une succession d'itérations. Nous décrivons dans la suite les diverses étapes d'une itération.

Calcul du paramètre K

Cette étape est effectuée si et seulement si¹ la machine n'est pas configurée en esclave (soit configuration maître ou indépendant). Elle consiste à extraire du signal de référence le paramètre de durée du segment inséré K . Le paramètre du point d'insertion est considéré comme nul : $I = 0$.

Dès qu'ils sont disponibles, les N_c premiers échantillons suivant l'origine de l'itération (c'est-à-dire après le résidu de l'itération précédente) sont placés en mémoire de programme (PM) et représentent le segment fixe. Les $2N_c$ premiers échantillons suivant l'origine de l'itération sont, quant à eux, placés en mémoire de données (DM) et représentent le segment glissant.

On prépare de la sorte les échantillons nécessaires à la fonction de corrélation afin de bénéficier de l'efficacité de la fonction MAC (multiplication-accumulation des 2 termes placés l'un en DM, l'autre en PM).

Cette fonction, basée sur le produit scalaire, est réalisée au vol, donc ne nécessite pas d'espace mémoire particulier. On conserve uniquement la valeur maximale de la mesure de similarité $C(k)$ ainsi que l'index correspondant k , à la fois pour les valeurs inférieures et supérieures à k_{limite} . Pour les valeurs supérieures à k_{limite} , une pondération (fonction linéaire calculée au vol) et un seuillage (fonction linéaire également calculée au vol) sont appliqués.

Un filtrage et un rapport d'énergie sont effectués pour détecter un éventuel transitoire pour $k > k_{limite}$.

Les $(Ordre_{FIR} + 1)$ coefficients du filtre étant stockés définitivement en mémoire PM la convolution est réalisée "in-place" (les échantillons issus du filtrage écrasent les données originales) avec les échantillons de DM en utilisant la fonction MAC.

Les énergies sont ensuite calculées sur les échantillons filtrés de DM, et le rapport est calculé. Le critère de décision nous donne enfin la longueur K du segment inséré.

A partir de ce moment, tout l'espace mémoire est disponible pour les transferts des échantillons des signaux originaux, ainsi que la création du segment inséré K_M . Les étapes suivantes sont réalisées quelle que soit la configuration : maître, indépendant ou esclave.

Calcul du segment inséré K_M

Pour $\alpha > 1$ (resp. $\alpha < 1$) le segment inséré K_M est calculé en additionnant le segment K_A (segment de longueur K situé à l'origine de l'itération) pondéré par la courbe de fondu-enchaîné croissant (resp. décroissant) et le segment K_B (segment de longueur K situé K échantillons après l'origine de l'itération) pondéré par la courbe de fondu-enchaîné décroissant (resp. croissant).

Pour cela, nous plaçons les échantillons correspondant au segment K_A en DM, et les échantillons correspondant au segment K_B en PM afin d'optimiser les cycles du processeur au moment du calcul. Nous devons convertir ces valeurs entières en valeurs flottantes.

1. Toutes les autres étapes sont effectuées quelle que soit la configuration du pilote.

Les valeurs de pondération du fondu-enchaîné sont calculées "au vol" par incrémentation de la valeur de pente de la droite (fondu-enchaîné linéaire), et le résultat (la somme des produits formant K_M) est placé en DM avant d'être converti à nouveau en valeurs entières.

Calcul de la durée R du résidu

A partir des valeurs de I et K , on est capable de déterminer la durée R du résidu, soit le nombre d'échantillons à transférer de la mémoire externe d'entrée vers celle de sortie avant d'effectuer une nouvelle itération (voir figure 3.15) :

$$R = -I + K \frac{r - 2\alpha}{\alpha - 1} \begin{cases} r = 3 & \text{pour } \alpha > 1 \\ r = 1 & \text{pour } \alpha < 1 \end{cases}$$

Transfert des échantillons

Le transfert des échantillons consiste à déplacer tels quels les échantillons non-modifiés de la mémoire externe d'entrée vers celle de sortie.

Pour $\alpha > 1$, il s'agit d'abord de transférer les $(I + K)$ premiers échantillons, avant d'insérer en sortie le segment mixé K_M de durée K , puis terminer de transférer les $(K + R)$ derniers échantillons. Ainsi, $L = (I + 2K + R)$ échantillons initiaux servent à construire $L' = (I + 3K + R)$ échantillons finaux, en accord avec la fonction de dilatation $L' = \alpha L$.

Pour $\alpha < 1$, il s'agit d'abord de transférer les I premiers échantillons, avant d'insérer en sortie le segment mixé K_M de durée K , puis terminer de transférer les R derniers échantillons. Ainsi, $L = I + 2K + R$ échantillons initiaux servent à construire $L' = I + K + R$ échantillons finaux, en accord avec la fonction de dilatation $L' = \alpha L$.

Tous les transferts passent par la mémoire interne. Nous avons mené une étude spécifique afin d'optimiser les vitesses de transfert de mémoire externe vers externe, externe vers interne et interne vers externe, et nous avons construit des routines adaptées.

4.2.3 Relations P0/P1

Nous traitons ici des relations entre les processeurs P0 et P1. Nous présentons l'automate d'état que nous avons conçu, le principe du fonctionnement multicanal avec le passage des paramètres entre processeurs au sein d'un même système, et à travers les systèmes reliés.

Automate d'état

Pour assurer une cohérence et une synchronisation entre l'interface graphique, gérée par P1, et le moteur de signal, géré par P0, il est nécessaire de mettre au point un automate d'état indiquant les différents états possibles ainsi que les transitions entre ces états. Le processeur P1 se charge de la responsabilité des transitions entre états, rendant ainsi P0 esclave de ses décisions.

On énumère les différents états créés, en détaillant sommairement leurs fonctions :

ARRET Dans cet état, la machine est arrêtée, et elle attend le signal adéquat provenant de l'IHM (validation d'un nouveau paramètre) pour passer dans l'état d'initialisation.

INIT Cet état marque la transition entre l'état d'arrêt et de marche. Il inclut une temporisation permettant de s'assurer qu'aucun problème n'est relevé avant le passage à l'état de marche.

MARCHE Dans cet état, P0 lance l'algorithme, et P1 effectue son travail de transmission de paramètres de synthèse vers les autres machines.

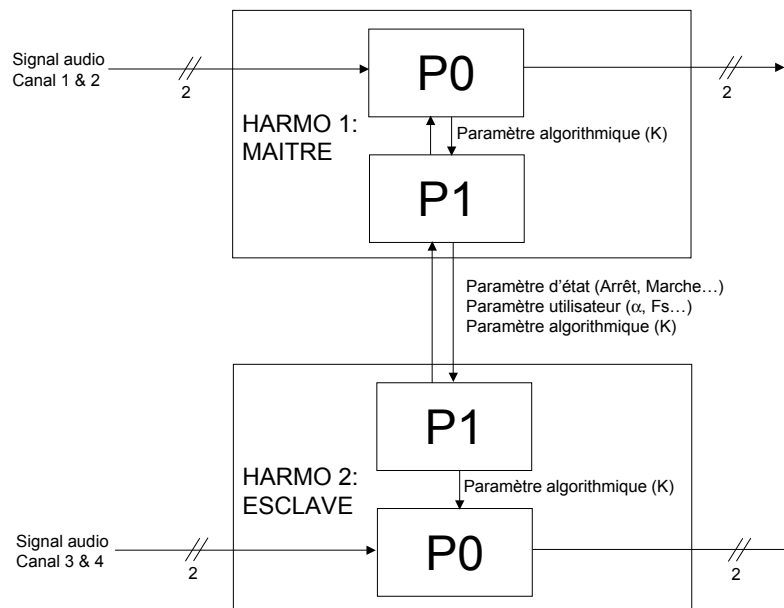


Figure 4.10 – Schéma de 2 systèmes reliés, avec passage des paramètres entre processeurs et entre systèmes

ERREUR Lorsqu'un problème est détecté, soit par P0 (problème algorithmique), soit par P1 (problème de configuration), un message d'erreur est envoyé afin que chacun des processeurs se mette dans l'état correspondant. On n'en sort que lorsqu'une action est effectuée (validation du bouton), mais on y retourne si cette action ne suffit pas à résoudre le problème.

Chacun de ces états est représenté par une combinaison unique d'allumage de diodes.

Fonctionnement multicanal

Plusieurs machines peuvent être reliées entre elles afin de fournir un traitement synchronisé, à la fois en termes d'horloge de sortie, de taux de transposition, mais aussi d'un point de vue algorithmique : en fonctionnement multicanal, les relations de phase existantes entre les canaux doivent être respectées sous peine de détruire l'image spatiale.

Cette contrainte est satisfaite grâce à la synchronisation du traitement : les paramètres d'insertion calculés sur un unique canal de référence sont utilisés de manière identique pour tous les canaux, s'assurant de la sorte du respect des phases relatives entre les canaux à chaque instant.

Le schéma 4.10 représente 2 systèmes stéréo reliés, l'un maître, l'autre esclave. On y indique les paramètres qui transitent entre processeurs au sein d'un système, ainsi que les paramètres qui transitent entre les systèmes.

Passage des paramètres entre processeurs

À l'issue de l'étape d'extraction des paramètres d'insertion réalisée sur le canal de référence du "maître" dans P0, il est nécessaire de communiquer ces paramètres à P1 qui se charge de les communiquer ensuite aux "esclaves".

Dans un souci de simplification, nous choisissons de réaliser l'étape de reconstruction du signal (transferts des échantillons et insertions) de manière unique, que l'on ait affaire à un "maître" ou à un "esclave". Ainsi, dans les deux cas, les paramètres d'insertion doivent transiter de P1 vers P0. Dans le cas du "maître", ces paramètres doivent donc aussi circuler de P0 vers P1.

D'autre part, les informations sur l'état de l'automate sont également transmis entre P0 et P1.

Cet échange d'informations est accompli à l'aide des registres IOP (de l'anglais Input/Output Processor) qui permettent un partage de mémoire très rapide (puisqu'au sein des processeurs) bien que très restreint en taille (8 registres par processeur).

Il est important de faire en sorte que chacun de ces registres ne soit autorisé en écriture que par un seul des deux processeurs, de manière à éviter une éventuelle perte d'information. Ainsi, on base la communication des paramètres sur un système de "drapeau" ("flag" en anglais), permettant de s'assurer de la réception des informations (état ou paramètres).

Passage des paramètres entre machines

Lorsque le processeur P1 du "maître" reçoit les paramètres d'insertion, il lui faut les transmettre aux autres machines qui lui sont reliées. De leur côté, les "esclaves" doivent pouvoir informer le "maître" d'un dysfonctionnement, afin que ce dernier prenne la décision d'envoyer un message d'erreur à toutes les machines pour qu'elles s'arrêtent.

Un protocole de communication est établi permettant à chacune des machines d'émettre, recevoir et traiter des signaux qui donnent une information sur les états (arrêt, init, marche, erreur), les paramètres de configuration (taux, horloge, référence, pilote) ou les paramètres d'insertions (I , K , FE).

4.3 Problématique temps-réel

Le fonctionnement en temps réel de cet algorithme n'est pas sans poser quelques problèmes, que nous détaillons dans la suite.

4.3.1 Difficultés spécifiques liées à l'algorithme

Optimisation de l'autocorrélation normalisée

La plus grosse charge en calcul de l'algorithme de dilatation-p est due à la fonction d'autocorrélation normalisée. Ce calcul peut être optimisé dans le domaine fréquentiel mais il requiert pour cela une trop grande capacité de mémoire interne pour notre application. En effet, les segments à corrélérer doivent être complétés à une puissance de 2 par un remplissage avec des zéros (appelé "zero-padding"), et un espace mémoire pour 2 FFT de cette taille est nécessaire.

D'autre part, dans l'optique d'une autocorrélation glissante, l'optimisation du calcul se révèle largement plus efficace dans le domaine temporel.

De plus, la mémoire de code nécessaire à la fonction FFT est loin d'être négligeable. Une meilleure optimisation de la puissance de calcul que la FFT semble alors être un sous-échantillonnage de la fonction d'autocorrélation (voir section 3.4.3).

C'est pourquoi nous calculons la fonction d'autocorrélation normalisée dans le domaine temporel.

La limite de puissance de calcul offerte par le DSP peut être atteinte pour un taux de transposition élevé, lorsque les segments insérés sont de petite taille. En effet, dans ce cas, les calculs de corrélation se répètent souvent. Il peut en résulter une accumulation de retards de traitements. Ces retards peuvent mener à une surcharge d'échantillons en entrée ou un manque d'échantillons en sortie.

Nous avons été confrontés à ce problème, et nous l'avons résolu en sacrifiant la qualité à des taux de transposition élevés pour lesquels le système n'est généralement pas utilisé. En pratique, nous faisons varier la valeur minimale du segment inséré de 4,5 ms à 5% jusqu'à 11 ms à 20%.

Surcharge/manque d'échantillons

La demande en puissance de calcul de l'algorithme retenu est fortement inégale au cours du temps. En effet, elle est maximale en début d'itération (au moment du calcul de la corrélation), mais elle est quasiment nulle en fin d'itération (lors du transfert des échantillons). De plus, elle dépend du signal, et bien évidemment du taux de transposition.

Lors du calcul de corrélation, le produit scalaire requiert à lui seul plusieurs millions d'opérations "multiplication/accumulation" (MAC), ce qui est effectué en plusieurs dizaines de millisecondes sur notre processeur cadencé à 60 MHz. Dans ce laps de temps, un grand nombre d'échantillons entrant doivent être stockés avant d'être traités. Comme nous travaillons avec un buffer circulaire, il est indispensable de ne pas écraser les données encore utiles, c'est pourquoi nous surveillons les positions relatives des pointeurs d'écriture et de lecture.

Lorsque le pointeur d'écriture (indiquant l'index des échantillons entrants) dépasse le pointeur de lecture (indiquant l'index des échantillons destinés au buffer de sortie) dans le buffer d'entrée, un message d'erreur est émis, provoquant l'arrêt du traitement.

Inversement, lorsque le pointeur d'écriture (indiquant l'index des échantillons provenant du buffer d'entrée) est dépassé par le pointeur de lecture (indiquant l'index des échantillons sortants) dans le buffer de sortie, un message d'erreur est également émis.

En pratique, si l'on dimensionne largement les buffers d'entrée/sortie, le contrôle sur les pointeurs devient inutile. Il faut de toute façon que l'algorithme fonctionne convenablement quel que soit le signal et quel que soit le taux de transposition utilisé. Ce contrôle est surtout nécessaire lors de la phase de débogage, mais il reste utile pour indiquer une erreur lorsque par exemple un rack "esclave" ne reçoit plus de paramètre d'insertion pour une raison quelconque.

Découpage des transferts directs

Lorsque le taux de transposition est faible, le nombre d'échantillons constituant le résidu peut être très élevé (plusieurs millions pour $\alpha = 0,1\%$). Il est dans ce cas impossible de transmettre tous ces échantillons d'un seul bloc à cause de la capacité de stockage, mais aussi et surtout à cause de la contrainte temps-réel.

C'est pourquoi les transferts d'une grande quantité d'échantillons sont réalisés par blocs de taille beaucoup plus restreinte, correspondant à la taille du buffer en mémoire PM (2048 échantillons), plus rapide pour les transferts que la mémoire DM sous certaines conditions.

4.3.2 Défauts et points forts de l'algorithme

Défauts de l'algorithme

Les utilisateurs sont parfois confrontés à un défaut auditif qu'ils appellent "bulles". Deux causes différentes semblent être à l'origine de cette sensation auditive.

La première cause se produit généralement sur la voix, et uniquement dans le sens $25 \rightarrow 24$ (vidéo vers cinéma), c'est-à-dire lorsque l'on transpose vers les aigus (élongation temporelle puis rééchantillonnage). Il s'agit de la duplication d'un long segment comportant un transitoire, qui mène à un redoublement de l'attaque. Ce cas se produit dans le cas spécifique de la parole dans un contexte bruité où sont présentes de très basses fréquences.

La deuxième cause se produit quel que soit le taux de transposition-p. Il s'agit d'une discontinuité de désynchronisation pour des sons à basses fréquences ("tenues de graves"). Elle se produit généralement lorsque des basses tenues ou des "nappes de bas-médium" sont situés dans une musique ou sur des dialogues. L'algorithme insère en effet des segments trop courts pour les grandes périodes fondamentales, et la discontinuité devient audible malgré le lissage du fondu-enchaîné.

Dans certaines séquences musicales rythmées, des irrégularités rythmiques (anisochronie) peuvent être entendus lorsque des basses fréquences sont présentes et nécessitent l'insertion d'un long segment.

Points forts de l'algorithme

Etant donné le type de fonctionnement itératif de l'algorithme, il arrive qu'un défaut (redoublement, discontinuité de partiel, défaut rythmique) provoqué sur un son à un moment donné, ne soit pas réitéré lors d'un second passage sur le même son. Cette observation explique l'amusante remarque d'utilisateurs selon laquelle ils ont l'impression que la machine "apprend ses défauts" puisque l'artefact ne se reproduit généralement pas la deuxième fois.

La machine produite est réputée pour être "transparente", c'est-à-dire qu'elle n'ajoute pas de coloration au son. On peut expliquer cette qualité par le fait que le signal modifié provient toujours de segments du signal original, et aucun filtrage n'est réalisé donc aucune coloration n'est audible.

D'autre part, au sein d'une chaîne numérique, aucune conversion analogique-numérique ou numérique-analogique n'est nécessaire (contrairement à la Lexicon 2400).

Un autre point fort du système HARMO est le synchronisme de son fonctionnement algorithmique, lui autorisant le traitement de bandes-son au format multicanal sans défauts émanant d'une modification des relations de phase entre canaux. De plus, son utilisation est extrêmement simple puisqu'aucun réglage concernant le type de son à traiter n'est nécessaire.

Chapitre 5

Conclusion

En guise de conclusion, un bilan de chacun des chapitres est donné, puis un accent est porté sur les contributions apportées et les perspectives offertes dans les domaines scientifiques et industriels.

Problématique

Un besoin spécifique a été formulé par l'industrie cinématographique européenne : il s'agit de ralentir ou d'accélérer les sons sans modifier leurs timbres, pour effectuer le transfert des bandes-son entre les formats cinéma (24 images/s) et vidéo (25 images/s). J'appelle "dilatation sous contraintes perceptives", ou plus simplement "dilatation-p", la technique associée à cette transformation, qui permet de dilater correctement (sur critères perceptifs) tous les éléments d'une bande-son (respect du timbre principalement).

Ce problème n'a pas de solution triviale tirée de la théorie du signal classique. Il ne peut être résolu qu'à travers des approches qui prennent en compte les spécificités de l'oreille (sensibilité aux composantes fréquentielles, aux transitoires), et bien qu'il soit l'objet de nombreuses recherches depuis plus de 75 ans pour des applications très variées, aucune des solutions proposées ne semble actuellement totalement satisfaisante.

L'objectif de cette thèse, à caractère industriel, est de développer un algorithme et une machine adaptée qui répondent à des contraintes technologiques (multicanal, temps-réel...) et à des exigences de qualité sonore imposées par la post-production audiovisuelle.

Classification des méthodes

Pour atteindre cet objectif, j'ai étudié la quasi-totalité des méthodes de dilatation-p proposées dans la littérature. J'en propose une classification distinguant les "méthodes temporelles" pour lesquelles des grains temporels (courts segments du signal original) sont déplacés mais pas modifiés, des "méthodes fréquentielles" pour lesquelles une modification des grains temporels est appliquée (cette modification est réalisée lors du passage temporaire dans le domaine fréquentiel, expliquant ainsi le nom donné à cette méthode).

Ce deuxième type de méthode peut être vu comme la transformation des données issues d'un banc de filtres de largeur de bande constante. Je propose dans cette classification la généralisation de ce type de méthode au cas où la répartition des fréquences des filtres n'est plus régulière mais mieux adaptée à l'oreille, que j'appelle "méthodes temps-fréquence".

Cette classification permet d'avoir une vue d'ensemble des méthodes de dilatation-p, de comprendre les relations parfois étroites existantes entre elles, et propose un point de départ à l'amélioration de ces techniques grâce au recensement et aux explications des défauts audibles.

Innovations algorithmiques et évaluations

La précédente classification me permet d'explorer quatre nouvelles voies basées, pour deux d'entre elles, sur une représentation temps-fréquence adaptée au fonctionnement de l'oreille, et pour les deux autres sur des méthodes couplées qui exploitent les avantages et évitent les défauts des deux principales classes de méthodes. Une étude sur l'anisochronie est également menée afin de déterminer la limite audible des défauts rythmiques.

Bien que les algorithmes auraient pu être optimisés, les exigences industrielles portant sur les délais de réalisation me pousse à sélectionner à un moment donné, l'algorithme qui offre les meilleurs résultats sonores. Ces évaluations sont effectuées par un jury peu nombreux mais entraîné, et avec une banque de sons spécialement élaborée pour ce problème.

La technique retenue, basée sur une méthode temporelle, est ensuite améliorée : d'une part la mesure de similarité, qui fournit la durée du segment inséré, est évaluée sur de longs segments de signaux, permettant ainsi d'insérer une longue période fondamentale (amélioration sur les sons très basses fréquences et les sons inharmoniques); d'autre part des critères basés à la fois sur les valeurs de l'autocorrélation normalisée et sur les rapports d'énergie des segments à mixer évitent l'insertion d'un segment contenant un transitoire, écartant ainsi la probabilité de percevoir un redoublement (amélioration sur les sons transitoires). Cette technique offre les meilleurs compromis entre qualité des sons basses fréquences et problèmes sur les transitoires.

Conceptions matérielle et logicielle

L'HARMO est une machine entièrement conçue par GENESIS, qui a été développée spécifiquement dans le cadre de cette étude. J'ai participé aux spécifications matérielles, et personnellement pris en charge l'implantation de l'algorithme et le développement du logiciel des processeurs de signaux numériques.

Un HARMO est un système stéréo autonome avec des entrées AES asynchrones qui réalise la transposition fréquentielle en temps-réel, pour des valeurs variant de -20% à +20% par pas de 0,1%. Son temps de latence est d'une seconde précise, quelles que soient les fréquences d'échantillonnage d'entrée et de sortie.

Plusieurs HARMO sont synchronisables en terme de taux de transposition, d'horloge numérique de sortie et surtout de paramètres de traitement permettant de respecter les relations de phase entre canaux. Cette synchronisation permet de traiter par exemple les trois canaux AES stéréo du format 5.1.

Contributions scientifiques et perspectives

La bibliographie et la classification présentées offrent une base de travail sur la dilatation-p et la transposition-p, quelles que soient les contraintes imposées. Elle révèle en outre des perspectives nouvelles quant à la mise en place de méthodes et implantations originales.

J'expose par ailleurs plusieurs méthodes innovantes : deux méthodes temps-fréquences, deux méthodes couplées, et la méthode HARMO retenue pour l'implantation.

Toutes ces méthodes apportent déjà des améliorations aux techniques présentes dans la littérature, et je pense de surcroît que leurs qualités pourraient être encore accrues : pour les méthodes temps-fréquences, des études pourraient être menées afin de trouver un équivalent du "verrouillage de phase" dans l'implantation en banc de filtres; pour les méthodes couplées, les décompositions temporelles pourraient être affinées.

L'algorithme HARMO développé spécifiquement pour le problème du transfert cinéma/vidéo peut être utilisé pour d'autres applications. Une version optimisée pour la parole est par exemple actuellement utilisée dans des expériences de sciences cognitives, pour lesquelles on étudie les réactions du cerveau à l'écoute de syllabes anormalement allongées grâce à une méthode de potentiels évoqués [YMF⁺03].

Contributions industrielles et perspectives

Le système HARMO est une réussite technologique, prouvée par son utilisation dans plusieurs studios de post-production en Europe, et la satisfaction de ses utilisateurs. Il s'agit du premier harmoniseur temps-réel adapté aux nouveaux formats multicanal comme le 5.1. Le code implanté a été conçu pour une version ultérieure huit canaux (dans une unique machine), pour laquelle de simples modifications matérielles sont nécessaires. Cette machine peut également être utilisée pour d'autres applications de traitement des signaux sonores stéréo ou multicanal. Par exemple, une application de synthétiseur piano stéréo est en cours de réalisation.

Bien qu'elle puisse être ponctuellement mise en défaut sur certains sons particuliers, la qualité sonore de l'algorithme est une réussite (se référer aux exemples sonores du CD audio accompagnant ce document, notamment la pièce musicale électroacoustique correspondants aux sons [1, 2]). Ses avantages permettent d'entrevoir la possibilité de le décliner sous forme d'insérable (ou "plugin") afin de l'utiliser dans des contextes plus variés que la post-production cinématographique.

L'annexe E présente un tableau recensant un échantillon des films qui ont été traités par l'HARMO, ainsi que des sons musicaux.

Annexe A

Dilatation et transposition avec conservation de l'énergie

La dilatation temporelle avec conservation d'énergie est donnée par l'équation suivante, où K est le terme de normalisation qui permet d'avoir égalité des normes entre le signal original et le signal dilaté :

$$D_\alpha[s](t) = K s\left(\frac{t}{\alpha}\right)$$

La norme de l'opérateur D_α est la suivante :

$$\begin{aligned} \|D_\alpha[s]\|_2 &= \sqrt{\int_{-\infty}^{+\infty} [D_\alpha[s](t)]^2 dt} \\ &= \sqrt{\int_{-\infty}^{+\infty} K^2 s^2(t/\alpha) dt} \\ &= K \sqrt{\int_{-\infty}^{+\infty} s^2(T) \alpha dT} \\ &= \sqrt{\alpha} K \sqrt{\int_{-\infty}^{+\infty} s^2(t) dt} \\ &= \sqrt{\alpha} K \|s\|_2 \end{aligned}$$

Il en résulte que l'on a conservation de l'énergie pour $K = \frac{1}{\sqrt{\alpha}}$. La dilatation temporelle avec conservation d'énergie est donc donnée par :

$$D_\alpha[s](t) = \frac{1}{\sqrt{\alpha}} s\left(\frac{t}{\alpha}\right) \quad (\text{A.1})$$

Avec une telle définition de la dilatation, la transformée de Fourier d'un signal dilaté est alors donnée par :

$$F[D_\alpha[s]](\omega) = \int_{-\infty}^{+\infty} D_\alpha[s](t) e^{-j\omega t} dt$$

$$\begin{aligned}
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\alpha}} s\left(\frac{t}{\alpha}\right) e^{-j\omega t} dt \\
&= \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{+\infty} s(T) e^{-j\omega \alpha T} \alpha dT \\
&= \sqrt{\alpha} F[s](\alpha\omega) \\
&= \sqrt{\alpha} S(\alpha\omega) \\
&= D_{\frac{1}{\alpha}}[S](\omega)
\end{aligned}$$

Il en résulte l'égalité suivante entre dilatation temporelle et transposition fréquentielle avec conservation d'énergie :

$$T_{\alpha} = D_{\frac{1}{\alpha}}. \quad (\text{A.2})$$

Annexe B

Les principaux systèmes et logiciels de transformation-p

Systèmes matériels

Processeurs d'effets indépendants

Les processeurs d'effets indépendants sont des machines autonomes possédant des entrées/sorties à un format standard et une interface utilisateur permettant le réglage des paramètres. Ils s'intègrent donc facilement dans la chaîne sonore existante de n'importe quel studio de production.

En 1975 a été commercialisé le premier harmoniseur temps-réel par la société Eventide (qui a elle-même déposé le nom "Harmonizer") : le H910. Le taux de transposition est de plus ou moins une octave.

Depuis, ce type d'algorithme se retrouve dans des produits plus récents comme l'Orville, l'Eclipse ou encore le H3000.

En 1986 est apparue la machine 2400 "Stereo Audio Compressor/Expander" commercialisée par la marque Lexicon. Il s'agit de "LA" référence en matière de dilatation-p dans les studios de post-production audiovisuelle.

Quelques unes des caractéristiques de cette machine :

- 2 entrées/sorties analogiques.
- Un traitement interne numérique.
- Un taux de dilatation/transposition variant de -25% à +33%.
- 2 modes algorithmiques : "solo" et "polyphonic".

L'algorithme utilisé dans la 2400 a été plus récemment intégré à la machine PCM 81, mais aussi sous la forme d'une carte d'extension pour la machine PCM 80 [Lex02].

En 1999, T.C. Electronic introduit sur le marché son "System 6000" [Ele02], équipé de l'algorithme VP2.

Le taux de transposition est de plus ou moins une octave.

Il s'agit de la première machine capable d'effectuer une transposition-p simultanément sur plus de 2 canaux, tout en respectant les relations de phase.

En 2002, Dolby commercialise son Model 585 "Time Scaling Processor" [Dol02b]. L'appareil est doté de 8 entrées/sorties numériques. Le taux de dilatation-p / transposition-p est de +/- 15%. La latence est réglable par l'utilisateur entre 400 et 480 ms.

C'est le premier appareil capable d'effectuer un ralentissement temporel en temps-réel grâce à son unité de stockage (limitée à 3 minutes pour huit canaux).

Stations de travail

Les stations de travail sont le coeur d'une installation de production sonore. Elles rassemblent les fonctions de stockage, d'édition, de traitement et de mixage du matériau sonore. Leur utilisation remet généralement en cause toute l'organisation de la production.

Ces stations sont généralement dotés d'outils de dilatation-p et de transposition-p internes. Nous citons ici les plus réputées.

Depuis 1985, la société Digidesign vend des systèmes d'enregistrement multipiste ("Protools"). L'algorithme "DDP-1" ("Digidesign Pitch Processor") permet de réaliser des transpositions-p dépassant 4 octaves.

En 1988, avec l'apparition de sa station de travail "SonicStudio" [Son02], Sonic Solution propose un traitement de dilatation-p nommé "TimeTwist" [Son88].

En 1996, la société E-mu commercialise une machine 8-pistes nommée "Darwin", qui est équipée dès 1997 d'un algorithme de compression/expansion temporelle [Em02].

Logiciels commerciaux

Du fait d'un rapport puissance/prix en constante augmentation et d'une qualité sonore des cartes-son convenable, les ordinateurs personnels (PC et Mac) sont actuellement beaucoup utilisés comme stations de travail dans les petites structures de production, les "home-studios" et aussi les centres de recherche.

Ce nouveau marché a permis la naissance de nombreux logiciels de dilatation-p et de transposition-p, sous forme autonome, intégré à un logiciel multipistes, d'édition sonore, de création musicale, ou encore sous forme insérable (module d'un logiciel) et donc compatible avec différents logiciels partageant le même standard.

Logiciels autonomes (ou "stand-alone")

Les logiciels autonomes de dilatation-p / transposition-p sont des programmes spécifiques à la transformation-p sonore. Ils ne requièrent que la présence d'un système d'exploitation compatible (généralement Windows ou MacOS) pour fonctionner.

Applications vocales

La société californienne Antares propose de nombreux outils de transformation de la voix, dont le célèbre "Autotune" [Ant02], disponible sous forme autonome ou insérable pour Mac et PC.

Ce dernier permet, en temps réel, de corriger des erreurs de fréquence fondamentale en remplaçant automatiquement une fausse note par la plus proche note appartenant à la gamme sélectionnée.

Celemony représente dans son logiciel "Melodyne" [Cel02] toutes les notes (durées et fréquences) d'une mélodie, ce qui permet de les replacer à des temps et des hauteurs précis en conservant (éventuellement) les formants.

Applications musicales

La société allemande Prosoniq [Pro02] propose un algorithme propriétaire de très haute qualité disponible dans ses produits autonomes "TimeFactory" (disponible sous PC et Mac) et "EZtimeStretch" (version allégée de "TimeFactory" sous PC uniquement). Cette société propose également son algorithme sous forme d'insérable nommé "TimeDesigner" pour son éditeur de sons "sonicWORX" sous Mac.

L'unique caractéristique dévoilée de cet algorithme, nommé MPEX ("Minimum Perceived Loss Time Compression/Expansion"), est "qu'il utilise un réseau de neurones artificiels pour la prédiction des séries temporelles dans le domaine de l'espace d'échelle" [Pro02].

Celui-ci fonctionne actuellement uniquement hors-temps réel, mais de manière totalement automatique (aucun paramètre n'a besoin d'être ajusté par l'utilisateur) et sur des signaux musicaux aussi bien monophoniques que polyphoniques.

Les taux de dilatation-p / transposition-p possibles varient entre -33% et +33%.

La transposition-p offre le choix entre une transformation avec ou sans modification des formants.

L'IRCAM distribue le logiciel "AudioSculpt" [Aud02], conçu par Depalle et Poirot [DP91], basé sur le moteur d'analyse-synthèse SVP (Super Vocodeur de Phase) fonctionnant lui-même sur le principe de la TFCT (Transformée de Fourier à Court Terme). La dilatation-p et la transposition-p sont réalisées après un réglage minutieux des paramètres d'analyse par l'utilisateur.

Logiciels d'enregistrement multipiste et d'édition sonore

Les enregistreurs multipistes sur disque dur ("direct-to-disk"), tout comme les éditeurs de sons, intègrent des outils de dilatation-p et de transposition-p. Nous citons les plus réputés.

Steinberg [Ste02] est une société connue des possesseurs de "home-studios". Elle produit les logiciels "Cubase", "Nuendo" et "Wavelab", tous équipés de traitement de dilatation-p et de transposition-p de qualité moyenne.

"Digital Performer" est un logiciel multipistes de la société MOTU (Mark Of The Unicorn) qui propose un algorithme de dilatation-p, de transposition-p avec modification de formants ("Standard pitch shifting"), mais aussi sans modification de formants ("PureDSP").

Sonic Foundry propose un éditeur de sons nommé "SoundForge" dans lequel est proposé un traitement de compression/expansion temporelle variable entre -50% et +500% et un traitement de transposition-p variant de plus ou moins une octave, tous deux dotés de modes spécifiques pour la musique, la parole, les instruments solo, et les percussions.

Bias propose également dans son logiciel "Peak" les transformations habituelles.

Emagic propose aussi dans son logiciel "Logic" la dilatation-p et la transposition-p, avec et sans modification de formants.

Logiciels de création musicale

De nombreux logiciels de création musicale ont vu le jour sous l'impulsion récente des musiques électroniques. Ils permettent, entre autres choses, de conformer les sons entre eux de manière à homogénéiser leurs durées et leurs rythmes mais aussi leurs fréquences.

Par exemple, Sonic Foundry propose le logiciel "ACID", et la société Arturia un logiciel baptisé "Storm".

Logiciels insérables (ou "plug-ins")

Les logiciels insérables sont de petits modules de transformation sonore capables de fonctionner uniquement au sein d'un programme appelé "hôte" (généralement un logiciel d'enregistrement multipiste ou d'édition sonore) donnant ainsi une possibilité de traitement supplémentaire à un environnement de travail ouvert.

Les formats d'insérables les plus couramment utilisés sont les formats VST ("Virtual Studio Technology" de Steinberg) et Direct-X (de Microsoft) pour les PC, et les formats (H)TDM ("(Host) Time Domain Multiplex" de Digidesign), RTAS ("Real-Time AudioSuite" de Digidesign), et MAS ("MOTU Audio System") pour les Mac.

De nombreux autres formats d'insérables existent, parmi lesquels Audio Units (Apple), DXi (Cakewalk), JACK (Linux), LADSPA (Linux), MFX (Cakewalk), OPT (Yamaha), ReWire (Propellerheads). Pour plus de détails sur ces formats, on pourra consulter [For02].

Des logiciels assurant la lecture des fichiers dans des formats communément répandus et permettant de transmettre de l'information sonore en temps réel sur internet comme RealPlayer [Rea03], Windows Media Player [Mic03], se voient dotés d'insérables assurant la dilatation-p comme c'est le cas de 2xAV de la société Enounce [Eno02].

Serato propose un insérable baptisé "Pitch'n Time" qui ne nécessite aucun réglage utilisateur, fonctionne en temps-réel. L'algorithme semble basé sur le brevet de Hoek [Hoe01].

Wavemechanics propose plusieurs insérables ("Speed", "Soundblender", "PitchDoctor", "PurePitch") dont la transposition-p permet de conserver ou non les formants.

Codes et logiciels institutionnels et gratuits

Le livre "DAFX: Digital Audio Effects" [Zöl02] donne de nombreux exemples de code réalisant différents types de dilatation-p et de transposition-p.

Logiciels basés sur des méthodes temporelles

Un algorithme de type WSOLA (voir la section 2.2.3) écrit en C par Flax est disponible dans le logiciel "MFFM Time Scale Modification for Audio" [Fla02]. Ellis propose un algorithme de type SOLAFS écrit en Matlab [Ell02b], et Birr et Cuadra en C [BC02].

Le logiciel "Pacemaker", écrit par Parviainen [Par02], est un insérable au format Winamp (un logiciel de lecture de fichier MP3 [Nul03]) basé sur une méthode temporelle.

Logiciels basés sur des méthodes fréquentielles

De nombreux logiciels basés sur la Transformée de Fourier à Court Terme (souvent appelé "vocodeur de phase") sont disponibles sur Internet, généralement accompagnés de leur code-source. On peut citer parmi eux "PVOC-EX" [Dob02], "CDP" [CDP02], "Phase Vocoder

in Matlab" [Ell02a], "PVC" [PJ02], "Soundhack" [Erb02], "Sculptor" [Scu02].

Logiciels basés sur des méthodes de modélisation

Le logiciel Lemur [Lem02], basé sur le modèle sinusoïdal de Quatieri et McAulay, est capable de réaliser une dilatation-p ou une transposition-p.

Le logiciel SMS ("Spectral Modeling Synthesis") [Ser02], basé sur le modèle "sinus+bruit" de Serra [Ser89], effectue entre autres choses les mêmes transformations.

Annexe C

Méthode PSOLA

La méthode PSOLA ("Pitch Synchronous OverLap-Add" ou addition-recouvrement synchronisé à la période fondamentale) est une technique développée pour la parole [CS86, Cha88, CM89, MC90], qui permet de réaliser simultanément des opérations de dilatation-p et de transposition-p. Nous étudions dans la suite les trois principales méthodes: TD-PSOLA ("Time-Domain PSOLA"), TDI-PSOLA ("Time-Domain Interpolation PSOLA"), FDI-PSOLA ("Frequency-Domain Interpolation PSOLA"). Nous indiquons par la suite quelques méthodes dérivées des précédentes.

Principe général de la méthode PSOLA

Le principe général de la méthode PSOLA repose sur des phases d'analyse, de transformation et de synthèse consistant à manipuler des marques de lecture et d'écriture ainsi que les grains temporels eux-mêmes. Puisque le principe est basé sur l'hypothèse d'un signal possédant un et un seul "pitch" (correspondant à la fréquence fondamentale d'un son harmonique, donc excluant tous les signaux inharmoniques et polyphoniques) et que des décisions de catégorisation de son "voisé/non voisé" sont nécessaires, nous apparentons cette méthode à une méthode paramétrique.

Il peut être fait appel uniquement à la représentation temporelle du signal, sans modification (TD-PSOLA) ou avec modification (TDI-PSOLA) du grain temporel, mais également à une représentation fréquentielle pour modifier le grain temporel (FDI-PSOLA).

L'étape d'analyse consiste à placer des marques de lecture sur le signal original, en fonction des caractéristiques locales de ses composantes (périodique, aléatoire, transitoire). Ces marques de lecture sont positionnées sur les maxima locaux d'énergie du signal, correspondant généralement aux impulsions pour les sons de type source-filtre, et plus particulièrement aux impulsions glottales dans le cas de la voix [HDdC00].

Le signal est ainsi segmenté en signaux élémentaires constitués de fenêtres se chevauchant, généralement de type Hanning, de longueur égale à deux ou quatre fois la période fondamentale et centrées sur les marques de lecture [Pee98]. Chacune des fenêtres doit être centrée sur le maximum d'énergie local.

L'étape de synthèse consiste à placer des marques d'écriture, qui, pour un son uniquement dilaté, se retrouvent espacés de la même distance que les marques de lecture, afin de conserver la fréquence fondamentale originale.

L'index de correspondance, qui est la position correspondante aux marques d'écriture sur le signal original, permet de sélectionner le signal élémentaire qui sera utilisé à cette marque

d'écriture (un signal élémentaire déjà existant (TD-PSOLA), ou interpolé soit dans le domaine temporel (TDI-PSOLA), soit dans le domaine fréquentiel (FDI-PSOLA) [Pee98]).

Marques de lecture, écriture et période fondamentale

Les caractéristiques des variables pour les méthodes PSOLA sont les suivantes :

→ Les marques de lecture sont placées aux instants suivants :

$$L_i = L_{i-1} + P_i$$

avec $P_i = P(L_i)$ la période fondamentale locale du signal autour de l'instant L_i . Il s'agit d'un processus itératif, synchronisé à la période fondamentale du signal.

→ Les marques d'écriture sont placées aux instants suivants :

$$E_j = E_{j-1} + P'_j$$

avec $P'_j = P'(E_j)$ la période fondamentale locale du signal autour de l'instant E_j . Il s'agit également d'un processus itératif synchronisé à la période fondamentale du signal, d'où le nom de la méthode [VR93].

→ Les longueurs des fenêtres équivalentes de granulation temporelle sont données par :

$$H_e = P_i$$

Soit, pour des fenêtres triangulaires ou de Hanning :

$$H = 2P_i$$

→ La période fondamentale P_i est déterminée à l'aide d'un algorithme de détection de période fondamentale.

Les formules des méthodes PSOLA sont les suivantes :

⇒ Formule de granulation temporelle :

$$g_i(t) = h_i(t)s(t + L_i) \quad (\text{C.1})$$

⇒ Formule de construction temporelle :

$$s'(t) = \sum_j g'_j(t - E_j) \quad (\text{C.2})$$

Les grains temporels $g'_j(t)$ sont soit des grains originaux (TD-PSOLA), soit des grains modifiés dans le domaine temporel (TDI-PSOLA) ou fréquentiel (FDI-PSOLA). Leur nombre est différent du nombre de grains originaux car certains sont répétés ou supprimés, pour la dilatation-p (la duplication explique l'allongement, comme pour les méthodes temporelles) comme pour la transposition-p (l'espacement entre les grains étant modifié, il faut dupliquer ou supprimer certains grains pour conserver la durée originale).

La différence entre dilatation-p et transposition-p réside dans la conservation ou non de l'espacement entre les grains qui définit ainsi la période fondamentale du signal.

Dans les parties non-voisées du signal vocal, les marques de lecture et d'écriture sont placées régulièrement. De nombreuses techniques ont été développées pour éviter les colorations du son [Cha88, Pee98, Pee01].

Marques de correspondance

Pour une dilatation-p, des marques de correspondance notées Lc_i sont déduites des marques de lecture grâce à la fonction de dilatation inverse :

$$Lc_j = D^{-1}(E_j)$$

Ces marques sont une indication pour déterminer les grains $g_i(t)$ (correspondant aux marques de lecture L_i) à partir desquels les grains $g'_j(t)$ (correspondant à la marque d'écriture E_j) sont construits :

- Pour la méthode TD-PSOLA, $g'_j(t) = g_i(t)$ tel que L_i est la marque la plus proche de Lc_j .
- Pour la méthode TDI-PSOLA, $g'_j(t)$ est un grain issu de l'interpolation temporelle entre $g_i(t)$ et $g_{i+1}(t)$ tel que $L_i < Lc_j < L_{i+1}$.
- Pour la méthode FDI-PSOLA, $g'_j(t)$ est un grain issu de l'interpolation fréquentielle entre $g_i(t)$ et $g_{i+1}(t)$ tel que $L_i < Lc_j < L_{i+1}$.

Pour une dilatation-p, on a $P'_j = P(Lc_j)$. Si l'on estime que la fréquence ne varie pas beaucoup entre 2 marques de lecture, on a alors $P'_j = P(L_i)$.

Cette méthode induit un nombre de marques de lecture et d'écriture différent. Pour la méthode TD-PSOLA, cela implique la duplication de certains grains comme l'illustre la figure C.1 avec la répétition du deuxième grain.

Pour une transposition-p, on a $P'_j = P(Lc_j)/\alpha$. L'écartement des impulsions glottales entraîne la modification de fréquence désirée. De la même manière, cela induit un nombre de marques de lecture et d'écriture différent. Pour la méthode TD-PSOLA, cela implique la duplication de certains grains comme l'illustre la figure C.2 avec la répétition du deuxième grain.

Il est évidemment possible de cumuler dilatation-p et transposition-p en une seule et unique transformation en jouant simultanément sur P'_j et sur les marques de correspondance $Lc_j = D^{-1}(E_j)$. On remarque que pour $\alpha = 2$, la méthode TD-PSOLA est identique à la méthode TDHS.

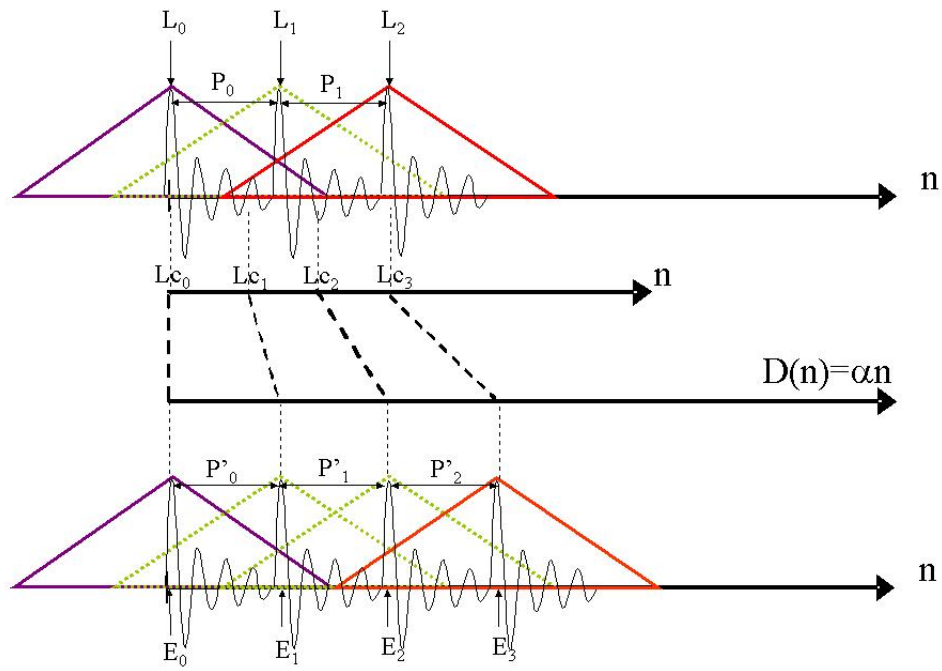
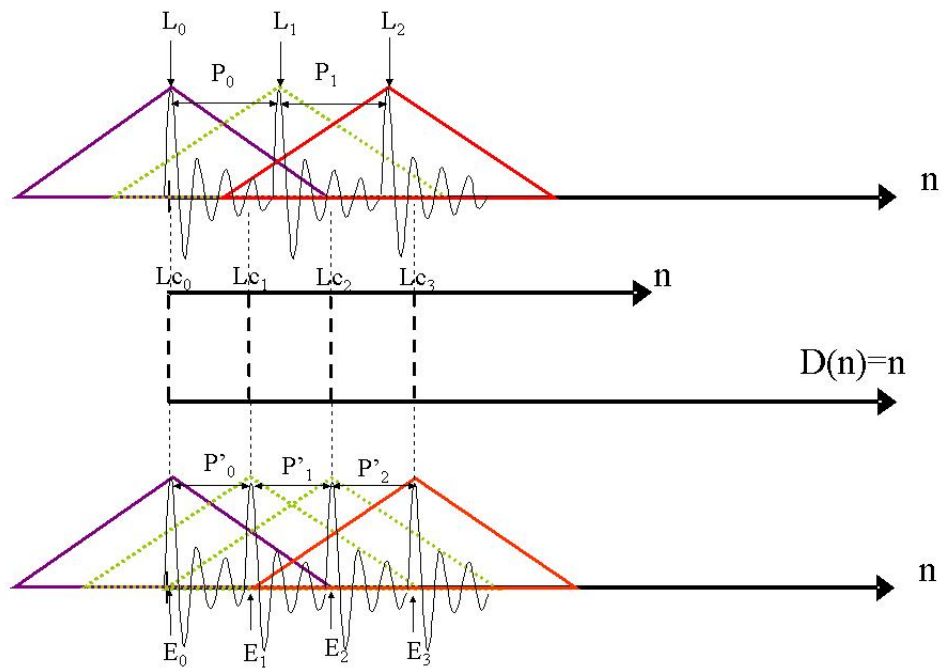
Avantages et inconvénients des méthodes PSOLA

L'avantage principal de la méthode PSOLA est qu'elle est très bien adaptée au signal vocal (l'hypothèse d'une période fondamentale unique est respectée dans les zones voisées). Elle autorise ainsi des manipulations de dilatation-p et de transposition-p de grande qualité.

La transposition-p réalisée par une méthode PSOLA conserve généralement les formants [BJ95]. En effet, on isole par l'application des fenêtres temporelles chaque impulsion glottale suivie de la résonance du conduit vocal. Comme ces grains temporels ne sont pas modifiés fondamentalement (seules des interpolations sont éventuellement appliquées), les formants sont conservés intacts, même après l'étape de transposition.

Pour des taux de dilatation supérieurs à 2, la répétition régulière de grains temporels identiques dans les parties non voisées introduit une corrélation à court terme perçue comme un bruit tonal. Une solution consiste à rendre aléatoire le spectre de phase des grains temporels non-voisés [MC90].

Des problèmes sont associés à cette méthode :

Figure C.1 – Dilatation- p par une méthode PSOLA ($\alpha = 1,5$)Figure C.2 – Transposition- p par une méthode PSOLA ($\alpha = 0,75$)

Des critères de décision permettant de caractériser les zones voisées/non voisées doivent être appliqués, et cette segmentation peut poser des problèmes.

La qualité du traitement dépend fortement du placement des marques de lecture [Kor97].

Pour un signal musical, l'hypothèse d'une seule période fondamentale n'est plus respectée et la qualité en souffre [Pee98].

La fenêtre de granulation doit être centrée assez précisément sur l'impulsion glottale, sous peine d'entendre une dégradation du signal. Lorsque le décentrage excède 30% de la période fondamentale, le son devient rauque [MC90].

Méthodes dérivées de PSOLA

De nombreuses méthodes inspirées de PSOLA ont été proposées. Parmi celles-ci, on peut citer :

LP-PSOLA : "Linear Predictive - PSOLA" [MC90].

MBR-PSOLA : "Multi-Band Resynthesis - PSOLA" [Dut93].

PIOLA : "Pitch Inflected OLA" [MRK⁺93].

MBROLA : "Multi-Band Resynthesis OLA" [DPP⁺96].

FS-OLA : "Frequency-shifted OLA" [PR99].

SINOLA : Basé sur "Sinusoidal additive" et "OLA/PSOLA" [PR99].

Annexe D

Magnétophone à têtes tournantes

Ces machines, dont le principe sous-jacent est identique à celui des appareils optiques, reposent sur la transformation d'un magnétophone classique en une machine possédant plusieurs têtes de lecture placées sur un cylindre rotatif (voir figure D.1).

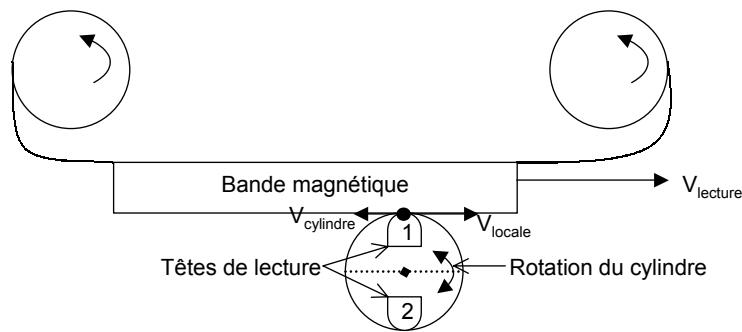


Figure D.1 – Schéma du magnétophone à têtes tournantes

Principe généralisé

Le principe généralisé de ce type de machine¹, schématisé en figure D.2, se décompose en trois étapes :

1. Enregistrement.

Le signal $s_1(t)$ est enregistré sur la bande A par un magnétophone tournant à une vitesse v_1 .

2. Transformation-p, par lecture, modification puis enregistrement.

- (a) La lecture de la bande A par le magnétophone modifié tournant à une vitesse v_2 qui, s'il était équipé d'une tête fixe permettrait d'obtenir le signal $s_{Tete\ Fixe}(t)$ donné par l'équation suivante :

$$s_{Tete\ Fixe}(t) = R_\beta[s_1](t) \quad (D.1)$$

On est donc confronté ici à une opération de lecture à vitesse variable, avec β donné par l'équation 1.5 :

$$\beta = \frac{v_1}{v_2} \quad (D.2)$$

1. Dans un souci de généralisation, l'exposé de ce principe peut paraître compliqué, mais il permet de retrouver sous un même et unique formalisme tous les cas particuliers des machines existantes.

- (b) La modification proprement dite est la combinaison de deux phénomènes : une lecture locale, par les têtes tournantes, plus ou moins rapidement qu'avec la tête fixe, et une dilatation-p due à la duplication/suppression de certains segments de bande magnétique.

La lecture locale engendre à nouveau une opération de lecture à vitesse variable :

$$s_2(t) = R_\gamma[s_{Tete\ Fixe}](t) \quad (D.3)$$

La vitesse de lecture locale v_{locale} étant donnée par la différence des vitesses v_2 et $v_{cylindre}$ ($v_{cylindre}$ est négatif si le cylindre tourne dans le sens inverse de celui de la bande), γ est donné par :

$$\gamma = \frac{v_2}{v_{locale}} = \frac{v_2}{v_2 - v_{cylindre}} \quad (D.4)$$

La dilatation-p fournit quant à elle le signal $s_3(t)$ suivant :

$$s_3(t) = D_\alpha[s_2](t) \quad (D.5)$$

α correspond au facteur de dilatation, soit la longueur globale de bande parcourue par les têtes par unité de longueur de bande :

$$\alpha = \frac{v_{locale} \cdot t}{v_2 \cdot t} = \frac{v_2 - v_{cylindre}}{v_2} = 1 - \frac{v_{cylindre}}{v_2} \quad (D.6)$$

Il ressort des équations D.4 et D.6 que :

$$\alpha = \frac{1}{\gamma} \quad (D.7)$$

Il en résulte que la modification globale due à la rotation du cylindre de lecture correspond à une tranposition-p. En effet, il découle des équations D.3, D.5, D.7 et 1.7 :

$$\begin{aligned} Dp_\alpha R_\gamma &= Dp_\alpha R_{\frac{1}{\alpha}} \\ &= Tp_\alpha \end{aligned}$$

Soit d la distance linéaire inter-tête, déterminé par la construction de la machine, ($d = \frac{2\pi R}{N}$ avec N le nombre de têtes de lecture disposées régulièrement autour de ce cylindre et R le rayon du cylindre rotatif) et $v_{cylindre}$ la vitesse linéaire en périphérie de cylindre ($v_{cylindre} = R\omega_{cylindre}$ avec $\omega_{cylindre}$ la vitesse angulaire du cylindre), la durée K des segments dupliqués/supprimés est donnée par l'équation D.8.

$$K = \frac{d}{v_1} = \frac{d}{\beta v_2} = \frac{d}{\frac{\alpha}{\delta} v_2} = \frac{dv_3}{\alpha v_2 v_4} \quad (D.8)$$

- (c) Le signal $s_3(t)$ est stocké temporairement sur une bande B enregistrée à la vitesse v_3 .

3. Lecture

Le signal stocké sur la bande B est lu par un magnétophone tournant à une vitesse v_4 , qui peut être différente de v_3 . On est donc à nouveau confronté à une opération de lecture à vitesse variable. Le signal sortant est alors donné par s_4 :

$$s_4(t) = R_\delta[s_3](t) \quad (D.9)$$

avec

$$\delta = \frac{v_3}{v_4} \quad (D.10)$$

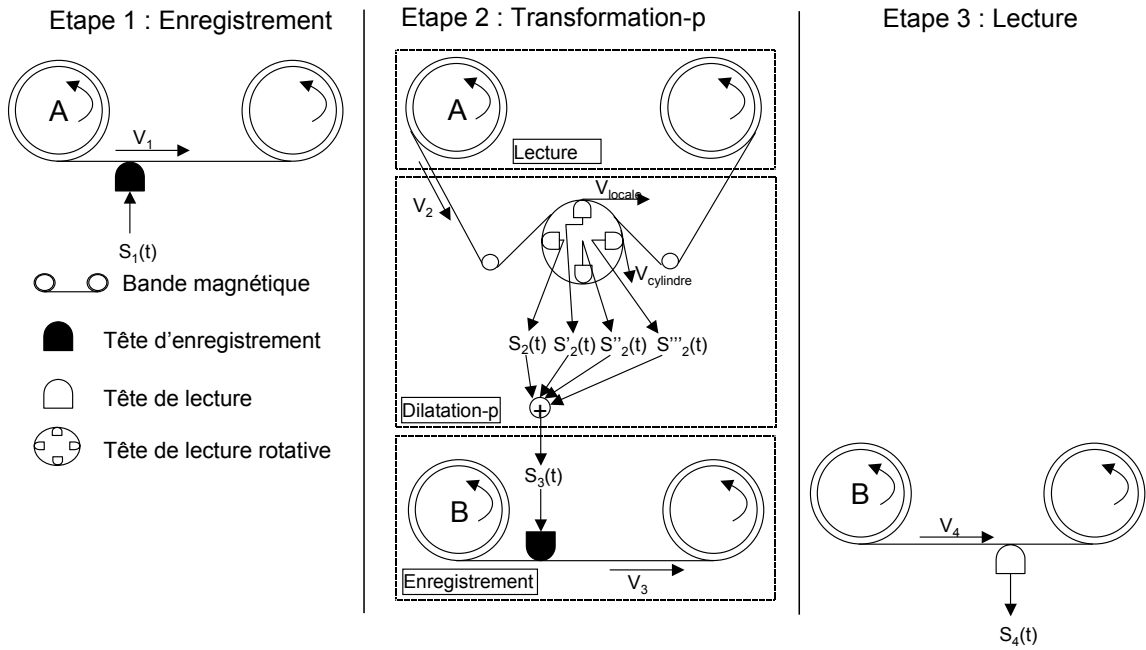


Figure D.2 – Principe de la dilatation-p par un magnétophone à têtes tournantes

On tire de ces 3 étapes le bilan suivant.

D'une part, l'équation D.6 nous indique que :

- Pour $\alpha = 1$ (aucune dilatation), $v_{cylindre} = 0$. Ce cas équivaut à une utilisation classique du magnétophone avec une tête de lecture fixe.
- Pour $\alpha > 1$ (élongation), $v_{cylindre} < 0$: le cylindre tourne dans le sens opposé au sens de défilement de la bande.
- Pour $0 < \alpha < 1$ (contraction), $0 < v_{cylindre} < v_2$: le cylindre tourne dans le sens de défilement de la bande.
- Si $v_{cylindre} = v_2$, il y a glissement nul entre la tête et la bande, donc $v_{locale} = 0$ d'où une absence de son.
- Si $v_{cylindre} > v_2$, alors $v_{locale} < 0$, il en résulte un son constitué d'une succession de fragments lus à l'envers, cas particulier de ce que peut fournir un algorithme de "brassage" comme le programme "BRAGE" [Ges98].

D'autre part, le signal s_4 lu à l'étape 3 peut s'exprimer en fonction du signal original s_1 grâce aux équations D.1, D.3, D.5, D.9 :

$$s_4(t) = R_\delta \left[Dp_\alpha \left[R_\gamma \left[R_\beta [s_1] \right] \right] \right] (t) \quad (D.11)$$

La dilatation-p est obtenue lorsque les opérateurs R annulent leurs effets de transposition, c'est-à-dire pour

$$\delta\gamma\beta = \delta\frac{1}{\alpha}\beta = 1 \quad (D.12)$$

Cette équation est vérifiée lorsque l'on a l'égalité suivante :

$$\alpha = \frac{v_1 v_3}{v_2 v_4} \quad (\text{D.13})$$

Réglages de vitesse de défilement

L'équation précédente nous montre qu'il existe plusieurs possibilités de fixer les différentes vitesses de défilement de bande pour obtenir une dilatation-p d'un taux donné. Nous exprimons ces possibilités à travers les deux exemples suivants.

Fairbanks

Dans l'appareil de Fairbanks *et al.* [FEJ54, FEJ59], les vitesses d'enregistrement et de lecture des étapes 1 et 2 sont égales ($v_1 = v_2$). Cela permet de réaliser le traitement "en direct", c'est-à-dire que l'on n'est pas obligé d'enregistrer le signal sur bande avant d'effectuer la transformation-p (cette dernière peut être exécutée au moment de l'enregistrement, c'est-à-dire en temps-réel). Le résultat de la transformation obtenue à l'étape 2 est une transposition-p. Pour réaliser une dilatation-p, l'équation D.13 nous indique que l'étape 3 doit remplir l'égalité suivante :

$$v_4 = \frac{1}{\alpha} v_3 \quad (\text{D.14})$$

Dans ce cas, l'équation D.8 nous donne :

$$K = \frac{dv_3}{\alpha v_2 v_4} = \frac{d}{v_2} = \frac{d(1 - \alpha)}{v_{cylindre}} \quad (\text{D.15})$$

On en conclut que la durée des segments dupliqués/supprimés peut être ajustée selon le type de matériau sonore à traiter en agissant sur la vitesse v_2 (liée à $v_{cylindre}$ par l'équation D.6).

"Tempo-Regulator"

Pour l'appareil "Tempo-Regulator" [Sco67], les vitesses v_3 et v_4 sont égales. Cela permet d'écouter directement la dilatation-p en sortie de l'appareil à têtes tournantes. Cependant, le signal original doit auparavant être enregistré sur bande afin d'adapter la vitesse v_2 au taux de dilatation désiré. En effet, l'équation D.13 nous donne :

$$v_2 = \frac{v_1}{\alpha} \quad (\text{D.16})$$

La durée des segments dupliqués/supprimés est alors donnée par l'équation suivante :

$$K = \frac{dv_3}{\alpha v_2 v_4} = \frac{d}{\alpha v_2} = \frac{d}{v_1} \quad (\text{D.17})$$

On en conclut que la vitesse d'enregistrement du matériau sonore à traiter détermine totalement la durée des segments dupliqués/supprimés. Il est cependant possible, dans les limites de bande passante et de fonctionnement de la machine, d'ajuster la vitesse d'enregistrement v_1 pour obtenir les longueurs de duplication/suppression désirées.

Dans cette technique, le fondu-enchaîné s'effectue de façon magnétique et électrique, par un éloignement progressif de la bande par rapport à la tête de lecture magnétique (entraînant

la diminution de l'intensité) et la sommation des courants électriques issus de chaque tête. La forme de la fenêtre de pondération assurant le fondu-enchaîné est donc déterminée par construction de l'appareil. La lecture d'un signal constant doit donner en sortie un signal constant quelle que soit la vitesse de rotation du cylindre, sans quoi une modulation d'amplitude peut être audible sur des sons stationnaires.

Bien entendu, ce système peut être employé pour effectuer une transposition-p mélangée ou non à une dilatation-p. Par exemple, si l'on souhaite cumuler une dilatation-p de taux ξ et une transposition-p de taux χ , on doit avoir la relation suivante :

$$\begin{aligned} R_\delta Dp_\alpha R_\gamma R_\beta &= Dp_\xi T p_\chi \\ &= R_\chi Dp_{\frac{\xi}{\chi}} \end{aligned}$$

d'où

$$\begin{aligned} \alpha &= \frac{\xi}{\chi} \\ \delta\gamma\beta &= \chi \end{aligned}$$

Nombre de têtes de lecture simultanées

Le nombre absolu de têtes de lecture disposées sur le cylindre rotatif importe peu dans ce type d'appareil; ce qui compte, c'est le nombre de têtes de lecture simultanément en contact avec la bande magnétique : ceci détermine le nombre de fenêtres d'écriture se chevauchant.

Supposons que le fondu-enchaîné soit inexistant ($FE = 0$). Dans ce cas, les fenêtres de granulation temporelle sont rectangulaires et de durée notée H :

$$h(t) = 1 \quad -\frac{H}{2} < t < \frac{H}{2}$$

Pour que la somme des fenêtres de granulation temporelle soit égale à l'unité, on doit avoir :

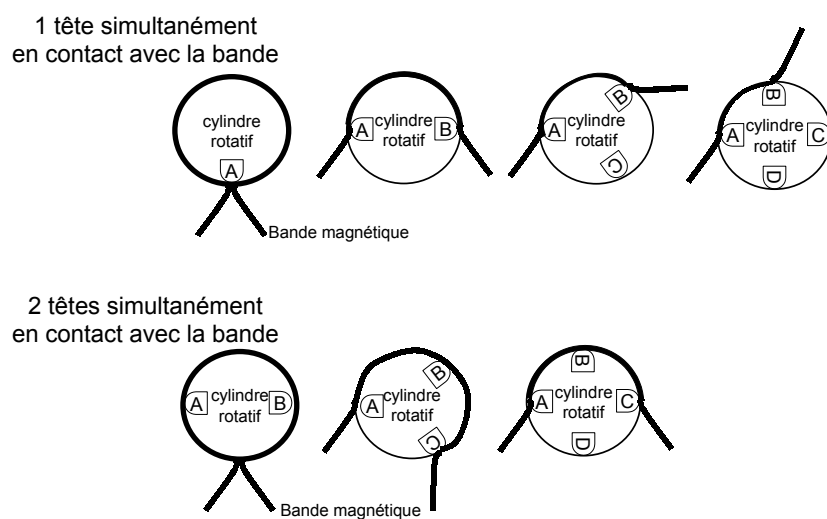
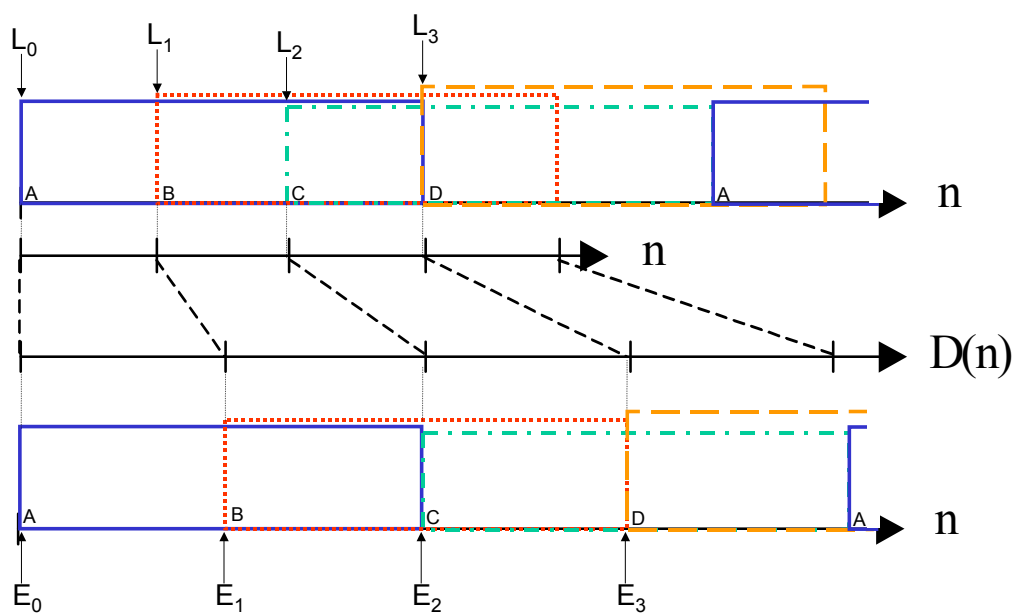
$$H = q.E \quad q \in \mathbb{N}$$

avec E l'espacement entre 2 marques d'écriture.

Ainsi, pour $q = 1$, les fenêtres de granulation temporelle se juxtaposent en sortie car leur durée correspond à la durée entre deux marques d'écriture. Une seule et unique tête est en contact avec la bande magnétique à tout moment. Il est possible de mettre en pratique cette méthode à l'aide d'un cylindre rotatif comportant une seule tête de lecture.

Pour $q = 2$, 2 têtes de lecture au minimum sont requises, mais le résultat est identique avec 2, 3, 4 ou N têtes de lecture, tant que leurs points de contact avec la bande s'effectuent aux endroits adéquats.

La figure D.3 schématise ces différentes possibilités pour $q = 1$ et $q = 2$, et la figure D.4 schématise le positionnement des fenêtres pour $q = 2$.

Figure D.3 – Schéma des cylindres rotatifs pour $q = 1$ et $q = 2$ Figure D.4 – Schéma de positionnement des fenêtres d'un appareil magnétique pour $q = 2$ ($\alpha = 1,5$)

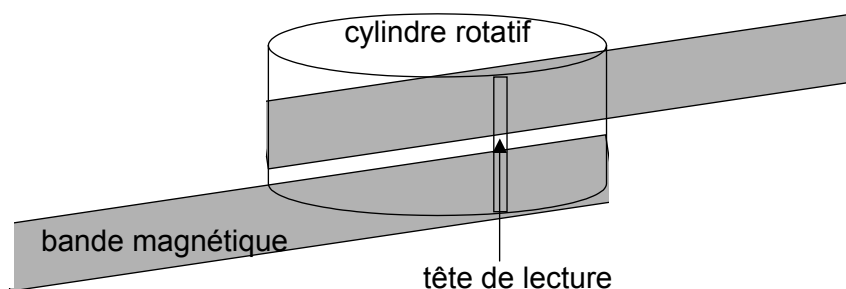


Figure D.5 – Schéma d'une réalisation avec une seule tête de lecture

Fondu-enchaîné

Le fondu-enchaîné est généralement réalisé en écartant progressivement la bande magnétique du cylindre rotatif [LD97] afin que le champs magnétique reçu par la tête de lecture s'atténue.

La fenêtre de granulation temporelle n'est plus rectangulaire, et possède des pentes d'entrée et de sortie progressives égales, et de durée FE . Classiquement, on est amené à utiliser des fenêtres de granulation temporelle possédant des formes trapézoïdales ou, à l'extrême, triangulaires.

Il est également possible de réaliser le fondu-enchaîné en décalant la bande magnétique de l'axe du cylindre. Ceci nous amène à une idée originale (mais obsolète à l'heure du numérique) qui consiste à n'utiliser qu'une seule tête de lecture placée sur un cylindre rotatif dont le schéma est présenté en figure D.5. La durée du fondu enchaîné peut être réglée grâce à l'angle entre la tête de lecture et la bande. Cette mise en œuvre possède l'avantage d'être moins compliquée à mettre au point que ses "concurrentes" puisqu'elle ne nécessite qu'une seule tête de lecture.

Annexe E

Films et sons traités par l'HARMO

Films traités par l'HARMO

Les films présentés dans le tableau suivant représentent uniquement les premières bandes-son traitées par le studio Cinéstéréo. Il s'agit donc seulement des films utilisant la première version de l'HARMO datant de décembre 2000, mais ce studio en traite en permanence, avec une moyenne de deux par jour.

<i>Titre du film</i>	<i>Réalisateur</i>	<i>Transfert</i>
Le stade de Wimbledon	Mathieu Almaric	24 → 25
L'emploi du temps	Laurent Cantet	24 → 25
Va savoir	Jacques Rivette	24 → 25
Une hirondelle a fait le printemps	Christian Carion	24 → 25
Mauvais genre	Francis Girod	24 → 25
Trouble every day	Claire Denis	24 → 25
Eloge de l'amour	Jean-Luc Godart	24 → 25
Les visiteurs en Amérique (just visiting)	Jean-Marie Gaubert	24 → 25
Yamakasi	Ariel Zeitoun	24 → 25
Intimité (Intimacy)	Patrice Chereau	24 → 25
Barnie et ses petites contrariétés	Bruno Chiche	24 → 25
Le pacte des loups	Christophe Gans	24 → 25
Calle 54	Fernando Trueba	24 → 25
Harry un ami qui vous veut du bien	Dominik Moll	24 → 25
Taxi 2	Gérard Krawczyk	24 → 25
Taxi 1	Gérard Pirès	24 → 25
Super 8 stories	Emir Kusturica	25 → 24
Loin	André Techiné	25 → 24
Too much flesh	Jean-Marc Barr	25 → 24
Les glaneurs et la glaneuse	Agnès Varda	25 → 24

On remarque que certains films sont traités dans le sens d'un transfert cinéma vers vidéo (24 → 25, accélération du film sans modification des fréquences) et d'autres dans le sens d'un transfert vidéo vers cinéma (25 → 24, ralentissement du film sans modification des fréquences). Les premiers ont en effet été tourné (au moins partiellement) à 24 images/s avec des caméras "classiques" (support argentique), les seconds ont été tournés (au moins partiellement) à 25 images/s avec des caméras vidéo (support magnétique ou numérique).

Sons musicaux traités par l'HARMO

Nous présentons ici trois exemples sonores illustrant la qualité du système HARMO sur des sons musicaux.

Le son original (son [1]) de la pièce électroacoustique "Enorien" de G. Pallone est rééchantillonné de -4% et transposé par l'HARMO de -4% (son [2]). Il en résulte une légère accélération sans modification des fréquences.

Le son original (son [90]) d'un passage de "La Flûte Enchantée" de W.A. Mozart est rééchantillonné de -20% (son [91]) et +20% (son [93]). Ces sons sont ensuite transposés par l'HARMO respectivement de -20% (son [92]) et +20% (son [94]) afin de compenser la modification de fréquence introduite par le rééchantillonnage.

Le son original (son [95]) d'un passage de "Vesoul" de J. Brel est rééchantillonné de -20% (son [96]) puis transposé par l'HARMO de -20% (son [97]).

Références bibliographiques

- [AB95] G. Assayag and G. Bloch. Quantification et création musicale. In *Colloque Modèles: Esthétique, analyse, sémiotique. Université des Sciences Humaines de Strasbourg*, 22 Avril 1995.
URL: <http://www.ircam.fr/equipes/repmus/RMPapers/Assayag95b/>.
- [AD98] D. Arfib and N. Delprat. Selective transformations of sounds using time-frequency representations: Application to the vibrato modification. In *Proc. 104th AES Convention, Amsterdam, The Netherlands, preprint 4652*, 1998.
- [AES92] AES3-1992. AES Recommended practise for digital audio engineering - Serial transmission format for two-channel linearly represented digital audio data. ANSI S4.40-1992, 1992.
URL: <http://ftp.aessc.org/pub/aes3-1992.pdf>.
- [AH71] B. Atal and S. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, 50(2), pages 637–655, 1971.
- [AKM92] D. Arfib and R. Kronland-Martinet. The hyperbaric voice as a musical sound. In *Proceedings of the Int. Workshop "Speech processing in adverse conditions", Cannes-Mandelieu*, pages 243–246, 10-13 Nov. 1992.
- [AKZ02] D. Arfib, F. Keiler, and U. Zölzer. *"Time-frequency processing" in DAFx: Digital Audio Effects*. Chichester: John Wiley, 2002.
- [All77] J.B. Allen. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25, pages 235–238, 1977.
- [AMI98] AMI. FS6131: Line-Locked Clock Generator IC. Preliminary Information, Jul. 1998.
URL: <http://www.amis.com/pdf/fs6131-01.pdf>.
- [Ana98] AnalogDevice. ADSP-21065L. User's Manual & Technical Reference, Sep. 1998.
URL: http://www.analog.com/Analog_Root/static/library/dspManuals/21065L-UM.html.
- [ANS60] American National Standards Institute: ANSI. USA Standard Acoustical Terminology. S1.1-1960 (R1976), New York, 1960.
- [Ant02] Antares. Antares web site. Logiciel commercial, 2002.
URL: <http://www.antarestech.com/products/auto-tune.html>.
- [AOP02] R. André-Obrecht and J. Pinquier. Reconnaissance et indexation de documents sonores. In *Journée AIM, Bordeaux, France*, 2002.
URL: http://julien.pinquier.free.fr/Articles/AIM_oral.pdf.
- [APB⁺00] A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and Cohen G. Using audio time scale modification for video browsing. In *33rd Hawaii International Conference on System Sciences-Volume 3, Maui, Hawaii*, 4-7 Jan. 2000.

- [AR77] J.B. Allen and L.R. Rabiner. A unified approach to short-time Fourier analysis and synthesis. In *Proceedings of the IEEE*, volume 65, pages 1558–1564, Nov. 1977.
- [Arf79] D. Arfib. Digital synthesis of complex spectra by means of multiplication of non-linear distorted sine waves. *Journal of Audio Engineering Society*, 27, pages 757–768, 1979.
- [Aro92] B. Arons. Techniques, perception, and application of time-compressed speech. In *Proc. Conf. American Voice I/O Society*, pages 169–177, Sep. 1992.
- [Aro97] B. Arons. SpeechSkimmer: A system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction*, 4(1), pages 338, Mar. 1997.
URL: <http://xbean.cs.ccu.edu.tw/~dan/papers/ACMPapers/acmSpeechSkimmer.pdf>.
- [AS84] L.B. Almeida and F.M. Silva. Variable-frequency synthesis: An improved harmonic coding scheme. In *Proc. ICASSP '84*, pages 27.5.1–27.5.4, 1984.
- [Ass85] P.Y. Asselin. *Musique et tempérament*. Editions Costallat, Paris, 1985.
- [Aud02] Audiosculpt. Logiciel institutionnel, 2002.
URL: <http://www.ircam.fr/produits/logiciels/>.
- [Bai95] P. Bailhache. La musique, une pratique cachée de l'arithmétique? In *Studia Leibniziana, Actes du colloque "L'actualité de Leibniz: Les deux labyrinthes"*. Cerisy, 15-22 juin, 1995.
URL: <http://bailhache.humana.univ-nantes.fr/thmusique/leibniz.html>.
- [Bat98] M. Battier. De la machine à l'oreille. Le paradoxe de la musique concrète. In *Colloque: La Musique Concrète jubile à Paris. Ecole Normale de Musique de Paris - Alfred Cortot*, 8 Oct. 1998.
URL: http://homestudio.thing.net/revue/content/battier_concrete.htm.
- [BC02] J. Birr and T. Cuadra. Real-time pitch shifting. EEL6586 Automatic Speech Processing & EEL6935 DSP Programming, Spring 2002.
URL: <http://www.ecel.ufl.edu/~jbirr/voicewarp/appendix.html>.
- [Ben88] J. Benson. *Audio engineering handbook*. McGraw-Hill, N.Y., 1988.
- [Bes93] R.E. Best. *Phase-locked loops*. McGraw-Hill, 1993.
- [Bes98] A. Besse. Reproduction sonore cinématographique (1ère partie). Les dossiers techniques de la CST, n°5, Commission Supérieure Technique de l'Image et du Son, Mars 1998.
URL: <http://www.cst.fr/dtech/05-mar98/index.html>.
- [BHA98] D. Berkani, H. Hassanein, and J.P. Adoul. A single DSP system for high quality enhancement of diver's speech. *IEICE Transactions, Neural Networks/Signal Processing/Information Storage*, E81-A(10), pages 2151–2158, Nov. 1998.
- [BJ95] R. Bristow-Johnson. A detailed analysis of a time-domain formant-corrected pitch-shifting algorithm. *Journal of Audio Engineering Society*, 43(5), pages 340–352, May 1995.
- [BJB01] R. Bristow-Johnson and K. Bogdanowicz. Intraframe time-scaling of nonstationary sinusoids within the phase vocoder. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA01)*, Mohonk Mountain Resort, NY, 21-24 Oct. 2001.
- [Bla97] J. Blauert. *Spatial Hearing: The psychophysics of human sound localization*. M.I.T. Press, Cambridge, MA, 1997.
- [BN93] M. Basseville and I.V. Nikiforov. *Detection of abrupt changes - Theory and application*. Prentice-Hall, Englewood Cliffs, N.J., 1993.
URL: <http://www.irisa.fr/sigma2/kniga/>.

- [Bod84] H. Bode. History of electronic sound modification. *Journal of Audio Engineering Society*, 32(10), pages 730–739, Oct. 1984.
- [BOGC68] J.W. Beauchamp, A.B. Otis, G.R. Grossman, and J.A. Cuomo. Four sound processing programs for the ILLIAC II computer and D/A converter. Technical Reports from the University of Illinois at Urbana-Champaign, 1968.
- [Bon00a] J. Bonada. Automatic technique in frequency domain for near-lossless time-scale modification of audio. In *Proceedings of the International Computer Music Conference, Berlin*, pages 396–399, 2000.
URL: <http://www.isl.uiuc.edu/~kenchen/Ref/2Bonada.pdf>.
- [Bon00b] P. Bonfils. Techniques numériques et animation. Les dossiers techniques de la CST, n°29, Commission Supérieure Technique de l’Image et du Son, Déc. 2000.
URL: <http://www.cst.fr/dtech/29-dec00/index.html>.
- [Bou02] A. Boudier. Etude et amélioration des méthodes de restitution des sons de piano par échantillonnage. Rapport final, DEA d’Acoustique. Université d’Aix-Marseille II, Ecole Doctorale de Mécanique, Physique et Modélisation, 2002.
- [BP59] M.D. Burkhard and R.W. Peters. System for frequency modification of speech and other audio signals. *U.S. Patent N° 3,681,756*, May 12, 1959.
- [BP89] J. Brown and M. Puckette. Calculation of a narrowed autocorrelation function. *J. Acoust. Soc. Am.*, 85(4), pages 1595–1601, 1989.
- [BP00] S. Bosquillon and J. Pigeon. La HD numérique et le 24p. Les dossiers techniques de la CST, n°26, Commission Supérieure Technique de l’Image et du Son, Juin 2000.
URL: <http://www.cst.fr/dtech/26-juin00/index.html>.
- [Can91] G. Canévet. Les profils psychoacoustiques in “Genèse et perception des sons”. Publication n°128 du Laboratoire de Mécanique et d’Acoustique, Oct. 1991.
- [Can00] G. Canévet. Eléments de Psychoacoustique. Cours du DEA d’Acoustique, Université d’Aix-Marseille II, 2000.
- [Cap93] O. Cappe. Techniques de réduction de bruit pour la restauration d’enregistrements musicaux. Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, Paris, 1993.
- [Cau41] A.L. Cauchy. Mémoire sur diverses formules d’analyse. *Comptes Rendus Académie des Sciences, Paris*, 12, pages 63–78, 1841.
URL: <http://gallica.bnf.fr/Fonds.Tables/009/M0090186.htm>.
- [CCJ83] R.V. Cox, R.E. Crochiere, and J.D. Johnston. Real-time implementation of time domain harmonic scaling of speech for rate modification and coding. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31(1), pages 258–272, Feb. 1983.
- [CDP02] CDP. Composers Desktop Project. Logiciel institutionnel, 2002.
URL: <http://www.bath.ac.uk/~masjpf/CDP/>.
- [Cel02] Celemony. Melodyne. Logiciel commercial, 2002.
URL: <http://www.celemony.com/>.
- [Cha88] F. Charpentier. Traitement de la parole par analyse-synthèse de Fourier: Application à la synthèse par diphtonges. Thèse de Doctorat, ENST, 1988.
- [CM89] F. Charpentier and E. Moulines. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphtonges. In *Proceedings EUROSPEECH*, pages 13–19, 1989.
- [Coh95] L. Cohen. *Time-frequency analysis*. Prentice hall, 1995.

- [Col01] T. Collignon. La prise de vues haute définition numérique. Mémoire de fin d'études section cinéma, Ecole Nationale Supérieure Louis Lumière, 2001.
- [Coo90] D.A. Cook. *A history of narrative film*. 2nd Edition, p2. W.W. Norton, New York, 1990.
- [CR83] R.E. Crochiere and L.R. Rabiner. *Multirate digital signal processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1983.
- [Cro80] R.E. Crochiere. A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28, pages 99–102, 1980.
- [Cry98] Crystal. CS8420: Digital audio Sample Rate Converter. Preliminary Product Information, Oct. 1998.
URL: <http://www.sc-elec.demon.co.uk/cs8420.pdf>.
- [CS86] F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proceedings of ICASSP 86*, number 3, pages 2015–2018, 1986.
- [CWS97] M. Covell, M. Withgott, and M. Slaney. Mach1 for nonuniform time-scale modification of speech: Theory, technique, and comparisons. Interval Research Corporation Technical Report, 1997.
URL: <http://rvl4.ecn.purdue.edu/~malcolm/interval/1997-061/>.
- [CWS98] M. Covell, M. Withgott, and M. Slaney. Mach1: Nonuniform time-scale modification of speech. In *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, Seattle WA*, May 12-15 1998.
URL: <http://citeseer.nj.nec.com/covell98mach.html>.
- [Dat87] J. Dattorro. Using digital signal processor chips in a stereo audio time compressor/expander. In *Proc. 83rd AES Convention, New York, preprint 2500 (M-6)*, 1987.
- [Dat98] Dattorro. "Speeding speech and comprehension". Communication sur mailing-list, 1998.
URL: <http://mambo.ucsc.edu/psl/ccrmas/199806/19980625.html>.
- [Dau00] L. Daudet. Représentations structurelles de signaux audiophoniques. Méthodes hybrides pour des applications à la compression. Thèse de Doctorat, Université de Provence, Marseille, 2000.
- [DB93] C. Drake and M.C. Botte. Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception and Psychophysics*, 54, pages 277–286, 1993.
- [DCM00] M. Desainte-Catherine and S. Marchand. High-precision Fourier analysis of sounds using signal derivatives. *Journal of Audio Engineering Society*, 48(7/8), July/August 2000.
- [DDPZ02] P. Dutilleux, G. De Poli, and U. Zölzer. "Time-segment processing" in *DAFx: Digital Audio Effects*. Chichester: John Wiley, 2002.
- [DDS01] C. Duxbury, M.E. Davies, and M. Sandler. Separation of transient information in musical audio using multiresolution analysis techniques. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-01), Limerick, Ireland*, December 6-8, 2001.
URL: <http://www.csis.ul.ie/dafx01/proceedings/papers/duxbury.pdf>.
- [DDS02] C. Duxbury, M.E. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. In *Proc. AES 112th Convention, Munich, Germany*, 2002.

- [Del91] J.P. Delmas. *Éléments de théorie du signal: Les signaux déterministes*. Ellipse, Paris, 1991.
- [Dem89] L. Demany. "Perception de la hauteur tonale" in *Psychoacoustique et perception auditive*. INSERM, 1989.
- [Dep99] P. Depalle. Notes de cours d'analyse/synthèse. Cours du DEA ATIAM, 1999.
- [DF98] R. Di Federico. Waveform preserving time stretching and pitch shifting for sinusoidal models of sound. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-98), Barcelona, Spain*, pages 44–48, 1998.
URL: <http://www.iua.upf.es/dafx98/papers/>.
- [DGBA00] A. De Goetzen, N. Bernardini, and D. Arfib. Traditional (?) implementations of a phase-vocoder: The tricks of the trade. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-00), Verona, Italy*, pages 7–43, 2000.
- [Dob02] R. Dobson. PVOC-EX. Page web, 2002.
URL: <http://www.cs.bath.ac.uk/jpff/NOS-DREAM/researchdev/pvocex/pvocex.html>.
- [Dol86] M. Dolson. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4), pages 14–27, Winter 1986.
- [Dol00] M. Dolson. System for Fourier transformed-based modification of audio. *U.S. Patent N° 6,112,169*, Aug. 29, 2000.
- [Dol02a] Dolby. Dolby Home Page. Page web, 2002.
URL: <http://www.dolbydigital.com>.
- [Dol02b] Dolby. Model 585, Time Scaling Processor. Specifications, 2002.
URL: <http://www.dolby.com/products/Model585/>.
- [DP91] P. Depalle and G. Poirot. SVP: A modular system for analysis, processing and synthesis of sound signals. In *Proceedings of ICMC, Montréal, Canada*, 1991.
- [DPP⁺96] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, USA.*, 1996.
URL: <http://www.asel.udel.edu/icslp/cdrom/vol3/920/a920.pdf>.
- [DT02] L. Daudet and B. Torrèsani. Hybrid representations for audiophonic signal encoding. *Signal Processing, Special issue on Image and Video Coding Beyond Standards*, 82(11), pages 1595–1617, 2002.
- [DTS02] DTS. DTS Home Page. Page web, 2002.
URL: <http://www.dtsonline.com/>.
- [Dud39] H. Dudley. The Vocoder. *Bell Labs*, 18, pages 122–126, Dec. 1939.
- [Dud02] R. Dudas. Spectral envelope correction for real-time transposition: Proposal of a "floating-formant" method. In *Proceedings of the International Computer Music Conference, Göteborg, Sweden.*, pages 126–129, 2002.
- [Dut93] T. Dutoit. High quality text-to-speech synthesis of the french language. PhD dissertation, Faculte Polytechnique de Mons, 1993.
- [Ele02] T.C. Electronic. System 6000. Page web, 2002.
URL: <http://www.system6000.com/>.
- [Ell91] D. Ellis. "pvanal.c". Part of the Csound distribution, MIT, 1991.
URL: <http://www.sfu.ca/sca/Manuals/Csound/pvocinfo.html>.
- [Ell92] D.P. Ellis. Timescale modifications and wavelet representations. In *Proc. Int. Computer Music Conference, San José, USA*, pages 6–9, Jun. 1992.

- [Ell02a] Ellis. A Phase Vocoder in Matlab. Logiciel institutionnel, 2002.
URL: <http://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc-matlab.html>.
- [Ell02b] Ellis. SOLAFS in Matlab. Logiciel institutionnel, 2002.
URL: <http://www.ee.columbia.edu/~dpwe/resources/matlab/solafs-matlab.html>.
- [Em02] E-mu. Darwin. Specifications, 2002.
URL: <http://www.emu.com/products/darwin/darwin.html>.
- [Eng77] I. Engel. A minicomputer implementation of an isolated-word recognition system. M.Sc. thesis, Technion - I.I.T., 1977.
- [Eno02] Enounce. 2xAV. Page web, 2002.
URL: <http://www.enounce.com/>.
- [Erb02] T. Erbe. Soundhack. Logiciel, 2002.
URL: <http://www.soundhack.com/>.
- [Fav01] E. Favreau. Phase vocoder applications in GRM tools environment. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-01)*, Limerick, Ireland, Dec. 2001.
- [FEJ54] G. Fairbanks, W. Everitt, and R. Jaeger. Method for time or frequency compression-expansion of speech. *Transactions of the Institute of Radio Engineers, Professional Group on Audio*, AU-2, pages 7–12, 1954.
- [FEJ59] G. Fairbanks, W. Everitt, and R. Jaeger. Recording Device. *U.S. Patent N° 2,886,650*, May 12, 1959.
- [FG66] J.L. Flanagan and R.M. Golden. Phase Vocoder. *Bell System Technical Journal*, 45, pages 1493–1509, Nov. 1966.
URL: <http://www.ee.columbia.edu/~dpwe/e6820/papers/FlanG66.pdf>.
- [Fla98] P. Flandrin. *Temps-fréquence*. (2ème édition, revue et corrigée) Editions Hermes, Paris, 1998.
- [Fla02] M. Flax. MFFM Time scale modification for audio. Logiciel, 2002.
URL: <http://sourceforge.net/projects/mffmtimescale/>.
- [Fle40] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12, pages 47–65, Jan. 1940.
- [For02] Formats. Formats d'insérables. Page web, 2002.
URL: <http://www.audiomidi.com/plugins/index.cfm>.
- [Fou64] E. Foulke. The comprehension of rapid speech for the blind Part II. , KY: Nonvisual Perceptual Systems Laboratory, University of Louisville, 1964.
- [Fre35] B. Freund. Method of and apparatus for varying the length of sound records. *U.S. Patent N° 1,996,958*, Apr. 9, 1935.
- [FS95] A. Friberg and J. Sundberg. Time discrimination in a monotonic, isochronous sequence. *J. Acoust. Soc. Am.*, 98, pages 2524–2531, 1995.
- [FZ28] N.R. French and M.K. Zinn. Method of and apparatus for reducing width of transmission bands. *U.S. Patent N° 1,671,151*, May 29, 1928.
- [Gab46] D. Gabor. Theory of Communication. *J. of the Institute of Electrical Engineers*, 93(III), pages 429–457, 1946.
- [Gar53] W.D. Garvey. The intelligibility of abbreviated speech patterns. *Quarterly Journal of Speech*, 39, pages 296–306, 1953.
- [Ges98] Y. Geslin. Sound and music transformation environments: A twenty years experiment at the "Groupe de Recherches Musicales". In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-98)*, Barcelona, Spain, pages 241–248,

1998.
URL: <http://www.iaa.upf.es/dafx98/papers/>.
- [GKMM89] A. Grossmann, R. Kronland-Martinet, and J. Morlet. "Reading and understanding continuous wavelet transforms" in *Wavelets, time-frequency representations and phase space*. Ed. J.M Combes, A. Grossmann, P. Tchamitchian, Springer Verlag, pp. 2-20, 1989.
- [GL84] D.W. Griffin and J.S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2), pages 236–243, 1984.
- [GL88] D.W. Griffin and J.S. Lim. Multiband-excitation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(2), pages 236–243, 1988.
- [Gol62] B. Gold. Computer program for pitch extraction. *J. Acoust. Soc. Am.*, 34(7), pages 916–921, 1962.
- [Gri99] R. Gribonval. Approximations non-linéaires pour l'analyse des signaux sonores. Thèse, spécialité: Mathématiques Appliquées, Université de Paris IX Dauphine, 1999.
- [GS85] J.W. Gordon and John Strawn. "An introduction to the phase vocoder" in *Digital Audio Signal Processing: An Anthology*. John Strawn, ed., Los Altos, CA. W. Kaufmann, Inc., pp. 221-270, 1985.
- [GS98] J. Garas and P.C.W. Sommen. Time/pitch scaling using the constant-Q phase vocoder. In *Proc. CSSP-98, ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing, Mierlo, Netherlands*, Nov. 1998.
URL: <http://www.stw.nl/prorisc/prorisc98/proceedings/garas.pdf>.
- [Gui02] P. Guillemain. Analyse, synthèse et transformation des sons musicaux. Cours du DEA d'Acoustique de Marseille, 2002.
- [HBG⁺94] G. ten Hoopen, L. Boelaarts, A. Gruisen, I. Apon, K. Donders, N. Mul, and S. Akerboom. The detection of anisochrony in monaural and interaural sound sequences. *Perception and Psychophysics*, 56(1), pages 110–120, 1994.
- [HD82] A.R. Halpern and C.J. Darwin. Duration discrimination in a series of rhythmic events. *Perception and Psychophysics*, 31, pages 86–89, 1982.
- [HDdC00] N. Henrich, B. Doval, C. d'Alessandro, and M. Castellengo. Open quotient measurements on EGG, speech and singing signals. In *Proc. 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research, Jena*, Apr. 2000.
URL: <http://www.limsi.fr/Individu/henrich/>.
- [Hej90] D.J. Hejna. Real-time time-scale modification of speech via the synchronized overlap-add algorithm. Master's thesis, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, Mass., 1990.
- [HG00] L. He and A. Gupta. User benefits of non-linear time compression. Technical Report MSR-TR-2000-96, Microsoft Research, Sep. 2000.
URL: <http://citeseer.nj.nec.com/376022.html>.
- [Hib83] S. Hibi. Rhythm perception in repetitive sound sequence. *Journal of the Acoustical Society of Japan*, 4, pages 83–95, 1983.
- [HMC92] D.J. Hejna, B.R. Musicus, and A.S. Crowe. Method for time-scale modification of signals. *U.S. Patent N° 5,175,769*, Dec. 29, 1992.
- [Hoe01] S.M.J. Hoek. Method and apparatus for signal processing for time-scale and/or pitch modification of audio signals. *U.S. Patent N° 6,266,003*, July 24, 2001.

- [HTCT97] K. N. Hamdy, A. H. Tewfik, T. Chen, and S. Takagi. Time-scale modification of audio signals with combined harmonic and wavlet representations. In *IEEE Proc. Int. Conf. Acoust., Speech and Signal Processing*, Apr. 1997.
- [IK99] T. Itagaki and D. Knox. Multimedia application of time compress/stretch of sound by granulation. In *Proceedings of International Computer Music Conference, Beijing, China*, pages 512–514, Oct. 1999.
- [Ita98] T. Itagaki. Real-time sound synthesis on a multi-processor platform. Ph.D Thesis, University of Durham, 1998.
- [Ita00] T. Itagaki. Sound compression/interpolation by granulation. In *108th Audio Engineering Society Convention, Paris, FRANCE, preprint 5126 (J-5)*, Feb. 2000.
URL: <http://www.brunel.ac.uk/~eesttti/papers/aes108p.html>.
- [KIS⁺99] D. Knox, T. Itagaki, I. Stewart, A. Nesbitt, and I.J. Kemp. Preservation of local sound periodicity with variable-rate video. In *Proceedings of the 7th ACM Multimedia Conference, Orlando, USA*, pages 299–302, Oct. 1999.
URL: <http://www.kom.e-technik.tu-darmstadt.de/acmmm99/ep/knox/>.
- [KMMG87] R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of sound pattern through wavelet transform. *Inter. J. of Pattern Analysis and Artificial Intelligence*, 1(2), pages 273–302, 1987.
- [Kor97] R. Kortekaas. Physiological and psychoacoustical correlates of perceiving natural and modified speech. Ph.D. thesis, Technical University of Eindhoven, 1997.
- [Lar93] J. Laroche. Autocorrelation method for high quality time/pitch scaling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York*, 1993.
- [Lar95] J. Laroche. Traitement des signaux audio-fréquences. Support de cours, Département Signal, TELECOM Paris, 1995.
- [Lar98] J. Laroche. "Time and pitch scale modification of audio signals" in *Applications of digital signal processing to audio and acoustics*. M. Kahrs and K. Brandenburg, eds., Kluwer Academic Publishers, pp. 279-309, 1998.
- [Lar00] J. Laroche. Time-domain time/pitch scaling of speech or audio signals with transient handling. *U.S. Patent N° 6,049,766*, Apr. 11, 2000.
- [LD97] J. Laroche and M. Dolson. About this phasiness business. In *Proceedings of the International Computer Music Conference*, 1997.
- [LD99a] J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3), pages 223–232, May 1999.
- [LD99b] J. Laroche and M. Dolson. New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing, and other exotic audio modifications. *Journal of Audio Engineering Society*, 47(11), pages 928–936, Nov. 1999.
- [LD99c] J. Laroche and M. Dolson. New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York*, Oct. 17-20, 1999.
URL: <http://www.ee.columbia.edu/~dpwe/papers/LaroD99-pvoc.pdf>.
- [Lee72] F. Lee. Time compression and expansion of speech by the sampling method. *Journal of Audio Engineering Society*, 20(3), pages 738–742, 1972.
- [Lem02] Lemur. Page web, 2002.
URL: <http://www.cerlsoundgroup.org/Lemur/>.
- [Lev98] S. Levine. Audio representation for data compression and compressed domain processing. Ph.D. thesis, Stanford University, 1998.

- [Lex02] Lexicon. PCM80/81. Page web, 2002.
URL: <http://www.lexicon.com/pcm81/options.asp>.
- [LFG02] Y. Liang, N. Färber, and B. Girod. Adaptive playout scheduling and loss concealment for voice communication over IP networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 438–449, 2002.
URL: <http://www.citeseer.nj.nec.com/liang02adaptive.html>.
- [LKK97] S. Lee, H. D. Kim, and H. S. Kim. Variable time-scale modification of speech using transient information. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Munich*, 1997.
- [LT98] A.H.J. Lin and R.K.C. Tan. Time-scale modification algorithm for audio and speech signal applications. In *Proc. 104th AES Convention, Amsterdam, preprint 4644 (p.574)*, volume 46, 1998.
- [MA89] J.S. Marques and L.B. Almeida. Frequency-varying sinusoidal modeling of speech. 37(5), pages 763–765, 1989.
- [Mal79] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2), pages 121–133, 1979.
- [Man85] P. Manning. *Electronic & computer music*. Oxford: Clarendon Press, 1985.
- [Mas96] P. Masri. Computer modelling of sound for transformation and synthesis of musical signals. PhD Thesis, University of Bristol, 1996.
- [Mas98] D. Massie. "Wavetable sampling synthesis" in *Applications of digital signal processing to audio and acoustics*. M. Kahrs and K. Brandenburg, eds., Kluwer Academic Publishers, pp. 311–341, 1998.
- [Mat03] Matlab. Mathematical computation, analysis, visualization, algorithm development, and deployment. The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098, USA, 2003.
URL: <http://www.mathworks.com/>.
- [Max80] N. Maxemchuk. An experimental speech storage and editing facility. *Bell System Technical Journal*, 59(8), pages 1383–1395, 1980.
- [Max86] J. and Max. *Méthodes et techniques de traitement du signal et applications aux mesures physiques. Tome 2 Appareillages; exemples d'applications; méthodes nouvelles*. 3ème édition, Editions Masson, Paris, 1986.
- [MC90] E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6), pages 453–467, 1990.
- [MCC93] R.E. Malah, Crochiere, and R.V. Cox. Performance of transform and subband coding systems combined with harmonic scaling of speech. 41(12), pages 3397–3415, Dec. 1993.
- [MEJ86] J. Makhoul and A. El-Jaroudi. Time-scale modification in medium to low rate speech coding. In *IEEE Proc. Int. Conf. Acoust., Speech and Signal Processing*, pages 1705–1708, 1986.
- [Meu90] S. Meunier. Transformation de signaux vocaux en plongée hyperbare. Rapport de stage de Maîtrise de Physique, Equipe d'Informatique Musicale, LMA-CNRS, 1990.
- [MG76] J.D. Markel and A.H.Jr. Gray. *Linear prediction of speech*. Springer-Verlag, Berlin, Heidelberg, New York, 1976.
- [Mic64] J.A. Michon. Studies on subjective duration, I. Differential sensitivity in the perception of repeated temporal intervals. *Acta Psychologica*, 22, pages 441–450, 1964.

- [Mic03] Microsoft. Windows Media Player. Page web, 2003.
URL: <http://www.microsoft.com>.
- [ML50] G.A. Miller and J.C.R. Licklider. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.*, 22, pages 167–173, 1950.
- [ML95] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16, pages 175–205, 1995.
- [Moo78] J.A. Moorer. The use of the phase vocoder in computer music applications. *Journal of Audio Engineering Society*, 26(1/2), pages 42–45, 1978.
- [Moo90] F.R. Moore. *The phase vocoder in Elements of Computer Music*. Prentice-Hall, pp. 227–263, 1990.
- [Moo97] B.C.J. Moore. *An introduction to the psychology of hearing*. 4th ed., Academic Press, 1997.
- [MRK⁺93] P. Meyer, H.W. Rühl, M. Kugler, L.L.M. Vogten, A. Dirksen, and K. Belhoula. PHRITTS: A text-to-speech synthesizer for the german language. *Eurospeech'93, Berlin*, pages 877–890, 1993.
- [MT01] M. Mansour and A. Tewfik. Audio watermarking by time-scale modification. Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Salt Lake City, May 2001.
URL: <http://www.ece.umn.edu/users/mmansour/Publications.htm>.
- [MZ89] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. 37(5), pages 763–765, 1989.
- [Neu78] E.P. Neuburg. Simple pitch-dependent algorithm for high quality speech rate changing. *J. Acoust. Soc. Am.*, 63(2), pages 624–625, 1978.
- [Neu98] M.R. Neuman. Dolby Digital vs DTS for 5.1 delivery. Page web, 1998.
URL: <http://music1.csudh.edu/MUS450/Students/MichaelNeuman/dolbyvsdts.htm>.
- [Nul03] Nullsoft. Winamp. Page web, 2003.
URL: <http://www.winamp.com>.
- [Nyq28] H. Nyquist. Certain topics in telegraph transmission theory. *AIEE Trans.*, 47, pages 617–644, 1928.
- [OFW56] D.B. Orr, H.L. Friedman, and J.C.C. Williams. *J. Educ. Psychol.*, 56, pages 148–156, 1956.
- [OS75] A.V. Oppenheim and R.W. Schaffer. *Digital signal processing*. Prentice-Hall International, 1975.
- [OS89] A.V. Oppenheim and R.W. Schaffer. *Discrete-time signal processing*. Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [Pal93] C. Palombini. Pierre Schaeffer, 1953: Towards an experimental music. *Music & Letters*, 74(4), pages 542–557, 1993.
- [Pal99] G. Pallone. Transposition fréquentielle pour des applications de post-production audiovisuelle. Mémoire de fin d'année, DEA ATIAM. Université d'Aix-Marseille II, Ecole Doctorale de Mécanique, Physique et Modélisation, 1999.
- [Par02] O. Parviainen. PaceMaker. Logiciel insérable, 2002.
URL: <http://www.sunpoint.net/~oparviain/pacemaker/>.
- [PBD⁺99] G. Pallone, P. Boussard, L. Daudet, P. Guillemain, and R. Kronland-Martinet. A wavelet based method for audio-video synchronization in broadcasting applications. In *Proc. of the COST G-6 Conference on Digital Audio Effects (DAFx-99)*, 9-11 Dec., Trondheim, Norway, pages 59–62, 1999.
URL: <http://citeseer.nj.nec.com/pallone99wavelet.html>.

- [PBKM00] G. Pallone, P. Boussard, and R. Kronland-Martinet. Transposition fréquentielle pour des applications de post-production cinématographique. In *Actes du 5ème congrès français d'acoustique, 3-6 Sep., EPFL, Lausanne, Suisse, Presses Polytechniques et Universitaires Romandes*, pages 595–598, 2000.
- [PDH99] J.G. Proakis, J.R. Deller, and J.H.L. Hansen. *Discrete time processing of speech signals*. Wiley, John and Sons, Inc., 1999.
- [Pee98] G. Peeters. Analyse et synthèse des sons musicaux par la méthode PSOLA. In *JIM98-Workshop, Agelonde, France*, Mai 1998.
- [Pee01] G. Peeters. Modèles et modification du signal sonore adaptés aux caractéristiques locales. Thèse, spécialité: Acoustique Traitement du signal et Informatique Appliqués à la Musique, Université de Paris 6, 2001.
- [Pie83] J.R. Pierce. *The science of musical sound*. Freeman, 1983.
- [PJ02] C. Penrose and A. Jaffe. PVC. Page web, 2002.
URL: <http://silvertone.princeton.edu/winham/PPSK/pvc.html>.
- [PMH00] G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of MPEG-7. In *Proceedings of International Computer Music Conference, Berlin, Germany*, Aug. 2000.
URL: <http://citeseer.nj.nec.com/peeters00instrument.html>.
- [Poh00] K.C. Pohlmann. *Principles of digital audio*. Mc Graw-Hill, fourth edition, pp. 22-23, 2000.
- [Por76] M.R. Portnoff. Implementation of the digital phase vocoder using the Fast Fourier Transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3), pages 243–248, 1976.
- [Por80] M.R. Portnoff. Time-frequency representation of digital signals and systems based on short-time Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(1), pages 55–69, Feb. 1980.
- [Por81a] M.R. Portnoff. Short-time Fourier analysis of sampled speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3), pages 364–373, 1981.
- [Por81b] M.R. Portnoff. Time-scale modifications of speech based on short-time Fourier analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3), pages 374–390, 1981.
- [PR99] G. Peeters and X. Rodet. SINOLA: A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum. In *Proceedings of International Computer Music Conference, Beijing, China*, pages 153–156, Oct. 1999.
- [Pre98] D. Pressnitzer. Perception de rugosité psychoacoustique: D'un attribut élémentaire de l'audition à l'écoute musicale. Thèse, spécialité: Acoustique Traitement du signal et Informatique Appliqués à la Musique, Université de Paris 6, 1998.
- [Pro02] Prosoniq. Prosoniq web site. Page web, 2002.
URL: <http://www.prosoniq.net/>.
- [Puc95] M. S. Puckette. Phase-locked vocoder. In *Proc. IEEE Conf. on Applications of Signal Processing to Audio and Acoustics, Mohonk*, 1995.
- [QDH95] T.F. Quatieri, R.B. Dunn, and T.E. Hanna. A subband approach to time-scale expansion of complex acoustic signals. *IEEE Transactions on Speech and Audio Processing*, 3(6), pages 515–519, 1995.
- [QM86] T. F. Quatieri and R.J. McAulay. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(6), pages 1449–1464, Aug. 1986.

- [RCRM76] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5), pages 399–418, Jul. 1976.
- [RD92] X. Rodet and P. Depalle. Spectral envelopes and inverse FFT synthesis. In *93rd Convention of the Audio Engineering Society, New York*, 1992.
- [Rea03] RealNetworks. RealPlayer. Page web, 2003.
URL: <http://www.real.com>.
- [Ris99] J.C. Risset. *Evolution des outils de création sonore in "Interfaces homme-machine et création musicale", sous la direction de H. Vinet et F. Delalande, pp. 17-36.* Hermès, Paris, 1999.
- [RJ01] X. Rodet and F. Jaillet. Detection and modeling of fast attack transients. In *Proceedings of ICMC, Cuba*, Sep. 2001.
- [Roa96] C. Roads. *The computer music tutorial*. M.I.T Press, 1996.
- [Roe90] C. Roehrig. Time and pitch scaling of audio signals. In *Proc. 89th AES Convention, Los Angeles, preprint 2954 (E-1)*, 1990.
- [Ros00] S. Rossignol. Segmentation et indexation des signaux sonores musicaux. Thèse, spécialité: Acoustique Traitement du signal et Informatique Appliqués à la Musique, Université de Paris 6, 2000.
- [RW85] S. Roucos and A. M. Wilgus. High quality time-scale modification of speech. In *IEEE Proc. Int. Conf. Acoust., Speech and Signal Processing, Tampa, FL*, Mar. 1985.
- [RW99] F. Rumsey and J. Watkinson. *Le guide des interfaces numériques*. Eyrolles, 1999.
- [SB86] M.J.T. Smith and T.P. Barnwell. Exact reconstruction techniques for tree-structured subband coders. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34, pages 434–441, June 1986.
- [Sch44] E. Schüller. Sound reproducer. *U.S. Patent N° 2,352,023*, Jun. 20, 1944.
- [Sch54] H. Schiesser. A device for time expansion used in sound recording. *Transactions of the Institute of Radio Engineers, Professional Group on Audio*, AU-2, pages 12, 1954.
- [Sch66a] P. Schaeffer. *Traité des objets musicaux*. Editions du Seuil, Paris, 1966.
- [Sch66b] M.R. Schroeder. Vocoders: Analysis and synthesis of speech. In *Proceedings of the IEEE*, volume 54(5), pages 720–734, 1966.
- [Sch78] H.H. Schulze. The detectability of local and global displacements in regular rhythmic patterns. *Psychological Research*, 40, pages 173–181, 1978.
- [Sco67] R.J. Scott. Time adjustment in speech synthesis. *J. Acoust. Soc. Am.*, 41(1), pages 60–65, 1967.
- [Scu02] Sculptor. A Real-Time Phase Vocoder for Linux. Page web, 2002.
URL: <http://sculptor.sourceforge.net/Sculptor/lj/lj.html>.
- [SDD02] SDDS. SDDS Home Page. Page web, 2002.
URL: <http://www.sdds.com/>.
- [Ser89] X. Serra. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Ph.D. Dissertation, Stanford University, 1989.
- [Ser02] X. Serra. SMS. Page web, 2002.
URL: <http://www.iua.upf.es/sms/>.
- [SG84] J.O. Smith and P. Gossett. A flexible sampling-rate conversion method. In *Proc. IEEE ICASSP-84, San Diego*, 1984.

- [Sha48] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, pages 379–423 and 623–656, Jul. and Oct. 1948.
URL: <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>.
- [SL95] Z. Settel and C. Lippe. Real-time musical applications using frequency domain signal processing. In *Proc. IEEE Conf. on Applications of Signal Processing to Audio and Acoustics, Mohonk*, 1995.
- [Son88] SonicStudio. Time Twist. Manual Version 5, Sonic Solution, 1988.
- [Son02] SonicSolution. SonicSolution Milestones. Page web, 2002.
URL: <http://www.sonic.com/corporate/milestones.htm>.
- [SP92] B. Sylvestre and Kabal P. Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, San Francisco, USA*, pages I–81–I–84, Mar. 1992.
URL: <http://www.tsp.ece.mcgill.ca/Kabal/papers/>.
- [Spr55] A. Springer. Ein akustischer Zeitregler. *Gravesaner Blätter*, 1, pages 32–37, 1955.
- [Spr02] S.M. Sprenger. Time and pitch scaling of audio signals. Page web, 2002.
URL: <http://dspdimension.com/html/timepitch.html>.
- [SR73] R.W. Schafer and L.R. Rabiner. Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis. *IEEE Transactions on Audio and Electroacoustics*, AU-21(3), pages 165–174, 1973.
- [SS90] X. Serra and J.O. Smith. Spectral Modeling Synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4), pages 12–24, 1990.
- [Ste02] Steinberg. Steinberg web site. Page web, 2002.
URL: <http://www.steinberg.net>.
- [Sti69] T.G. Sticht. Comprehension of repeated time-compressed recordings. *The Journal of Experimental Education*, 37(4), Summer 1969.
- [SVW94] G. Spleesters, W. Verhelst, and A. Wahl. On the application of automatic waveform editing for time warping digital and analog recordings. In *Proc. 96th AES Convention, Amsterdam, preprint 3843 (p. 11.3)*, 1994.
- [TB61] W.R. Tiffany and D.N. Bennett. *J. Speech & Hearing Res.*, 4, pages 248–258, 1961.
- [Ter02] D. Terez. In *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Orlando, USA*, 2002.
- [Tho82] D.J. Thomson. Spectrum estimation and harmonic analysis. In *Proceedings of the IEEE (Special issue on spectrum estimation)*, number 70, pages 1055–1096, 1982.
- [TL00] R.K.C. Tan and A.H.J. Lin. A time-scale modification algorithm based on the subband time-domain technique for broad-band signal applications. *Journal of Audio Engineering Society*, 48(5), pages 437–449, May 2000.
- [Tor91] B. Torresani. Wavelet associated with representations of the affine Weyl-Heisenberg group. *J. Math. Physics*, 32, pages 1273–1279, May 1991.
- [Tor98] B. Torresani. An overview of wavelet analysis and time-frequency analysis (a mini-course). In *Self-Similar Systems, Proceedings of the International Workshop, Dubna, Russia*, July 30 - August 7, 1998.
- [Tra90] H. Trautmüller. Analytical expressions for the tonotopic sensory scale. *J. Acoust. Soc. Am.*, 88, pages 97–100, 1990.
- [Tru94] B. Truax. Discovering inner complexity: Time-shifting and transposition with a real-time granulation technique. *Computer Music Journal*, 18(2), pages 38–48, 1994.

- [UQA96] UQAM. Dictionnaire des arts médiatiques. Groupe de recherche en arts médiatiques, 1996.
URL: <http://www.comm.uqam.ca/~GRAM/>.
- [VC93] C. Valette and C. Cuesta. *Mécanique de la corde vibrante*. Hermes, Traité des nouvelles technologies, 1993.
- [Ver00] W. Verhelst. Overlap-add methods for time-scaling of speech. *Speech Communication*, 30, pages 207–221, 2000.
- [VH96] R. Veldhuis and Haiyan He. Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform. *Speech Communications*, 18, pages 257–279, 1996.
- [Vil00] A. Villeval. La synchronisation de l'image et du son au cinéma. Les dossiers techniques de la CST, n°25, Commission Supérieure Technique de l'Image et du Son, Mai 2000.
URL: <http://www.cst.fr/dtech/25-mai00/index.html>.
- [VM98] T. Verma and T. Meng. Time scale modification using a Sines + Transients + Noise signal model. In *Digital Audio Effects Workshop, Barcelona, Spain*, 19-21 Nov. 1998.
URL: <http://www.acoustics.hut.fi/~tverma/publications/>.
- [VR93] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *IEEE Proceedings of ICASSP-93, Minneapolis, vol. II*, pages 554–557, Apr. 1993.
- [vS94] J. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8(2), pages 95–128, 1994.
- [VVAf97] P.G. Vos, M.A.L.M. Van Assen, and M. Franek. Perceived tempo change is dependent on base tempo and direction of change: Evidence for a generalized version of Schulze's (1978) Internal Beat model. *Psychological Research*, 59(4), pages 240–247, 1997.
- [Wat94] J. Watkinson. *The art of digital video - 2nd Edition*. Oxford, Boston: Focal Press, 1994.
- [Whi02] J.C. Whitaker. *Standard handbook of video and television engineering, 3rd ed., chap. 19.7*. Mc Graw-Hill, 2002.
- [WJG77] C.C. Wier, W. Jesteadt, and D.M. Green. Frequency discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.*, 61, pages 178–184, 1977.
- [Woo51] H. Woodrow. "Time perception", *Handbook of experimental psychology*, pp. 1224–1236. S.S. Stevens, Wiley, New York, 1951.
- [WW88] J. L. Wayman and D. L. Wilson. Some improvements on the synchronized-overlap-add method of time-scale modification for use in real-time speech compression and noise filtering. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1), pages 139–140, January 1988.
- [Xil02] Xilinx. Spartan and Spartan-XL families Field Programmable Gate Arrays. Product Specification, Jun. 2002.
URL: <http://direct.xilinx.com/partinfo/ds060.pdf>.
- [YMF⁺03] S. Ystad, C. Magne, S. Farner, G. Pallone, V. Padeloup, R. Kronland-Martinet, and M. Besson. Influence of rhythmic, melodic, and semantic violations in language and music on the electrical activity in the brain. In *Stockholm Music Acoustics Conference 2003, Stockholm, Sweden*, 6 - 8 August 2003.

- [Yst89] S. Ystad. Sound modeling using a combination of physical and signal models. Thèse de Doctorat, Université d'Aix-Marseille II et Université de Trondheim (NTNU), 1989.
- [ZF90] E. Zwicker and H. Fastl. *Psychoacoustics facts and models*. Springer-Verlag Berlin Heidelberg, 1990.
- [Zöl97] U. Zölzer. *Digital audio signal processing*. John Wiley & Son Ltd, 1997.
- [Zöl02] U. Zölzer. *DAFx: Digital Audio Effects*. Chichester: John Wiley, 2002.

Références aux pistes audio

- [1] "Enorien" de G. Pallone : Pièce électroacoustique originale.
- [2] "Enorien" de G. Pallone : Pièce électroacoustique rééchantillonnée de -4% et transposé-p pour compensation de -4% par l'HARMO.
- [3] Voyelle "a" : Son original puis son dilaté entraînant simultanément une dilatation temporelle et une transposition fréquentielle.
- [4] Signal représentant 2 transitoires échantillonnés à 44,1 kHz puis les mêmes transitoires échantillonnés à 11 kHz.
- [5] Voix d'homme : Son original.
- [6] Voix d'homme : Son original rééchantillonné de +4,2% (donc ralenti et transposé vers le bas).
- [7] Voix d'homme : Son rééchantillonné de +4,2% et transposé-p pour compensation de +4,2% par l'HARMO.
- [8] Musique : Son original.
- [9] Musique : Son original rééchantillonné de -4% (donc accéléré et transposé vers le haut).
- [10] Musique : Son rééchantillonné de -4% et transposé-p pour compensation de -4% par l'HARMO.
- [11] Castagnettes : Son original.
- [12] Castagnettes : Son dilaté-p par la Lexicon 2400 de 18,9% (tierce mineure).
- [13] Castagnettes : Son dilaté-p par l'HARMO de 18,9% (tierce mineure).
- [14] Castagnettes : Son dilaté-p par une méthode aveugle de +20%.
- [15] "Cocktail-party" : Son original.
- [16] "Cocktail-party" : Son dilaté-p par la Lexicon 2400 de 18,9% (tierce mineure).
- [17] "Cocktail-party" : Son dilaté-p par l'HARMO de 18,9% (tierce mineure).
- [18] Note de piano : Son original.
- [19] Note de piano : Son dilaté-p par la Lexicon 2400 de 18,9% (tierce mineure).
- [20] Note de piano : Son dilaté-p par l'HARMO de 18,9% (tierce mineure).
- [21] Voix parlée féminine : Son original.
- [22] Voix parlée féminine : Son dilaté-p par la Lexicon 2400 de 18,9% (tierce mineure).
- [23] Voix parlée féminine : Son dilaté-p par l'HARMO de 18,9% (tierce mineure).
- [24] Sinus 110 Hz modulé en amplitude : Signal original (Période fondamentale = 9 ms).
- [25] Sinus 110 Hz modulé en amplitude : Son dilaté-p de 4,2% par une méthode de collage "aveugle". $K = 9$ ms, $FE = 0$ ms.
- [26] Sinus 110 Hz modulé en amplitude : Son dilaté-p de 4,2% par une méthode de collage "aveugle". $K = 9$ ms, $FE = 9$ ms.
- [27] Sinus 110 Hz : Signal original (Période fondamentale = 9 ms).

- [28] Sinus 110 Hz : Son dilaté-p de 4,2% par une méthode de collage "aveugle". $K = 4,5$ ms, $FE = 0$ ms.
- [29] Sinus 110 Hz : Son dilaté-p de 4,2% par une méthode de collage "aveugle". $K = 4,5$ ms, $FE = 9$ ms.
- [30] Voix parlée masculine : Son original.
- [31] Voix parlée masculine : Son dilaté-p +4,2% par une méthode de collage "aveugle". $K = 45$ ms, $FE = 45$ ms.
- [32] Son inharmonique de synthèse : Original.
- [33] Son inharmonique de synthèse : Son dilaté-p par l'HARMO de +4,2%.
- [34] Note de piano la2 : Original.
- [35] Note de piano la2 : Son dilaté-p par l'HARMO de +4,2%.
- [36] Voix d'homme : Son original.
- [37] Voix d'homme : Son dilaté-p (+100%) en modifiant uniquement le spectrogramme (pas de modification du phasogramme).
- [38] Voix d'homme : Son transposé-p (+100%) en modifiant uniquement le spectrogramme (pas de modification du phasogramme).
- [39] Voix d'homme : Son dilaté-p (+100%) par vocodeur de phase classique.
- [40] Bruit blanc original.
- [41] Bruit blanc dilaté-p de +4,2% par une méthode aveugle classique. Pas d'analyse = 491, pas de synthèse = 512, longueur de fenêtre = 2048.
- [42] Bruit blanc dilaté-p de +4,2% par une méthode à pas d'analyse et de synthèse identiques. Pas d'analyse = pas de synthèse = 512, longueur de fenêtre = 2048.
- [43] Orchestre : Son original.
- [44] Orchestre : Son dilaté-p de +20% par une méthode aveugle classique. Pas d'analyse = 205, pas de synthèse = 256, longueur de fenêtre = 1024.
- [45] Castagnettes : Son dilaté-p de +20% par une méthode aveugle classique. Pas d'analyse = 205, pas de synthèse = 256, longueur de fenêtre = 1024.
- [46] Pulsations régulières originales (IOI = 240 ms).
- [47] Pulsations avec un décalage de 6 ms de la 7ème pulsation.
- [48] Pulsations avec un décalage de 12 ms de la 7ème pulsation.
- [49] Pulsations avec un décalage de 24 ms de la 7ème pulsation.
- [50] Pulsations avec un décalage (entraînant le décalage de toutes les pulsations suivantes) de 6 ms à la 7ème pulsation.
- [51] Pulsations avec un décalage (entraînant le décalage de toutes les pulsations suivantes) de 12 ms à la 7ème pulsation.
- [52] Pulsations avec un décalage (entraînant le décalage de toutes les pulsations suivantes) de 24 ms à la 7ème pulsation.
- [53] Pulsations avec un décalage de 6 ms et changement de tempo à la 7ème pulsation.
- [54] Pulsations avec un décalage de 12 ms et changement de tempo à la 7ème pulsation.
- [55] Pulsations avec un décalage de 24 ms et changement de tempo à la 7ème pulsation.
- [56] Orchestre : Son dilaté-p de +20% par une méthode de transposition-p adaptée à l'audition suivie d'un rééchantillonnage (donc durée identique à l'original).
- [57] Orchestre : Son dilaté-p de +20% par une méthode de dilatation-p adaptée à l'audition.
- [58] Castagnettes : Son dilaté-p de +20% par une méthode de transposition-p adaptée à l'audition suivie d'un rééchantillonnage (donc durée identique à l'original).

- [59] Castagnettes : Son dilaté-p de +20% par une méthode de dilatation-p adaptée à l'audition.
- [60] Rock : Son original, puis sous-bande basse-fréquence suivie de sa version dilatée, puis sous-bande haute-fréquence suivie de sa version dilatée ($k_{max} = 2$ ms), et enfin somme des deux signaux dilatés.
- [61] Rock : Son original, puis sous-bande basse-fréquence suivie de sa version dilatée, puis sous-bande haute-fréquence suivie de sa version dilatée ($k_{max} = 10$ ms), et enfin somme des deux signaux dilatés.
- [62] Sinusoïde 500 Hz modulée en fréquence : Son original, puis sous-bande basse-fréquence suivie de sa version dilatée, puis sous-bande haute-fréquence suivie de sa version dilatée ($k_{max} = 10$ ms), et enfin somme des deux signaux dilatés.
- [63] Percussions : Son original, puis sous-bande basse-fréquence suivie de sa version dilatée, puis sous-bande haute-fréquence suivie de sa version dilatée ($k_{max} = 2$ ms), et enfin somme des deux signaux dilatés.
- [64] Percussions : Son original, puis sous-bande basse-fréquence suivie de sa version dilatée, puis sous-bande haute-fréquence suivie de sa version dilatée ($k_{max} = 10$ ms), et enfin somme des deux signaux dilatés.
- [65] Percussions : Son original, puis partie transitoire suivie de sa version dilatée, puis partie résiduelle suivie de sa version dilatée, et enfin somme des deux signaux dilatés.
- [66] Note de piano synthétique (harmonique) la0 (27,5 Hz) : Son dilaté-p de +4,2%, $K_{max} = 25$ ms.
- [67] Note de piano synthétique (harmonique) la0 (27,5 Hz) : Son dilaté-p de +4,2%, $K_{max} = 40$ ms.
- [68] Horloge comtoise (inharmonique) : Son original.
- [69] Horloge comtoise (inharmonique) : Son dilaté-p de +4,2%, $K_{max} = 45$ ms.
- [70] Signal sinusoïdal traité avec une optimisation pour signaux corrélés.
- [71] Signal sinusoïdal traité avec une optimisation pour signaux décorrélés.
- [72] Bruit blanc traité avec une optimisation pour signaux décorrélés.
- [73] Bruit blanc traité avec une optimisation pour signaux corrélés.
- [74] Voix chantée féminine ("Tom's dinner" de Suzanne Vega).
- [75] Son de clavicorde.
- [76] Son de synthèse harmonique décroissant de corde.
- [77] Voix masculines avec bruit de fond.
- [78] Voix féminine avec bruit de fond.
- [79] Accordéon.
- [80] Percussions rythmiques.
- [81] Impulsions de synthèse de fréquence 1 Hz.
- [82] Impulsions de synthèse de fréquence 10 Hz.
- [83] Basse et guitare/batterie ("Muscle Museum" de Muse).
- [84] Sinusoïde pure de 220 Hz modulé en amplitude à 4Hz.
- [85] Sinusoïde pure de 440 Hz.
- [86] Son d'horloge.
- [87] Son de pendule.
- [88] Son de montée d'orchestre.
- [89] Signal constant (Attention au haut-parleur!).

- [90] "La Flûte Enchantée" de W.A. Mozart : Pièce de musique classique originale.
- [91] "La Flûte Enchantée" de W.A. Mozart : Pièce de musique classique originale rééchantillonnée de -20% (donc accéléré et transposé vers le haut).
- [92] "La Flûte Enchantée" de W.A. Mozart : Pièce de musique classique rééchantillonnée de -20% et transposé-p pour compensation de -20% par l'HARMO.
- [93] "La Flûte Enchantée" de W.A. Mozart : Pièce de musique classique originale rééchantillonnée de +20% (donc ralenti et transposé vers le bas).
- [94] "La Flûte Enchantée" de W.A. Mozart : Pièce de musique classique rééchantillonnée de +20% et transposé-p pour compensation de +20% par l'HARMO.
- [95] "Vesoul" de J. Brel : Pièce de musique populaire originale.
- [96] "Vesoul" de J. Brel : Pièce de musique populaire originale rééchantillonnée de -20% (donc accéléré et transposé vers le haut).
- [97] "Vesoul" de J. Brel : Pièce de musique populaire rééchantillonnée de -20% et transposé-p pour compensation de -20% par l'HARMO.

Title

Time-stretching and pitch-shifting of audio signals: Application to cinema/video conversion.

Summary

Coexistence of different formats for cinema (24 frames/s) and video (25 frames/s) involves speeding up or slowing down the soundtrack when converting from one format to another. This causes a temporal modification of the sound signal, and therefore a spectral modification with a change in timbre. Audiovisual post-production studios have to compensate this effect by an appropriate sound transformation.

The aim of this work is to propose to the audiovisual industry a system which allows the counteraction of timbre modification caused by a change in the playback rate. This system consists of a processing algorithm and a machine on which it is implemented. The algorithm is designed to respect sound quality and multichannel compatibility constraints. The machine, named HARMO, is designed for this purpose by the company GENESIS. It is based on digital signal processors and has to respect real-time constraints. The commercial aspect of the project is linked to economic and timing constraints.

A state of the art based on a quasi-exhaustive bibliography leads to an original classification of existing time-stretching and pitch-shifting methods. Well-known time-domain and frequency-domain methods are studied, and time-frequency methods are introduced. This classification allows the creation of several innovative methods:

- two time-frequency methods using an analysis technique adapted to the human ear,
- two coupled methods using advantages of both time- and frequency-domain methods,
- a method which proposes an improvement of time-domain methods.

Algorithms are evaluated using a bank of test sounds specially designed to highlight characteristic artifacts. The time-domain approach is selected and optimized thanks to criteria based on normalized autocorrelation and detection of transients. This algorithm is integrated into a software designed for multichannel real-time running, and implemented on the HARMO hardware.

Keywords

Time-stretching, pitch-shifting, sound transformation, audio digital signal processing, multimedia conversion, real-time, computer music, audio effects.

Résumé

La coexistence de deux formats : cinéma à 24 images/s et vidéo à 25 images/s, implique l'accélération ou le ralentissement de la bande-son lors du transfert d'un format vers l'autre. Ceci provoque une modification temporelle du signal sonore, et par conséquent une modification spectrale avec altération du timbre. Les studios de post-production audiovisuelle souhaitent compenser cet effet par l'application d'une transformation sonore adéquate.

L'objectif de ce travail est de fournir à l'industrie audiovisuelle un système permettant de pallier la modification de timbre engendrée par le changement de vitesse de lecture. Ce système se compose d'une part d'un algorithme de traitement et d'autre part d'une machine sur lequel il est implanté. L'algorithme est conçu et développé pour répondre aux contraintes liées à la qualité sonore et à la compatibilité multicanal. La machine, baptisée HARMO, est conçue spécifiquement par la société GENESIS sur la base de processeurs de signaux numériques, et doit répondre à la contrainte de temps-réel. Cet aspect "valorisation" conduit à intégrer dans le projet les contraintes de coût et de délai de réalisation.

Un état de l'art basé sur une bibliographie quasi-exhaustive aboutit à une classification originale des méthodes de dilatation et de transposition existantes. Ceci nous amène à distinguer et à étudier les méthodes classiques temporelles et fréquentielles, et à introduire les méthodes temps-fréquence. Cette classification est à la base de plusieurs méthodes innovantes :

- deux méthodes temps-fréquence dont l'analyse est adaptée à l'audition,
- deux méthodes couplées qui associent les avantages des méthodes temporelles et fréquentielles,
- une méthode temporelle basée sur une amélioration des méthodes existantes.

Les algorithmes sont évalués grâce à une banque de sons-test spécifiquement élaborée pour mettre en évidence les défauts caractéristiques des algorithmes. Notre choix final s'est porté sur l'approche temporelle, que nous optimisons par l'adjonction de critères de segmentation basés sur l'autocorrélation normalisée et la détection de transitoires. Cet algorithme s'intègre dans un logiciel qui a été structuré pour un fonctionnement temps-réel et multicanal sur le système HARMO.

Mots clefs

Dilatation, transposition, transformation des sons, traitement du signal audio, conversion multimedia, temps réel, informatique musicale, effets audionumériques.

Discipline

Acoustique, Traitement du signal et Informatique Appliqués à la Musique (ATIAM).

Laboratoire d'accueil

Laboratoire de Mécanique et d'Acoustique

CNRS - UPR 7051

31 chemin Joseph-Aiguier, 13402 Marseille cedex 20.