Mémoire de stage de DEA ATIAM année 2003-2004

Application du modèle additif « shape invariant » pour la transformation de la voix

---

Gaël Champion, Université Paris VI.
Réalisé à l'IRCAM.

Responsable de stage : Axel Roebel

30 Juin 2004.

ircam
Centre Pompidou

Table des matières

Je tiens à remercier Xavier Rodet et toute l'équipe Analyse-synthèse de l'IRCAM pour leur accueil,
Axel Roebel pour sa disponibilité et son aide tout au long du stage,
Toute l'équipe du DEA ATIAM de nous offrir encore et toujours la possibilité unique d'une formation scientifique et musicale,
Cyrille Defaye pour sa sympathique présence et tous les services qu'elle est toujours prête à rendre,
La promo du dernier DEA ATIAM pour cette année passée…

Introduction

Changer la durée, transposer un signal de parole sont des opérations délicates si l'on veut conserver le réalisme de la voix. Pour cela, il est nécessaire de bien prendre en compte les différentes natures d'un signal de parole, périodiques ou bruitées, en d'autres termes, voisées ou non voisées.

La synthèse additive est largement utilisée pour des transformations de type dilatation/compression temporelle (time stretch) ou transposition (pitch shift). Cependant, l'utilisation de techniques dites « shape invariant » (invariance de la forme d'onde) est nécessaire pour maintenir le réalisme des parties voisées. La première partie du projet consiste à étudier voire améliorer (on verra par la suite que le modèle a été modifié) la synthèse additive « shape invariant ».

Dans un deuxième temps, on s'intéresse aux régions bruitées, pour lesquelles le modèle additif est inadapté. On propose ici d'utiliser une génération de bruit qui soit robuste aux transformations, ce qui n'est pas le cas des méthodes plus classiques basées sur le résiduel.

Enfin, la prise en compte de la caractérisation voisée/non voisée du signal permet d'intégrer dans un seul modèle le « shape invariant » et la génération de bruit.
L'objectif final du projet est d'investiguer le modèle additif pour la transformation de la voix, d'en montrer éventuellement les limites, et les pistes qu'il serait nécessaire de poursuivre.

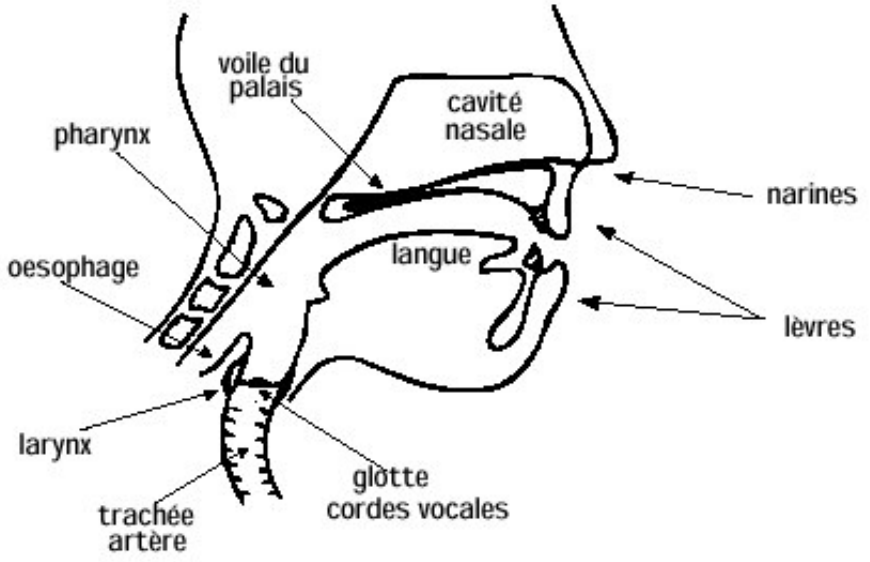
Nature du signal de parole

Pouvoir synthétiser un signal de parole réaliste, qu'il soit parlé ou chanté, présuppose de bien connaître toutes ses spécificités. C'est en comprenant comment l'organe vocal produit des sons que l'on peut dans un deuxième temps tenter d'approcher ces caractéristiques physiques par des caractéristiques sur les signaux. On verra notamment en quoi la distinction entre les sons voisés et les sons non-voisés impacte et complexifie les modèles de signaux liés à la parole (synthèse, reconnaissance, transformation, codage bas débit…).

1 La production de la parole [4][5]

1.1 Description de l'appareil phonatoire

La figure ci-dessous montre schématiquement l'appareil phonatoire.

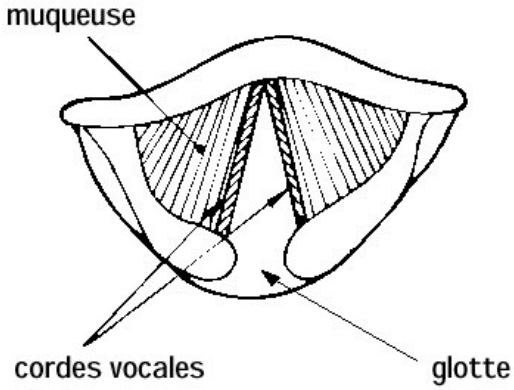


[Fig. 1] Schéma de l'appareil phonatoire, d'après [4]

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques.

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le larynx où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la

trachée (cf. figure suivante). Les cordes vocales sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée glotte. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la luette est en position basse, la cavité nasale vient s'y ajouter en dérivation.



[Fig. 2] Section du larynx, vu de haut, d'après [4]

1.2 Sons produits par l'organe vocal

Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur mode articulatoire. On distingue généralement trois classes principales : les voyelles, les semi-voyelles et les liquides, et les consonnes.

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la bouche se réduit alors à une modification du timbre vocalique. Si, au contraire, le passage se rétrécit par endroit, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. La bouche est dans ce cas un organe de production à part entière. Les semi-voyelles (dans le son « oui » en français, par exemple), quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes. Enfin, les liquides (les sons « l » et « R ») sont assez difficiles à classer. L'articulation du « l » ressemble à celle d'une voyelle, mais la position de la langue conduit à

une fermeture partielle du conduit vocal. Le son « R » (exemple : « rond »), quant à lui, admet plusieurs réalisations fort différentes.

Les voyelles se différencient principalement les unes des autres par leur lieu d'articulation, leur aperture, et leur nasalisation. On distingue ainsi, selon la localisation de la masse de la langue, les voyelles antérieures, les voyelles moyennes, et les voyelles postérieures, et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles fermées et ouvertes. Enfin, citons aussi le distinguo entre voyelles orales et voyelles nasales (par exemple, le son « on ») pour lesquelles le voile du palais est abaissé, mettant en parallèle les cavités nasales et buccale. Notons que, dans un contexte plus général que celui de la seule langue française, d'autres critères peuvent être nécessaires pour différencier les voyelles, comme leur labialisation, leur durée, leur tension, leur stabilité, leur glottalisation, voire même la direction du mouvement de l'air.

On classe principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, et de leur nasalisation. Comme pour les voyelles, d'autres critères de différenciation peuvent être nécessaires dans un contexte plus général : l'organe articulatoire, la source sonore, l'intensité, l'aspiration, la palatalisation, et la direction du mouvement de l'air.

En français, la distinction de mode d'articulation conduit à deux classes de consonne : les fricatives (ou constrictives) et les occlusives (ou plosives).

Les fricatives sont créées par une constriction du conduit vocal au niveau du lieu d'articulation, qui peut être le palais (par exemple, dans le mot « jeu »), les dents (par exemple, le son « s »), ou les lèvres (par exemple, dans le son « feu »). Les fricatives non-voisées sont caractérisées par un écoulement d'air turbulent à travers la glotte, tandis que les fricatives voisées combinent des composantes d'excitation périodique et turbulente : les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais complète.

Les occlusives correspondent quant à elles à des sons essentiellement dynamiques. Une forte pression est créée en amont d'une occlusion maintenue en un certain point du conduit vocal (qui peut ici aussi être le palais (par exemple, le son « k »), les dents (par exemple, le son « t »), ou les lèvres (par exemple, le son « p »)), puis relâché brusquement. La période d'occlusion est appelée la phase de tenue. Pour les occlusives voisées (par exemple, le son « b ») un son basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue; pour les occlusives non voisées (par exemple, le son « t »), la tenue est un silence.

Enfin, les consonnes nasales (par exemple, dans le mot « me ») font intervenir les cavités nasales par abaissement du voile du palais.

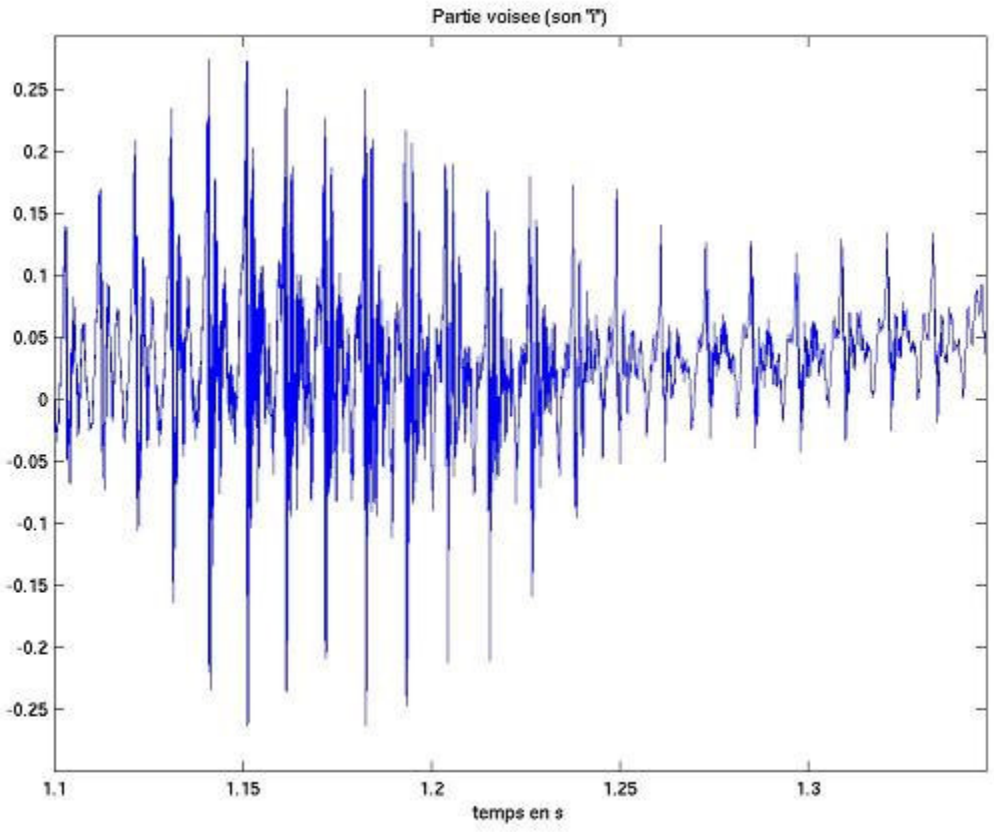
2 Caractéristiques du signal de parole

On peut relier certaines caractéristiques du signal à la production du son :

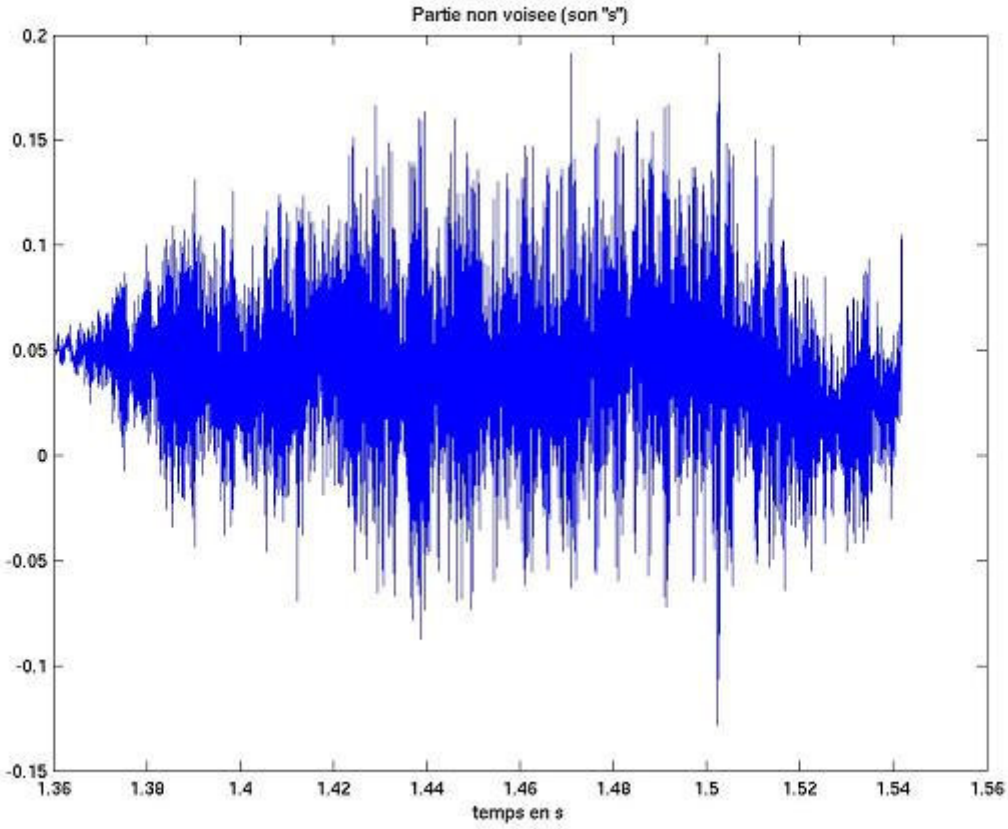
- L'intensité du son est liée à la pression de l'air en amont du larynx.
- Sa fréquence (le pitch), qui n'est rien d'autre que la fréquence du cycle d'ouverture/fermeture des cordes vocales, est déterminée par la tension de muscles qui les contrôlent.
- Son spectre résulte du filtrage dynamique du signal glottique (impulsions, bruit, ou combinaison des deux) par le conduit vocal, qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses.

La seule grandeur perceptuelle pour laquelle il n'existe aucun lien direct avec le signal est le timbre.

La première caractéristique d'un son de parole, visible sur le signal temporel est son voisement (cf. figures suivantes).



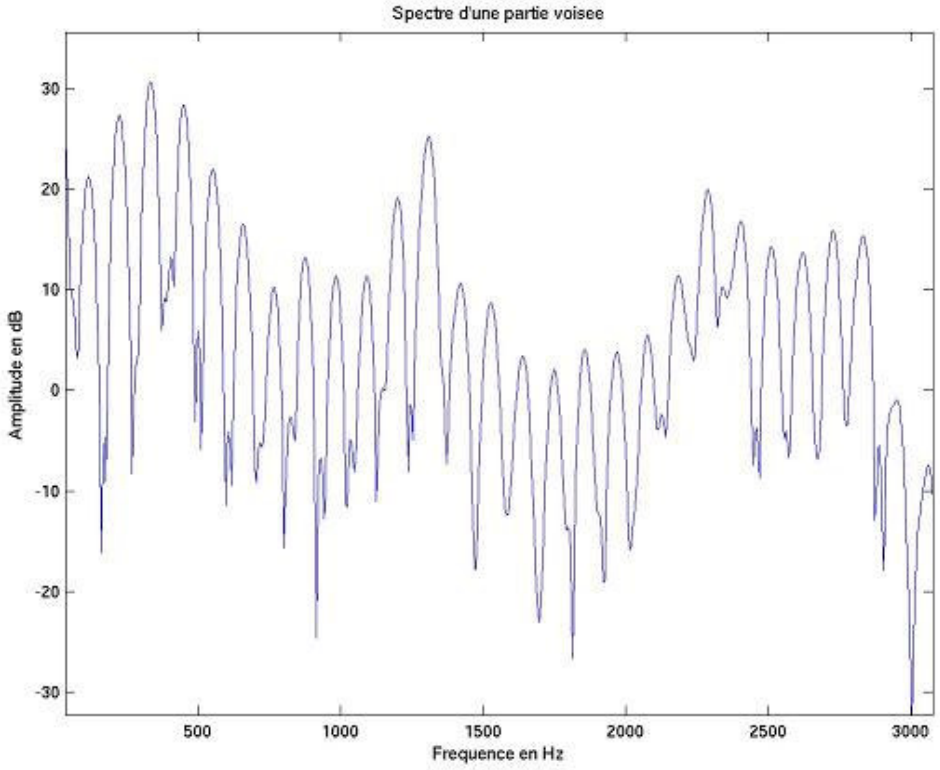
[Fig. 3] Extrait d'un signal de parole pour un son voisé (échantillonné à 44100Hz)



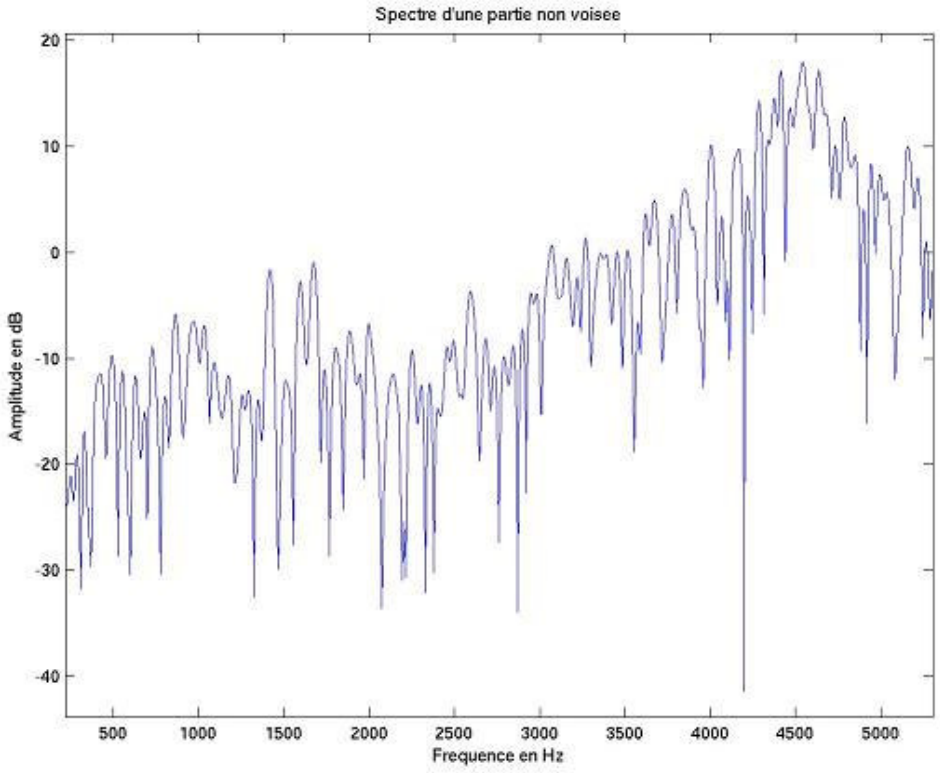
[Fig. 4] Extrait d'un signal de parole pour un son non voisé (échantillonné à 44100Hz)

Un son voisé est par nature périodique, tandis qu'un son non voisé est par nature bruité. Cependant, il faut garder en tête que certains sons sont à la fois voisés et non voisés et qu'on ne peut pas nécessairement considérer un son voisé comme l'étant sur toutes ces fréquences. On pourra d'ailleurs définir une fréquence maximale de voisement.

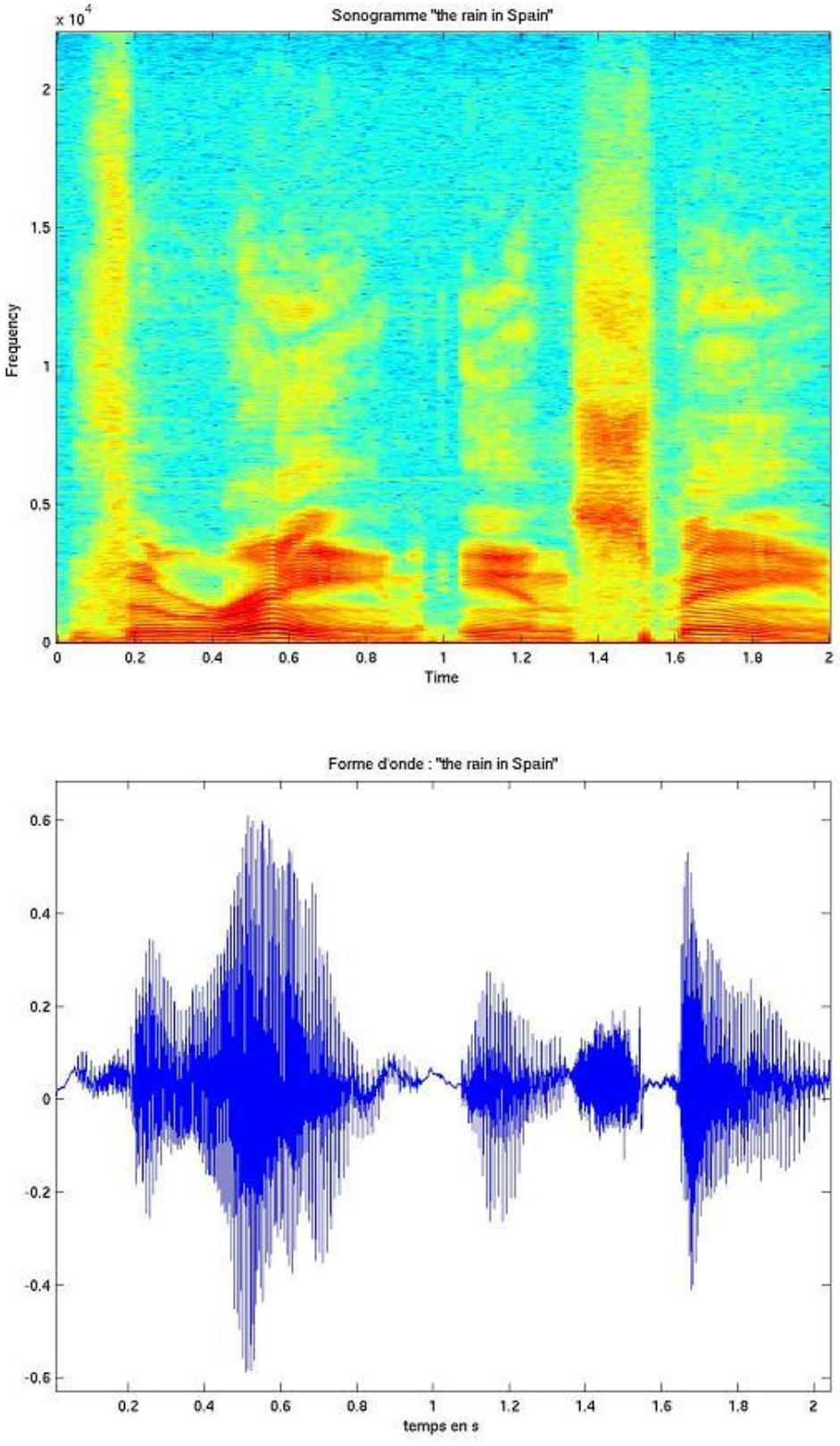
Spectralement, les parties voisées du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière. La forme générale de ces spectres, appelée enveloppe spectrale, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés formants et anti-formants. Enfin, l'évolution temporelle de leur fréquence centrale et de leur largeur de bande impacte le timbre du son.



[Fig. 5] Spectre d'une partie voisée



[Fig. 6] Spectre d'une partie non voisée



[Fig. 7] Sonogramme (en haut) et forme d'onde (en bas) de la phrase parlée par un homme : « the rain in Spain »

La fréquence fondamentale (ou pitch) évolue lentement dans le temps à l'intérieur des zones voisées. Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants.

De toutes ces caractéristiques, on comprend pourquoi le modèle source-filtre est particulièrement bien adapté à la production de la parole. La source est un signal périodique pour les sons voisés, elle est un bruit pour les sons non voisés. Enfin, le conduit vocal est modélisé par un filtrage variant lentement dans le temps.

Ainsi, dans une analyse additive, le modèle harmonique sera particulièrement approprié pour les sons voisés, tandis qu'un modèle avec générateur de bruit, éventuellement couplé à une analyse non harmonique sera nécessaire pour la production des sons voisés.

Le modèle additif

Ce modèle a été proposé par McAulay et Quatieri en 1986 [1]. L'idée est de modéliser un son par une collection de trajets sinusoïdaux. Un trajet décrit les variations dans le temps en amplitude, phase (ou fréquence) d'une composante sinusoïdale ou partiel du signal audio analysé. Ce modèle paraît par nature plus approprié pour des signaux harmoniques mais il peut finalement s'appliquer à tous types de signaux : harmoniques ou non, monophoniques ou polyphoniques mais plus difficilement à des signaux de nature bruitée.

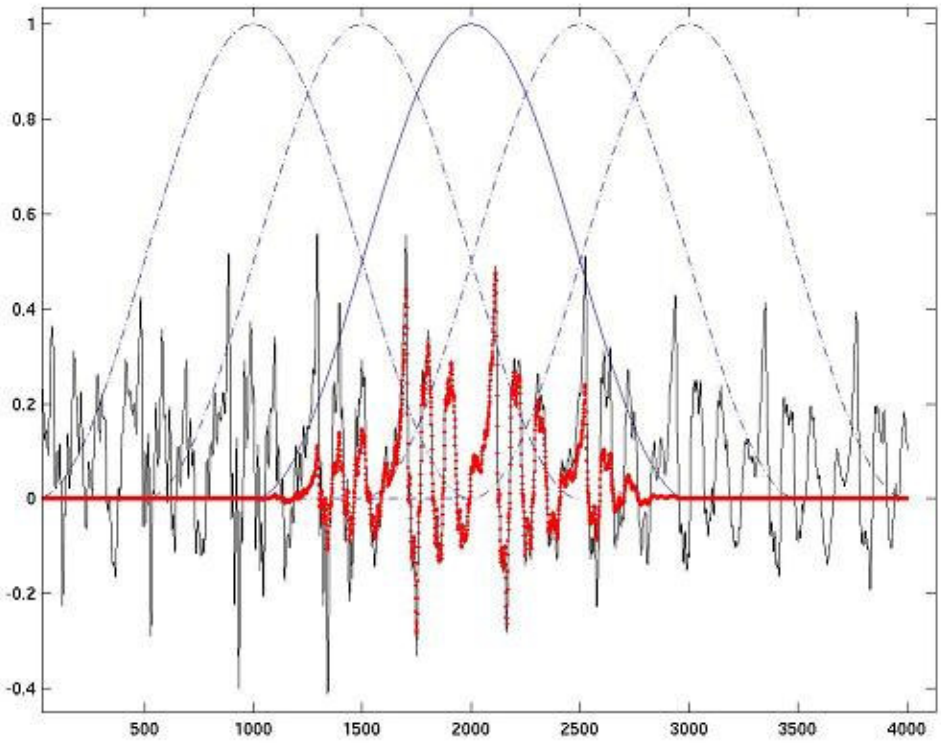
1 Principes

1.1 L'analyse

L'analyse est basée sur la transformée de Fourier à court terme (notée TFCT par la suite). L'objectif final étant d'en extraire un jeu de trois paramètres (amplitude, phase et fréquence) pour les composantes sinusoïdales constitutives du son.

Rappelons tout d'abord les principes de la TFCT. Considérons un signal temporel *ft).* La TFCT consiste à effectuer la transformée de Fourier du signal *f(t)* fenêtré par une fenêtre *g(t)* glissant dans le temps [11].

Sur la figure suivante, on voit en traits pleins le signal dont on cherche à calculer la TFCT, en pointillés la fenêtre glissante et au centre, en traits épais, on voit le résultat du fenêtrage sur le signal audio. Ici, le signal est un signal de parole, échantillonné à 44100 Hz, la fenêtre est une fenêtre de Hanning de 2000 échantillons et le recouvrement est de 1500 échantillons.



[Fig. 8] Illustration du fenêtrage pour la TFCT

A l'instant t_0, la TFCT *S(t,w)* décrit le spectre local de *f(t)* en fonction de ω. C'est la transformée de Fourier de $f(t)g(t-t_0)$ où *g(t)* est une fenêtre (généralement à support compact) centrée sur t_0. En déplaçant *g* au cours du temps, on obtient une image fréquentielle et temporelle de *f* (*f* et *g* sont des fonctions réelles) .

$$S(t,w) = \int_{\mathbb{R}} \left(f(s)g(s-t)e^{-i\omega s} ds \right) \tag{1.1}$$

Son équivalent en temps et fréquence discrets (où f_s est la fréquence d'échantillonnage et ***L***, le nombre de points pour la transformée de Fourier) est donné par la formule suivante :

$$S(n,\omega_k) = \sum_{s=-\infty}^{+\infty} f(s)g(s-n).e^{-i\omega_k s} \text{ avec } 2\pi\omega_k = \frac{k}{L} f_s \tag{1.2}$$

Généralement, on fixe les paramètres de la TFCT suivant pour l'analyse additive :

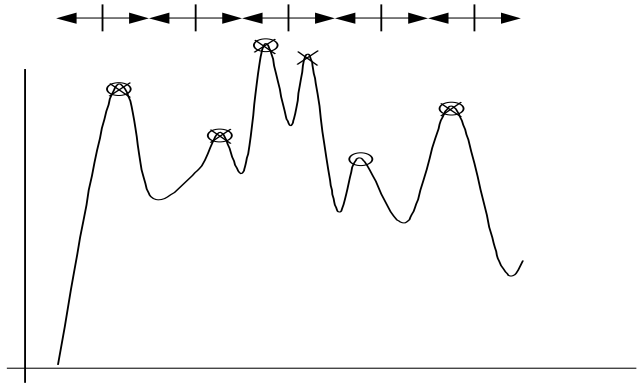
- Fenêtre de longueur *N* au moins égale à 3 fois la période fondamentale maximale (on prendra plutôt de 3 à 4 fois cette période). Les fenêtre couramment utilisées sont les fenêtre de Hanning ou de Blackman
- Recouvrement temporel (overlap) d'environ 3*N/4* (le pas d'avancement de l'analyse est de *N/4*)

Une fois les TFCT calculées, on cherche à en extraire des trajets de partiels. On cherche à extraire en général pour chaque TFCT un même nombre maximal de partiels noté M (et qui ne pourra pas toujours être atteint). On extrait couramment moins d'une centaine de partiels.

Deux types d'analyse sont à distinguer ici :

- L'analyse harmonique : elle nécessite une estimation de la fréquence fondamentale f_0 du signal. Puis, elle n'extrait qu'une sinusoïde pour chaque multiple de f_0, la recherche de l'harmonique k ($k \leq M \leq \frac{f_s}{f_0}$) se faisant dans une fenêtre de largeur inférieure ou égale à f_0 et centrée sur $k.f_0$
- L'analyse non harmonique : elle consiste généralement à sélectionner les M pics sinusoïdaux de plus fortes amplitudes.

La figure suivante illustre simplement les différents types d'analyse. Elle représente schématiquement une TFCT où la fréquence fondamentale (et ses harmoniques) estimés sont donnés par les petits traits verticaux tandis que la fenêtre de recherche de chaque harmonique est représentée par les intervalles entre double-flèches. Si l'on cherche à sélectionner 5 sinusoïdes, dans l'analyse harmonique, on retiendra les partiels illustrées d'un cercle (i.e. qu'on ne retiendra qu'une seule sinusoïde par fenêtre) et dans l'analyse non harmonique, on retiendra les 5 plus importants, ici notés d'une croix.



[Fig. 9] Schéma de principe d'une analyse additive harmonique et non harmonique

Une fois que l'on a sélectionné les pics que l'on souhaitait extraire, il faut calculer le jeu de paramètres correspondant. Bien entendu, l'estimation de ces paramètres est directement liée à la précision de la FFT. Ainsi, pour une analyse sur 4096 points et une fréquence d'échantillonnage de 44100Hz, on a une précision fréquentielle de seulement

44100/4096=10.77Hz ce qui est tout-à-fait insuffisant pour la synthèse additive.

La méthode couramment utilisée pour estimer la fréquence des pics est une interpolation polynomiale de degré 2 ou 3. Nous ne détaillerons pas cet aspect ici.

Une fois la fréquence du pic estimée, on peut simplement extraire la phase des pics par la formule suivante :

$$\varphi(n,k) = \arg(S(n,k)) + 2\pi \frac{N}{2f_s} f \text{ où } k = \frac{f_k}{f_s} L \qquad (1.3)$$

où f est la fréquence estimée, n l'indice en temps discret, f_s la fréquence d'échantillonnage, f_k la fréquence du pic, L la longueur de la FFT et N la longueur de la fenêtre. Une fois de plus, la phase sera estimée à la fréquence du pic extrait par une interpolation linéaire.

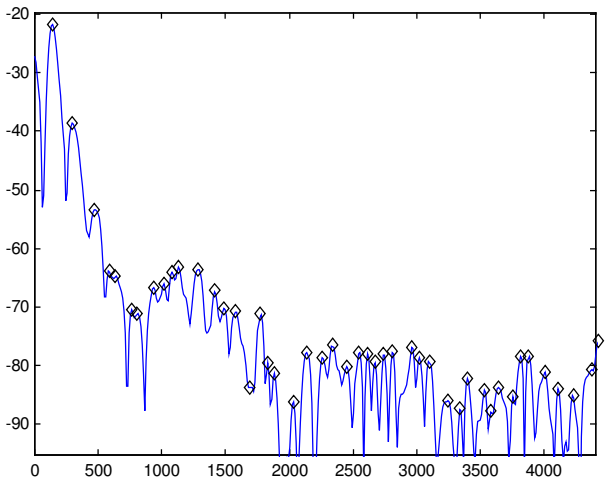
Enfin, l'amplitude est donnée par la formule suivante :

$$A(n,k) = \frac{2\|S(n,k)\|}{\sum_{i=1}^{N} g(i)} \text{ où } g \text{ est la fenêtre utilisée pour la TFCT} \qquad (1.4)$$

Bien entendu, il existe des méthodes plus sophistiquées et plus pertinentes pour estimer ces paramètres reposant par exemple sur :

- La discrimination sinusoïde-bruit
- La prise en compte de la pente de la trajectoire fréquentielle
- …

La figure ci-dessous illustre une partie de TFCT d'un signal audio et on a représenté par des points les pics extraits par l'analyse additive.



(fréquences en Hz en abscisses, et amplitude en dB en ordonnée)

[Fig. 10] Illustration sur la TFCT de l'extraction des pics pour l'analyse additive

1.2 Le suivi des trajets des partiels

A la synthèse (cf. partie suivante), on connectera entre eux les paramètres supposés correspondre à une même sinusoïde, i.e. décrivant un même trajet. Si la sinusoïde « meurt », elle ne doit plus être connectée.

Une des difficultés du modèle additif réside donc dans l'analyse et le suivi des trajets. On veut idéalement :

- Ne sélectionner que les pics sinusoïdaux : il peut y avoir confusion entre des pics de sinusoïdes et des zones bruitées qui émergent du spectre ou même entre des pics de sinusoïdes et les lobes secondaires de la TFCT d'une sinusoïde.
- Suivre tous les trajets qu'il est nécessaire de suivre mais ne pas connecter des sinusoïdes qui ne correspondent pas à un trajet

Le suivi est a priori plus simple dans une analyse harmonique. En effet, on est sûr alors de devoir connecter tous les pics de même rangs (qui sont supposés correspondre au trajet d'une même harmonique) [8].

Dans un modèle non harmonique, on relie en général les partiels dont les fréquences sont les plus proches (on peut aussi affiner ce critère en prenant en compte la proximité en amplitude) [1].

1.3 La synthèse

Une fois que l'on a extrait les trajets des partiels, on cherche à resynthétiser un signal temporel sous la forme :

$$s(n) = \sum_{i=1}^{M} A_i(n)\cos(\Phi(n)) \tag{1.5}$$

où les $A_i(n), f_i(n)$ et $\varphi_{i,p}$ sont respectivement les amplitudes, fréquences et phases initiales du $i^{ème}$ partiel à l'instant discret $n=pT$ où T est l'incrément temporel entre chaque TFCT ($T = N - O$ où N la longueur de la fenêtre d'analyse et O le nombre d'échantillons de recouvrement entre deux TFCT successives) et p l'indice de trame. $\Phi(n)$ est la phase synthétisée, fonction de $f_i(n)$ et $\varphi_{i,p.}$

On suppose pour l'instant qu'on resynthétise le signal sans transformation et on ne prend pas en compte un éventuel résiduel.

Pour l'instant, on ne dispose des valeurs d'amplitude, fréquence et phase qu'à chaque

instant de calcul de la TFCT.

On distingue trois cas :

- Il existe un trajet reliant le partiel i à l'instant d'analyse $(t = n) = pT$ au partiel i à l'instant d'analyse $(t = n') = (p+1)T$.

Soit $n \le t \le n + T$ et $m = t - n$,

On interpole les amplitudes linéairement :

$$A_i(m) = A_{i,p} + m\frac{A_{i,p+1} - A_{i,p}}{n' - n} \tag{1.6}$$

avec $A_{i,p}$ l'amplitude du pic de rang i extrait à la trame p.

Il faut désormais calculer le trajet de phase, hors les phases ne sont connues qu'à 2π près. Ainsi, une contrainte supplémentaire est nécessaire pour calculer la phase.

McAulay et Quatieri ont proposé une méthode basée sur une interpolation polynomiale de la phase en imposant de minimiser les variations de fréquence instantanée pendant une trame de synthèse.

On pose

$$\begin{aligned} \varphi(m) &= a + b.m + c.m^2 + d.m^3 \\ \omega(m) &= 2\pi f(m) = \dot{\varphi}(m) = b + 2c.m + 3d.m^2 \end{aligned} \tag{1.7}$$

avec les conditions initiales suivantes :

$$\begin{cases} \varphi(0) = a = \varphi_{i,p} \\ \omega(0) = b = 2\pi f_{i,p} \\ \varphi(n' - n) = \varphi_{i,p+1} + 2k\pi, \qquad k \in \mathbb{N} \\ \omega(n' - n) = 2\pi f_{i,p+1} \end{cases} \tag{1.8}$$

avec $f_{i,p}$ et $\varphi_{i,p}$ les fréquences et phases extraites du pic de rang i extrait à la trame p.

La solution (dépendant de k puisque la phase n'est connue qu'à $2k\pi$ près) est donnée par :

$$\begin{pmatrix} c(k) \\ d(k) \end{pmatrix} = \begin{pmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{pmatrix} \times \begin{pmatrix} \varphi_{i,p+1} - \varphi_{i,p} - \omega(0)T + 2k\pi \\ \omega(n' - n) - \omega(0) \end{pmatrix} \tag{1.9}$$

où $T=n'-n$.

On calcule ensuite k de manière à minimiser les variations fréquentielles. McAulay et

Quatieri ont fixé comme critère de minimiser la fonction suivante :

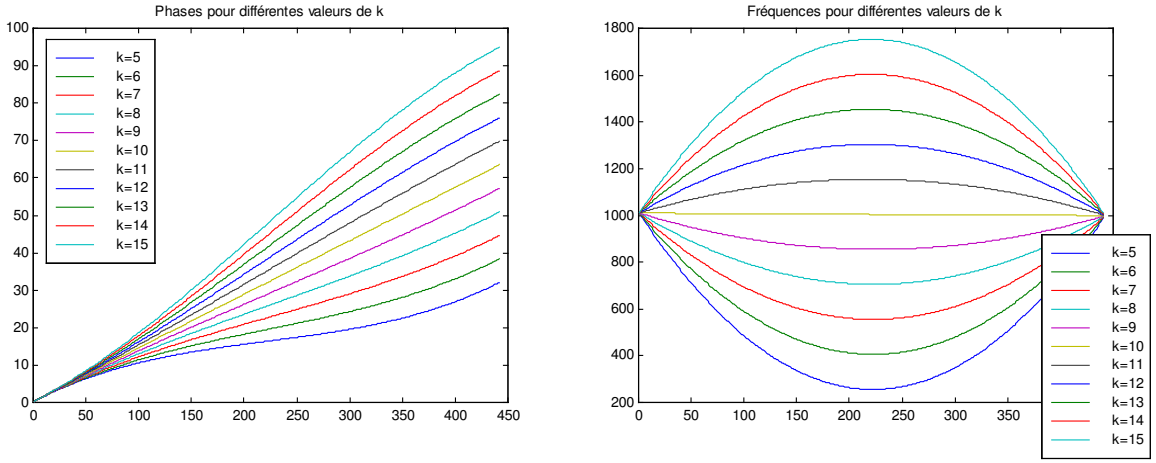
$$g(k) = \int_0^T \left(\dot{\omega}(t,k)\right)^2 dt \tag{1.10}$$

La solution mathématique est

$$k^* = \frac{1}{2\pi}\left[\left(\varphi_{i,p} + \omega(0)T - \varphi_{i,p+1}\right) + \left(\omega(n'-n) - \omega(0)\right)\frac{T}{2}\right] \tag{1.11}$$

et on prend pour valeur entière *k*, l'entier le plus proche de *k**.

Les figures suivantes illustrent les différents trajets de phase (à gauche) et les trajets de fréquences correspondant (à droite) répondant aux valeurs de l'analyse mais pour différentes valeurs de *k* (la valeur 10, dans ce cas étant l'optimale résultant du calcul).



[Fig. 11] Trajets de phase et fréquence synthétisés pour différentes valeur de k

- Le trajet meurt à l'instant d'analyse en cours (aucune connexion n'a été réalisée avec l'un des partiels de la trame suivante) .

  Dans ce cas, on synthétise :

$$\left(A_{i,p} - m\frac{A_{i,p}}{n'-n}\right)\cos\left(\theta(m)\right) \qquad \text{avec} \quad \theta(m) = \varphi_{i,p} + 2.\pi.f_{i,p}\frac{m}{f_s} \tag{1.12}$$

  Le signal a une amplitude qui décroît linéairement jusque 0 et une fréquence constante.

- Le trajet naît à l'instant d'analyse suivant (aucune connexion n'a été réalisée avec un partiel précédent) .

  Dans ce cas, on synthétise :

$$\left(m\frac{A_{i,p+1}}{n'-n}\right)\cos\left(\theta(m)\right)$$
$$\text{avec} \quad \theta(m)=\left(\varphi_{i,p+1}-2.\pi.f_{i,p+1}\frac{T}{f_s}\right)+2.\pi.f_{i,p+1}\frac{m}{f_s} \tag{1.13}$$

Le signal a une amplitude qui croît linéairement depuis 0 et une fréquence constante.

2 Transformations

Dans la plupart des cas (hormis les cas de signaux bruités), si les paramètres de l'analyse sont bien choisis et les partiels bien estimés, le signal synthétisé est perceptivement indistinguable de l'original. On a donc réussi à représenter le signal avec un nombre réduit de paramètres qu'il paraît facile de contrôler avant la synthèse.

S'il est évident qu'en choisissant délibérément de ne pas resynthétiser certaines fréquences, on réalise un filtrage : on peut aussi assez facilement réaliser des transpositions (ou pitch shift) sans changer le rythme et/ou des étirements ou compressions temporelles (time stretch) sans changer la hauteur.

- Transposition

Il suffit en fait de multiplier les fréquences issues de l'algorithme de synthèse $\omega(m)$ par un facteur de transposition ***B***.

- Etirement ou compression temporelle

Il suffit de réechantillonner les $A_i(m)$ et $\omega(m)$ aux instants $m'=D^{-1}m$ où D est le facteur de dilatation.

D'après (1.7), on aura donc pour la fréquence et la phase :

$$\omega(m')=2\pi f(m')=b+2c\frac{m'}{D}+3d\left(\frac{m'}{D}\right)^2$$
$$\varphi(m')=\int w(m')dm'=a+b.m'+\frac{c}{D}m'^2+\frac{d}{D^2}m'^3 \tag{1.14}$$

Enfin, on peut imaginer toute autre sorte de transformation en modifiant indépendamment amplitude, fréquence ou phase des partiels analysés.

3 Avantages et limitations

Les avantages du modèle additif se déduisent assez facilement de ses principes :

- C'est une méthode efficace de synthèse des sons et perceptivement de haute qualité
- Comparativement au vocodeur de phase, on ne garde qu'un nombre réduit de paramètres de l'analyse pour effectuer la synthèse
- On peut aisément réaliser des transformations sur le signal
- C'est une méthode qui peut être utilisée en débruitage (la sélection même des pics limite voire supprime la resynthèse du bruit)

Cependant, on peut lui opposer les inconvénients suivants :

- Elle est coûteuse en calcul (dans la phase d'analyse notamment) ce qui peut poser problème si l'on cherche des applications temps réel
- La partie analyse et suivi des trajets est délicate et conditionne la qualité du signal synthétisé : des artéfacts peuvent facilement apparaître
- Elle ne prend pas directement en compte le bruit. Il peut cependant être resynthétisé en ajoutant le résiduel (signal synthétisé – signal original) au signal synthétisé (à condition de ne pas opérer de transformations majeures sur le signal synthétisé)
- Elle n'est donc pas adapté aux signaux totalement bruités comme notamment les parties non voisées d'un signal de parole
- En analyse harmonique, elle nécessite une estimation de la fondamentale (ce qui est toujours une opération délicate et pas toujours très fiable)

La synthèse « shape invariant »

Les parties voisées d'un signal de parole, qui sont par nature harmoniques, sont générées par une excitation périodique pour laquelle on peut définir des instants de synchronisation des phases des différentes harmoniques. Le passage par le conduit vocal, dont les caractéristiques changent dans le temps, provoque un filtrage sur l'excitation ce qui casse le synchronisme des phases (le filtrage introduit un déphasage variable au cours du temps entre le fondamental et les harmoniques). Cependant, le réalisme du signal est garanti par le maintien de ce déphasage aux mêmes instants, il faut donc conserver pour ces instants de synchronisation le déphasage introduit par le conduit vocal.

Lorsque l'on applique une transformation de type time-stretch ou pitch-shift sans prendre de précaution particulière en utilisant les algorithmes de calcul de McAulay et Quatieri décrits précédemment, on casse ce synchronisme. En effet, dans ce cas, on force la phase de synthèse à être égale à la phase d'analyse et bien entendu, la valeur absolue de la phase ne traduit en rien le déphasage relatif entre l'harmonique et le fondamental. Perceptivement, cela se traduit par un son métallisé, réverbérant ou ressemblant au signal passé dans un effet de type flanger ou phaser, et qui perd son réalisme. Pour parer à cet inconvénient, il existe des techniques dites « shape invariant » qui garantissent l'invariance de la forme d'onde du signal synthétisé.

Les algorithmes décrits ci-après concernent uniquement les parties harmoniques d'un signal. De même, on ne s'intéresse ni à la modélisation du bruit ni au résiduel.

1 L'algorithme "shape invariant" de McAulay et Quatieri [2]

En 1992, McAulay et Quatieri ont publié un article proposant une méthode « shape invariant » pour la transformation de la voix.

Nous n'entrerons pas ici dans le détail de la technique car elle n'a pas été utilisée et se révèle plus compliquée que la méthode de Federico décrite ci-après, pourtant parfaitement équivalente [6].

Le signal de parole $s(t)$ peut être modélisé comme la sortie d'un système source-filtre, où $e(t)$ représente l'excitation et $v(t)$ la réponse impulsionnelle d'un filtre linéaire .

L'excitation est une série d'impulsions périodiques, dont la période correspond au « pitch » du locuteur, pour un signal voisé. Pour un signal non voisé, l'excitation est un bruit.

Le filtre représente les caractéristiques et les variations du conduit vocal.

Pour le modèle sinusoïdal, ***e(t)*** sera représenté par une somme de sinusoïdes.

$$e(t) = \sum_i A_{e,i}(t)\cos\left(\varphi_{e,i}(t)\right) \tag{2.1}$$

La réponse fréquentielle du filtre peut s'écrire :

$$V(\omega,t) = M(\omega,t)\exp\left[j\cdot\psi(\omega,t)\right] \tag{2.2}$$

On en déduit le gain et la phase donnés par le filtre pour chaque trajectoire de fréquence de l'excitation :

$$\begin{aligned} A_{v,i}(m) &= M(\omega_i,t) \\ \varphi_{v,i}(m) &= \psi(\omega_i,t) \end{aligned} \tag{2.3}$$

avec $\omega_i = 2\pi f_i$ où f_i est la fréquence du $i^{ème}$ partiel et m l'indice temporel discret correspondant à l'instant d'analyse.

On en déduit pour le signal synthétisé :

$$s(m) = \sum_i A_i(m)\cos(\varphi_i(m)) = A_{v,i}(m)A_{e,i}(m)\cos\left(\varphi_{e,i}(m)+\varphi_{v,i}(m)\right) \tag{2.4}$$

Pour les amplitudes, c'est directement l'amplitude extraite qui correspond au produit des amplitudes liées au conduit vocal et celles qui sont liées à l'excitation (il est inutile de chercher à les dissocier).

On s'intéresse aux parties voisées (il n'y a aucun besoin de « shape invariant » pour les parties non voisées).

Les composantes du signal d'excitation sont supposées être toutes en phase aux instants t_0 de fermeture de la glotte (le signal source est périodique et harmonique dans ce cas), ce qui impose donc dans cette méthode de les estimer. La phase du signal source à l'instant d'analyse m peut donc s'exprimer en terme de décalage par rapport à ces instants t_0 :

$$\varphi_{e,i} = (m - t_0).\omega_i(m) \tag{2.5}$$

Pour obtenir la phase du filtre, il suffit alors de soustraire $\varphi_{e,i}$ à la phase d'analyse.

Lors d'une dilatation d'un facteur D, de nouveaux instants t_0' (obtenus par sous-échantillonnage d'un facteur D des périodes définissant t_0) de fermeture de la glotte sont définis. La préservation de la forme d'onde est obtenue en gardant les phases du signal source égales entre elles aux instants t_0'.

$$\hat{\varphi}_{e,i}(m') = (m' - t_0').\hat{\omega}_i(m') \tag{2.6}$$

La contribution de phase du filtre $\varphi_{v,i}(m)$ reste inchangée et la phase du signal de synthèse s'exprime donc :

$$\hat{\varphi}_i(m') = \varphi_{v,i}(m') + (m' - t_0').\hat{\omega}_i(m') = \varphi_{v,i}(m') + D(m - t_0).T.\omega_i(m) \tag{2.7}$$

où *T* est le facteur de transposition.

L'inconvénient principal de cette méthode est qu'elle impose une recherche des instants de fermeture de la glotte, instants pour lesquels les harmoniques sont synchronisées.

2 L'algorithme "shape invariant" de Federico [3]

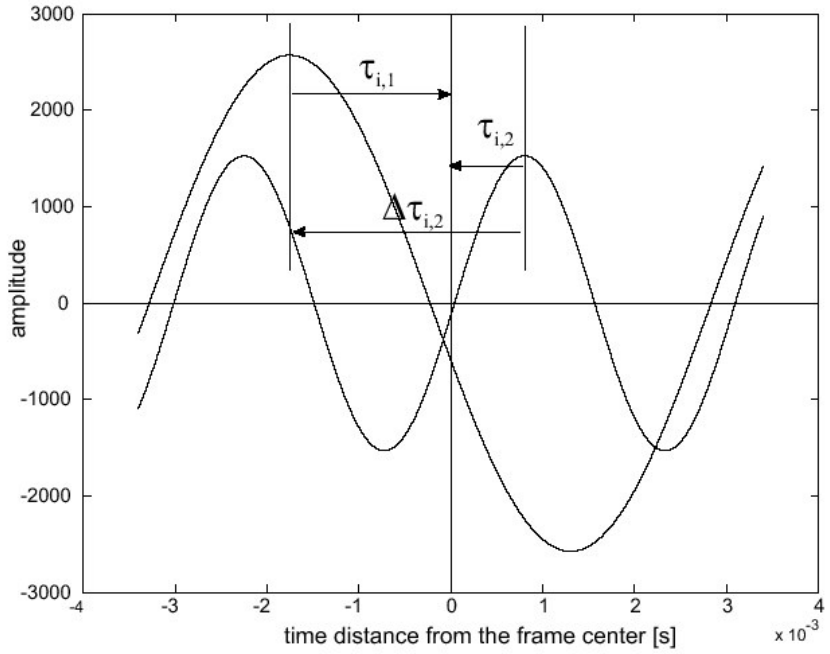
2.1 Principes

L'idée est de remplacer la recherche des instants de synchronisation (instants de fermeture de la glotte) par une synchronisation sur le maximum du fondamental. Ainsi, on remplace le jeu de paramètres additifs $A_i(n)$, $f_i(n)$ et $\varphi_i(n)$ par un nouveau jeu de 3 paramètres : $A_i(n)$, $f_i(n)$ et $\Delta\tau_i(n)$ où $\Delta\tau_i(n)$ est le retard de phase relatif du partiel ***i*** vis-à-vis du premier partiel (le fondamental).

On calcule d'abord le retard de phase de chaque partiel. Il est donné par :

$$\tau_i(n) = \frac{\varphi_i(n)}{\omega_i(n)} \tag{2.8}$$

Ce retard de phase représente en fait le temps qui sépare l'instant d'analyse du maximum de l'harmonique le plus proche, comme l'illustre la figure ci-dessous.



[Fig. 12] Illustration du retard de phase relatif, d'après [3]

Le retard de phase relatif (non normalisé) est simplement donné par :

$$\Delta\tilde{\tau}_i(n) = \tau_i(n) - \tau_1(n) \tag{2.9}$$

Reste à le normaliser de manière à ce qu'il soit compris dans l'intervalle $\left[0; \frac{2\pi}{\omega_i(n)}\right]$ (afin qu'il ne décrive pas un retard de plus d'une période) :

$$\Delta\tau_i(n) = \mathrm{mod}\left[\Delta\tilde{\tau}_i(n), \frac{2\pi}{\omega_i(n)}\right] \tag{2.10}$$

La synthèse du premier trajet reste identique à la synthèse additive classique sans « shape invariant ». On note $\varphi'_1(n)$ la phase de synthèse du premier trajet à l'instant ***n***.

Les phases des autres partiels sont données par la formule suivante :

$$\varphi'_i(n) = \mathrm{mod}\left[\left(\frac{\varphi'_1(n)}{\omega_1(n)} + \Delta\tau_i(n)\right).\omega_i(n)\ ,\ 2.\pi\right] \tag{2.11}$$

On remarque bien entendu que la formule (2.11) nous redonne bien $\varphi'_i(n) = \mathrm{mod}\left[\varphi_i(n)\ ,\ 2.\pi\right]$ si la phase du 1er partiel est inchangée à la synthèse.

Les avantages de cette méthode sont :

- On n'a pas besoin de marqueurs i.e. pas besoin de recherche d'instant de synchronisation : elle est implicitement faite dans le calcul du retard de phase.
- Quelle que soit la transformation opérée sur le premier partiel, l'application de la formule (2.11) suffit à garantir la conservation de la forme d'onde

2.2 Time-strech et pitch-shift

La transformation est appliquée au 1er partiel sans précaution particulière (cf. partie sur la synthèse additive). $\varphi'_1(m)$ est la phase du 1er trajet après transformation (transposition d'un facteur ***T*** et/ou dilatation d'un facteur ***D***).

La formule de calcul des autres phases devient :

$$\varphi'_i(n) = \mathrm{mod}\left[\left(\frac{\varphi'_1(n)}{T.\omega_1(n)} + \Delta\tau_i(n)\right).T.\omega_i(n)\ ,\ 2.\pi\right] \tag{2.12}$$

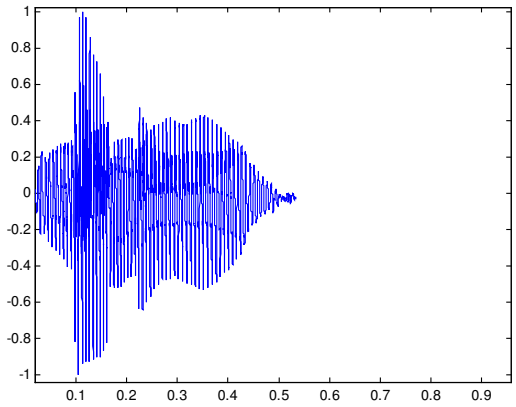
Puis on calcule la phase de synthèse et l'amplitude dans la nouvelle trame synthétisée à l'aide de l'interpolation cubique décrite dans la partie synthèse additive avec un axe temporel décrivant cette fois l'intervalle [0 ; D.(n'-n)].

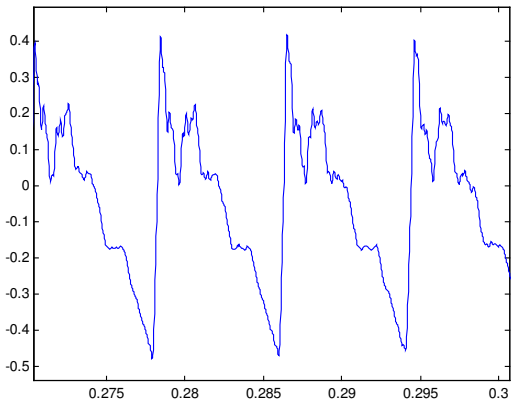
Toutefois, on verra dans la partie « modification de l'algorithme » en quoi cette formulation simple délivre un jeu de paramètres incohérent pour les signaux de parole.

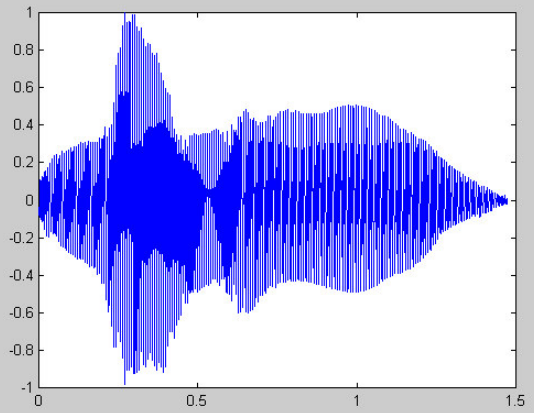
2.3 Résultats obtenus

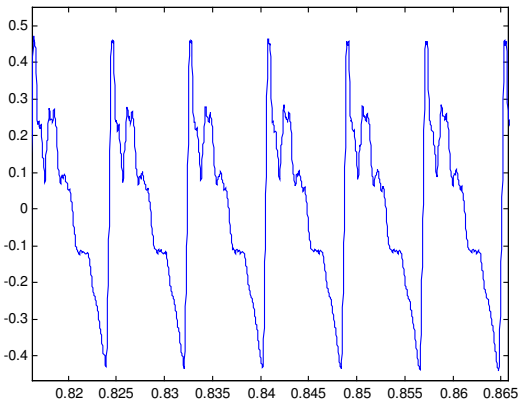
Les figures ci-dessous montrent « visuellement » l'effet du shape invariant. Le signal de départ (figures du haut) est un signal de parole, voisé en grande partie, échantillonné à 44100Hz. L'axe des abscisses est en secondes. Les figures de gauche montrent le signal dans sa globalité et les figures de droites un zoom sur une zone du signal.

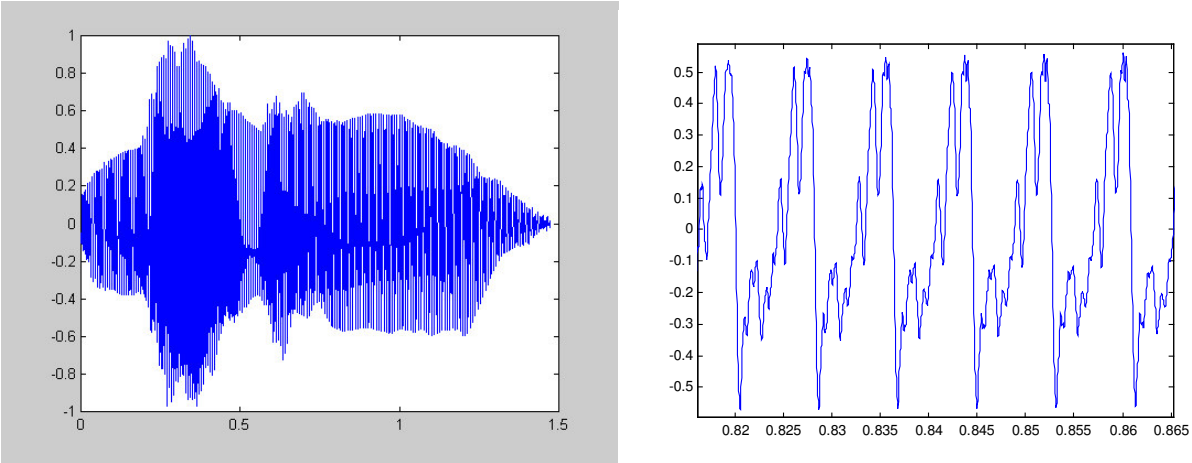
Les 2 premières figures sont le signal d'origine, les 2 suivantes le signal avec un time-stretch de 3 et l'utilisation de l'algorithme shape invariant de Federico. Enfin, les 2 dernières sont dans les mêmes conditions mais sans utiliser d'algorithme « shape invariant ». On voit très nettement l'effet de l'algorithme sur la forme d'onde…











[Fig. 13] Illustration temporelle du « shape invariant » sur un signal de parole

Perceptivement, la différence entre un signal traité avec et sans algorithme shape invariant est flagrante. On regagne sur les parties voisées un réalisme que l'on avait perdu lors de la transformation.

Cependant, pour les parties voisées, les résultats ont été en-dessous de nos attentes, laissant toujours au signal une partie réverbérante et/ou métalisante pas toujours très naturelle pour des transformations importantes (time stretch de 5 par exemple). Deux sources d'erreur ont été identifiées : l'incohérence possible des paramètres issus des algorithmes « shape invariant » de Federico ; la nécessité de prendre en compte une fréquence de coupure de voisement et de ne pas appliquer les algorithmes shape invariant sur tout le spectre.

C'est pourquoi, dans un premier temps, un long travail d'investigation a été fait pour déceler dans le modèle de Federico quelles incohérences il pouvait y avoir et comment on pouvait l'améliorer. Cela nous a permis de proposer la modification ci-après à l'algorithme de Federico qui a beaucoup plus de sens théorique pour une transformation sur des signaux de parole. Cependant, malgré ces efforts, et si cette modification s'avère certainement plus satisfaisante pour la cohérence des trajectoires, elle ne nous aura pas permis d'améliorer de manière sensible la qualité audio du signal resynthétisé.

Finalement, c'est donc vers une application sélective des algorithmes « shape invariant » en fonction de la fréquence et du degré de voisement que l'on se tourne pour espérer améliorer la qualité audio, qualité qui pourra encore être améliorée par l'utilisation de modèles de bruit.

Enfin, pour les parties non voisées, on se tournera aussi vers des modèles de bruit et on n'appliquera pas le modèle « shape invariant » qui tend à rendre périodique un son qui ne l'était pas (en réduisant l'inharmonicité, cf partie suivante).

2.4 Modifications de l'algorithme

Problématique

La contribution des changements de position du conduit vocal au signal de parole voisé, qui se traduit par un filtrage variant dans le temps, introduit des changements de phase et donc de fréquence. Il se crée une inharmonicité qui se traduit par une fréquence apparente pour le k^{ème} partiel de

$$f_k = k.f_0 + \Delta f \tag{2.13}$$

où Δf traduit l'inharmonicité de la k^{ème} harmonique.

Lors d'une dilatation temporelle d'un facteur ***D***, l'inharmonicité doit aussi être réduite du même facteur. En effet, le filtre évolue plus lentement, les changements de phase aussi, l'inharmonicité est donc logiquement réduite du facteur de dilatation.

On veut garder au nouvel instant de synthèse, un déphasage dû au conduit vocal identique à celui de l'analyse.

$$\varphi(T) = 2\pi \int_{t=0}^{T} f_k dt = 2\pi \int_{t=0}^{T} (k.f_0 + \Delta f)dt = 2\pi k.f_0 T + \Delta\varphi \text{ avec } \Delta\varphi = 2\pi \int_{t=0}^{T} \Delta f \tag{2.14}$$

Hors si l'on veut garder le même $\Delta\varphi$ à la synthèse après une dilatation temporelle, on doit resynthétiser en fait :

$$\varphi(DT) = 2\pi k f_0 DT + \Delta\varphi = 2\pi k f_0 DT + 2\pi \int_{t=0}^{T} \Delta f dt \simeq 2\pi \int_{t=0}^{DT} (k.f_0 + \frac{\Delta f}{D})t \tag{2.15}$$

Il faut donc à la synthèse réduire l'inharmonicité (i.e. l'écart entre la fréquence analysée et la fréquence théorique de l'harmonique) et en tenir compte dans l'algorithme de Federico.

De même, lors d'une transposition d'un facteur P, il faut aussi réduire l'inhamonicité de la même manière :

$$\varphi(T) = 2\pi k P f_0 T + \Delta\varphi = 2\pi k P f_0 T + 2\pi \int_{t=0}^{T} \Delta f dt = 2\pi \int_{t=0}^{T} (k.f_0 + \frac{\Delta f}{P})Pt \tag{2.16}$$

Algorithme proposé

<u>NB</u> : Nous ne présentons pas ici toutes les investigations faites autour de l'algorithme shape invariant mais uniquement l'algorithme que l'on a finalement retenu.

La première étape consiste à modifier les fréquences issues de l'analyse pour réduire l'inharmonicité d'un facteur *D*. Il faut donc avoir une estimation robuste de la fréquence

fondamentale. On ne peut pas se baser sur le premier partiel puisque d'une part, lui aussi subit une inharmonicité, et d'autre part, sa précision peut ne pas être suffisante si l'amplitude du fondamental est faible. Pour garder une cohérence entre les paramètres de l'analyse additive et une estimation de f_0, on choisit d'utiliser une sorte de moyenne pondérée des partiels, plutôt que de se baser sur un estimateur indépendant.

Ainsi, on estime une fondamentale moyenne de :

$$\overline{f_0} = \frac{\sum_{i=1}^{M} A_i . \frac{F_i}{i}}{\sum_{i=1}^{M} A_i} \tag{2.17}$$

avec les hypothèses suivantes :

- $i = \text{round}(F_i / f_0)$ où f_0 est la fondamentale estimée de manière indépendante à l'analyse-synthèse additive (la méthode utilisée ici est celle de YIN publiée par A. Cheveigné, et H. Kawahara [15])
- De même, si le $i^{ème}$ partiel est absent, il n'est bien sûr pas pris en compte dans le calcul de $\overline{f_0}$

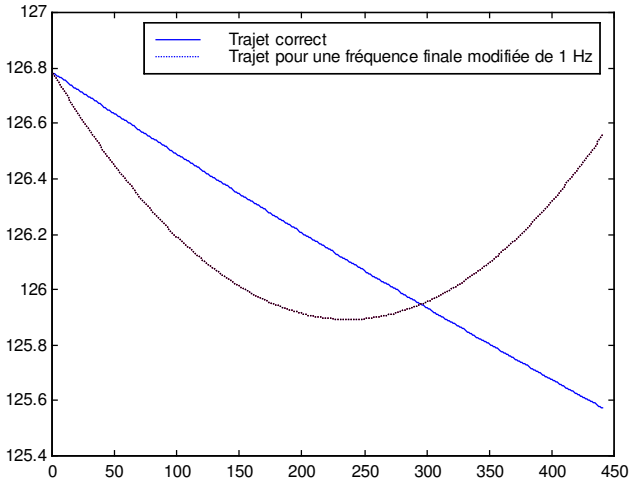
La deuxième étape consiste à modifier toutes les fréquences (y compris celle du premier trajet relativement auquel se calculeront les nouvelles phases) issues de l'analyse de la manière suivante (sous les mêmes hypothèses pour le calcul de i) afin de réduire l'inharmonicité du facteur de dilatation :

$$F_i' = i.\overline{f_0} + \frac{1}{D}\left[F_i - i.\overline{f_0}\right] \tag{2.18}$$

A ce stade, on dispose d'un jeu de paramètres qui ne concordent plus. Le premier trajet fréquentiel sur lequel s'appuie le reste de l'algorithme « shape invariant » est décrit d'une part par les phases (et amplitudes) issues de l'analyse, d'autre part, par des nouvelles valeurs de fréquences pour lesquelles l'inharmonicité a été réduite (si transformation il y a).

Hors, l'algorithme utilisé pour calculer le trajet de phase, ne peut pas correctement fonctionner si on modifie indépendamment phase ou fréquence sans répercuter ce changement sur l'autre des 2 valeurs. N'oublions pas que la phase est, à un facteur près, l'intégrale de la fréquence.

Ainsi, puisque l'algorithme de calcul de phase est basé sur les conditions initiales et finales de phase et de fréquence issues de l'analyse, le désaccord des paramètres sera compensé par une modulation de fréquence. Cela est illustré par la figure ci-dessous où l'on montre en trait continu le trajet en fréquence issu du calcul de synthèse additive avec les paramètres corrects et le trajet fréquentiel calculé avec une fréquence finale modifiée de 1Hz en traits pointillés. La modulation de fréquence apparaît nettement.

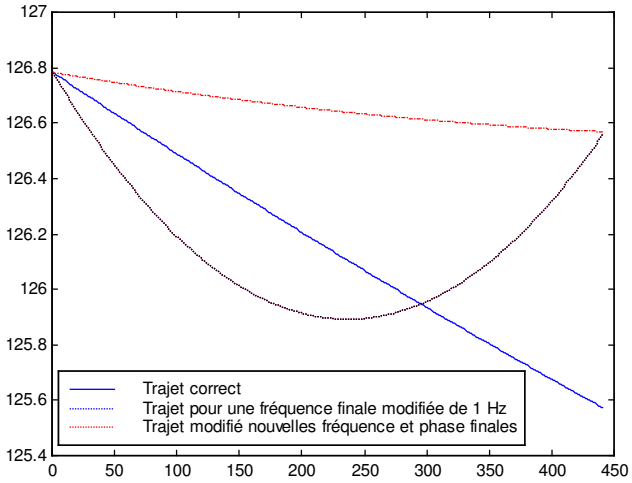


[Fig. 14] Problématique des modulations de fréquence pour un changement d'inharmonicité

Afin d'éviter ces apparitions de modulation de fréquence (même s'il n'est pas prouvé qu'elles peuvent avoir un impact perceptif sur la qualité), on modifie les valeurs de phases associées au trajet du 1er partiel pour qu'elles prennent en compte le changement de fréquence. On utilise la formule suivante :

$$\varphi'_{1,p+1} = \varphi'_{1,p} + \left(\varphi_{1,p+1} - \varphi_{1,p}\right) + 2\pi.T\left[\frac{F_1'(n+1) + F_1'(n)}{2} - \frac{F_1(n+1) + F_1(n)}{2}\right] \quad (2.19)$$

La figure suivante illustre la correction qui revient en fait à avoir le même trajet de fréquence que l'original (avec le même niveau de modulation s'il y en avait un) mais pour rejoindre cette fois-ci les fréquences modifiées (traits pointillés larges).



[Fig. 15] Correction proposée pour supprimer les modulations de fréquence dues au changement d'inharmonicité du 1er trajet

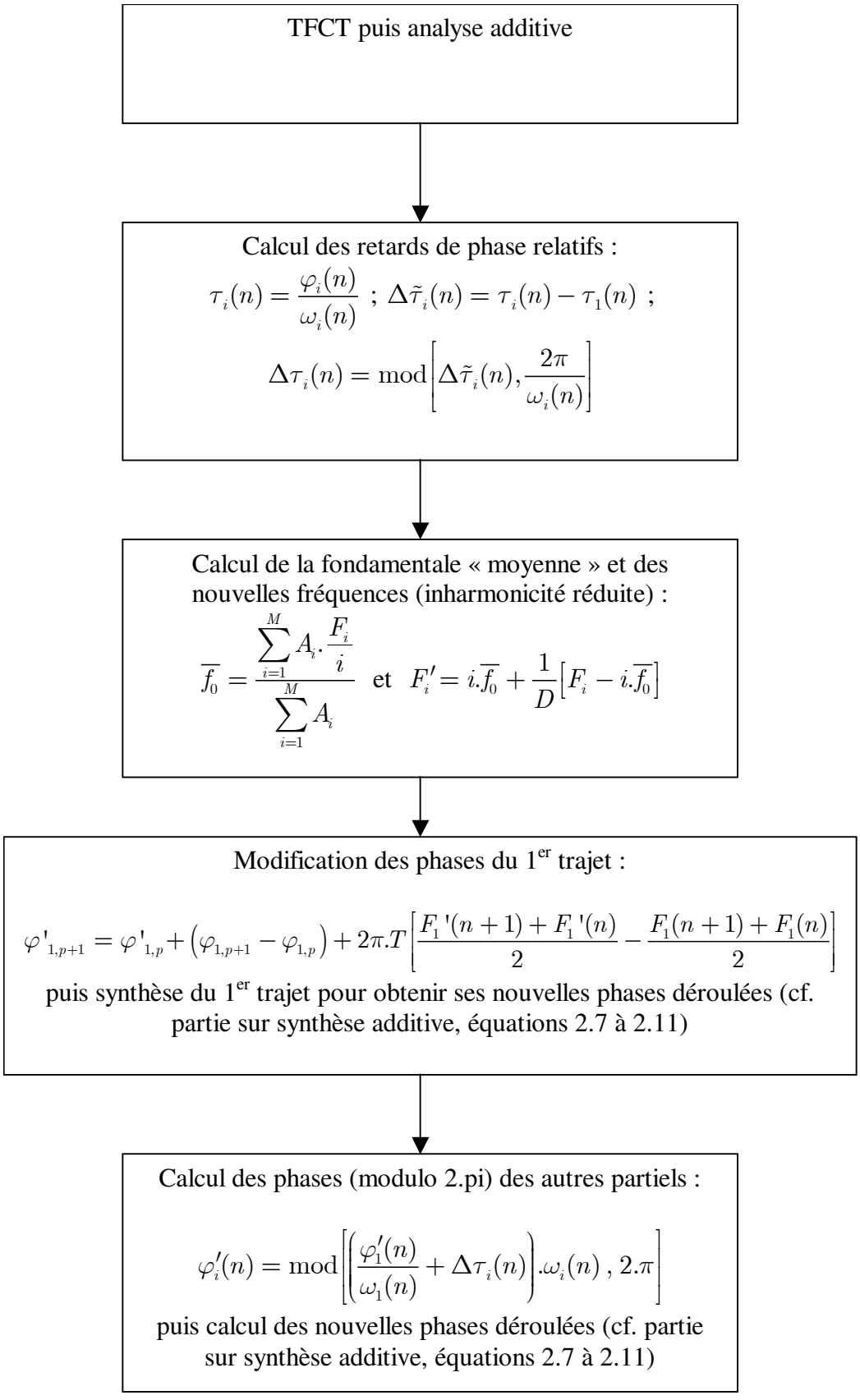
Ainsi, on est sûr de générer un premier trajet cohérent par rapport à l'analyse, tout en tenant compte de la correction de l'inharmonicité.

Puisque, ensuite, pour les autres partiels, la phase est imposée relativement à la phase du 1er trajet, on peut appliquer les formules de Federico sans précaution particulière pour les nouvelles fréquences avec l'inharmonicité modifiée.

Enfin, une fois toutes ces phases modulo[2.pi] et fréquences calculées, il ne reste plus qu'à calculer les trajets de phase (équations (1.7) à (1.11)) et les trajets des partiels comme décrits dans la partie sur la synthèse additive.

Schéma synoptique de l'algorithme proposé

L'algorithme complet (dans le cas d'une dilatation d'un facteur *D*) est résumé dans le schéma ci-après.



[Fig. 16] Synoptique résumant l'algorithme de « shape invariant » utilisé

Synthèse additive du bruit

La synthèse additive, et ce n'est pas un hasard si elle est utilisée dans des applications de débruitage, ne permet pas de prendre en compte correctement les informations de bruit comprises dans le signal d'origine. Si la prise en compte du bruit résiduel dans la synthèse pallie en partie ce problème, maintenir son réalisme devient alors délicat lorsqu'on applique des transformations. Kelly Fitz a proposé un modèle additif bruité qui permet de combler ce manque et c'est cet algorithme qui a été utilisé notamment pour les parties non voisées des signaux de parole.

1 Prise en compte du bruit par le résiduel [8][10]

Le résiduel (portant en grande partie l'information de bruit) est en fait la différence entre le signal d'origine et le signal synthétisé sans transformation. Il est donné par (***u*** étant le signal d'origine):

$$r(n) = u(n) - s(n) = u(n) - \sum_{i=1}^{M} A_i(n)\cos(\Phi(n)) \tag{3.1}$$

Ajouter le résiduel au signal synthétisé (après transformation) n'est bien sûr pas cohérent. Aussi, la plupart des méthodes basées sur le résiduel proposent de calculer la fonction de transfert d'un filtre ***h(n)*** (estimé trame d'analyse par trame d'analyse) qui, ayant pour entrée un bruit blanc ***b(n)***, génère en sortie un signal proche du résiduel.

$$b(n) * h(n) \simeq r(n) \tag{3.2}$$

Si cette méthode paraît convenir pour des compressions/dilatations temporelles, elle n'est en revanche pas forcément adaptée pour des transpositions, puisque le résiduel est toujours considéré constant. Ce qui par exemple poserait problème dans le cas d'un son de parole à la fois voisé et non voisé, le déplacement des harmoniques par transposition ferait alors apparaître des trous dans le spectre.

C'est pourquoi, nous avons choisi d'utiliser un algorithme plus robuste proposé par Fitz et décrit ci-après.

2 Le modèle additif amélioré de Kelly Fitz [12] [13] [14]

2.1 Principes

Il doit s'appliquer à tous types de signaux. L'idée est d'associer à chaque oscillateur sinusoïdal du modèle additif, un générateur de bruit filtré.

Ainsi, chaque oscillateur s'écrit désormais :

$$y_{i,k}(n) = G_{i,k}\left[\sqrt{1-K_{i,k}} + \sqrt{2K_{i,k}}.\left(b_{i,k} * h_{i,k}\right)[n]\right].\cos(\Phi(n)) \tag{3.3}$$

où i est l'indice de trame, k le rang du partiel, n l'indice temporel discret, $b_{i,k}$ un bruit blanc (gaussien en l'occurrence), $h_{i,k}$ un filtre, et $2\pi\omega_{i,k}$ la fréquence du $k^{ème}$ partiel. $G_{i,k}$ permet de régler le gain de l'oscillateur. Enfin, $K_{i,k}$ est compris entre 0 et 1 et permet de régler le niveau de bruit injecté dans l'oscillateur ($K_{i,k} = 0$ pour un oscillateur pur non bruité, $K_{i,k} = 1$ pour un bruit pur).

La phase et la fréquence du modèle sinusoïdal restent inchangés, tandis que $G_{i,k}$ et $K_{i,k}$ doivent vérifier :

$$G_{i,k}\sqrt{1-K_{i,k}} = A_{i,k} \tag{3.4}$$

où $A_{i,k}$ est l'amplitude du pic sinusoïdal considéré.

On a donc désormais pour chaque oscillateur et à chaque instant d'analyse 3 nouveaux paramètres dont un filtre à évaluer.

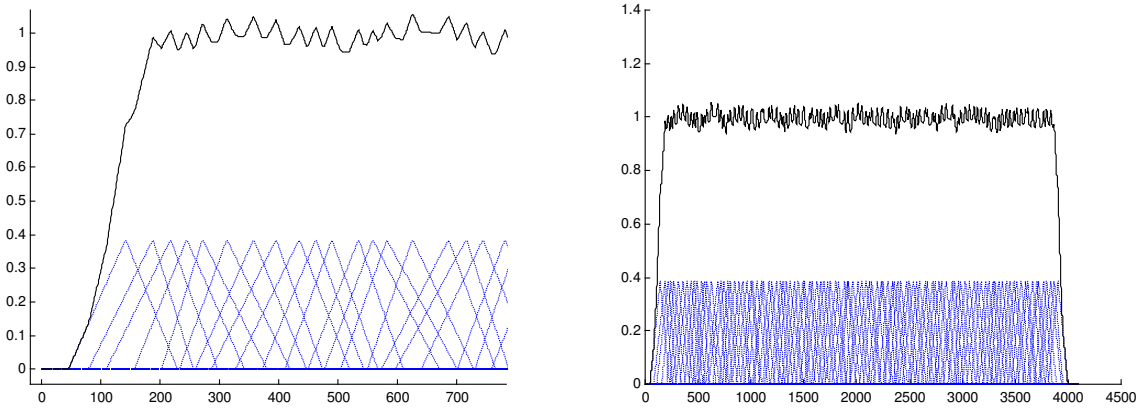
L'évaluation des paramètres se fait en calculant la puissance de bruit dans une fenêtre spectrale comprenant l'harmonique en question et en la rendant égale à celle de y. D'autre part, Kelly Fitz suggère aussi d'utiliser des fenêtres larges et recouvrantes pour de meilleurs résultats perceptifs et une plus grande robustesse aux transformations.

La méthode utilisée par Fitz pour évaluer les paramètres n'est pas décrite dans la littérature.

2.2 Algorithme proposé

Afin de simplifier l'algorithme, on prend pour hypothèse un modèle harmonique, ce qui nous permettra d'avoir des fenêtres d'évaluation des puissances quasi-identiques et relativement régulièrement bien espacées. En effet, chaque fenêtre a une même largeur et est centrée sur le partiel dont on cherche à estimer le niveau de bruit. De plus, on prend pour filtre de chaque oscillateur, un filtre dont la réponse fréquentielle approxime la fenêtre choisie.

On peut voir sur les figures suivantes en traits pointillés, les fenêtre élémentaires dans lesquelles le bruit sera mesuré puis généré et en trait plein la somme de ses fenêtres qui oscille autour de 1. L'axe des abscisses s'étend de 0 à 4096 pour 0 à 22050Hz.



[Fig. 17] Fenêtres de mesure et de génération du bruit

Une fois la fenêtre posée, on évalue la puissance du bruit dans le fenêtre à l'aide de la TFCT du signal d'entrée.

On rappelle les relations suivantes pour le calcul d'énergie E (que l'on doit à Parseval) :

$$E = \sum_{i=1}^{N} \|x.g\|^2 = \frac{1}{n_{fft}} \sum_{i=1}^{n_{fft}} \|\mathrm{fft}(x.g)\|^2 \tag{3.5}$$

De plus, dans le cas d'un oscillateur sinusoïdal, on a :

$$x = A\cos(2\pi f \frac{n}{f_s}) \Rightarrow E = \frac{A^2}{2} \sum_{i=1}^{N} \|g\|^2 \tag{3.6}$$

où g est la fenêtre temporelle de longueur N utilisée pour la FFT, et telle que $n_{fft} \geq N$.

Appliquons ces formules à notre TFCT S, multipliée par une fenêtre spectrale W de longueur L, (la fenêtre étant appliquée en puissance).

$$E = E_{bruit} + E_{sinus}$$
$$E = \sum_{i=1}^{nbfft} \frac{\|S\|^2 .W}{nbfft} \tag{3.7}$$

D'autre part,

$$E_{sinus} = \sum_{k} \frac{A_{i,k}{}^2 W_k}{2} \sum_{i=1}^{N} \|g\|^2 \tag{3.8}$$

où $W_{k'}$ est l'amplitude de la fenêtre coïncidant aux pics $A_{i,k}$.

On en déduit :

$$E_{bruit} = \sum_{i=1}^{n_{fft}} \frac{\|S\|^2 .W}{n_{fft}} - \sum_{k} \frac{A_{i,k}{}^2 W_k}{2} \sum_{i=1}^{N} \|g\|^2 \tag{3.9}$$

Il faut désormais calculer les paramètres que $G_{i,k}$ et $K_{i,k}$ pour que l'énergie du bruit de notre nouvel oscillateur égale E_{bruit}. L'énergie de bruit de notre oscillateur est donnée ci-après :

$$\begin{aligned} & y_{i,k}(n) = G_{i,k}\left[\sqrt{1-K_{i,k}} + \sqrt{2K_{i,k}}.\left(b_{i,k} * h_{i,k}\right)[n]\right].\cos(\Phi(n)) \\ & \Rightarrow E_{bruit} = E\left[G_{i,k}.\sqrt{2K_{i,k}}.\left(b_{i,k} * h_{i,k}\right)[n].\cos(\Phi(n))\right] \\ & \Rightarrow E_{bruit} = \left(G_{i,k}\cdot\sqrt{2K_{i,k}}\right)^2 \cdot \sum_n \left[h_{i,k}(n)\right]^2 \cdot \frac{1}{2} = {G_{i,k}}^2 \cdot K_{i,k} \sum_n \left[h_{i,k}(n)\right]^2 \end{aligned} \tag{3.10}$$

Finalement (d'après ce qui précède et l'équation (3.4)), l'oscillateur bruité doit être solution du système d'équation suivant :

$$\begin{cases} G_{i,k}\sqrt{1-K_{i,k}} = A_{i,k} \\ E_{bruit} = \sum_{i=1}^{n_{fft}} \frac{\|S\|^2.W}{n_{fft}} - \sum_k \frac{{A_{i,k}}^2 W_k}{2} \sum_{i=1}^{N} \|g\|^2 = {G_{i,k}}^2 \cdot K_{i,k} \sum_n \left[h_{i,k}(n)\right]^2 \end{cases} \tag{3.11}$$

qui admet comme solutions :

$$\begin{cases} K_{i,k} = \frac{\left({G_{i,k}}^2 - {A_{i,k}}^2\right)}{{G_{i,k}}^2} \\ E_{bruit} = \left({G_{i,k}}^2 - {A_{i,k}}^2\right) \sum_n \left[h_{i,k}(n)\right]^2 \end{cases} \Leftrightarrow \begin{cases} G_{i,k} = \sqrt{\frac{E_{bruit}}{\sum_n \left[h_{i,k}(n)\right]^2} + {A_{i,k}}^2} \\ K_{i,k} = 1 - \frac{{A_{i,k}}^2}{{G_{i,k}}^2} \end{cases} \tag{3.12}$$

2.3 Mise en oeuvre pratique

On a pu se rendre compte qu'il fallait utiliser des filtres assez large pour qu'il y ait un recouvrement important mais d'une largeur de bande raisonnable pour pouvoir « modeler » la forme du bruit générer (avec une fenêtre de 2000Hz, par exemple, on n'est plus capable de modeler un bruit finement, on aura un bruit beaucoup plus uniforme).

On a finalement utilisé un filtre de 400Hz de large, ce qui, suivant les trames, donne 2 à 3 partiels par fenêtre. D'autre part, afin de synthétiser le bruit sur l'ensemble du spectre, on ajoute si besoin des partiels « virtuels » $F_{i,bruit}$, décrivant uniquement le bruit et couvrant la bande non couverte par les M partiels issues de l'analyse additive : $F_{i,bruit} = \left\{ i \cdot f_0 \;\middle|\; i > M \text{ et } i \cdot f_0 \leq \frac{f_s}{2} \right\}$.

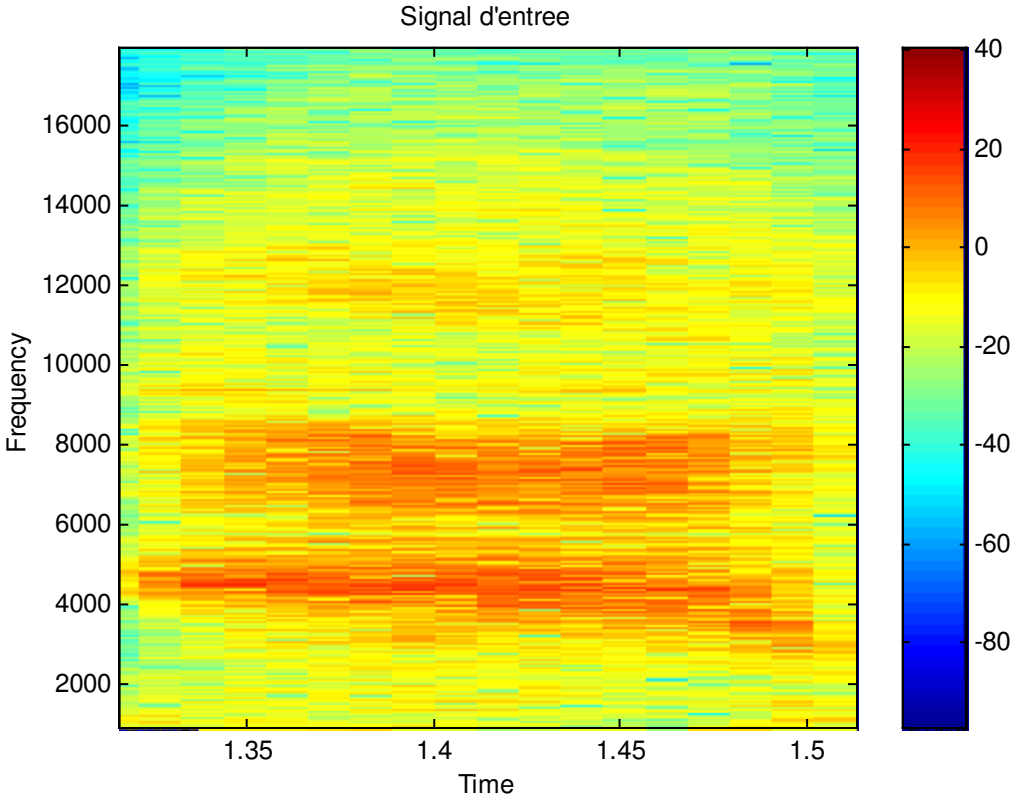
2.4 Résultats

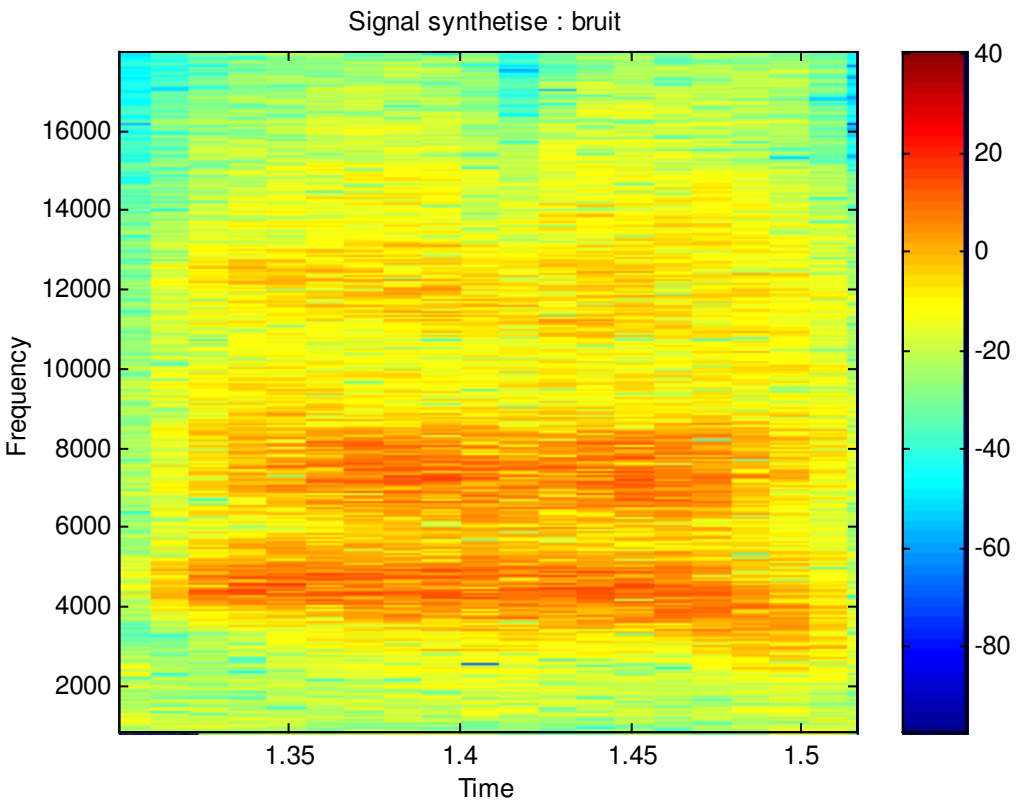
D'un point de vue perceptif, on arrive tout-à-fait correctement à générer les parties non voisées comme les sifflantes par exemple. Cependant, une erreur d'estimation de l'amplitude des pics extraits peut transformer une partie de l'énergie sinusoïdale en énergie de bruit (supplémentaire) et qui suit les partiels. Cela crée un son très peu naturel et il peut s'avérer utile de majorer légèrement les amplitudes des pics extraits avant l'analyse de bruit pour éviter ce problème.

Cependant, mieux vaut avoir un modèle prenant en compte la fréquence de coupure de voisement et moduler le bruit généré en fonction de celle-ci (cf. partie résultats).

Les figures suivantes montrent le spectrogramme d'une sifflante (le sons « s ») du signal original puis du signal de bruit synthétisé.

Par contre, on verra dans le chapitre suivant, que pour les sons nécessitant une partie voisée et une partie bruitée, l'intégration des 2 modèles reste délicate pour maintenir un certain réalisme.





[Fig. 18] Spectrogrammes d'une sifflante sur le signal original en haut et sur le signal synthétisé par le modèle de bruit en bas

Résultats

Une fois le « shape invariant » et les modèles de bruit implantés, l'étape suivante consiste à fusionner ces 2 parties afin de converger vers un modèle réaliste de synthèse de la voix. L'idée sous-jacente est de prendre en compte une estimation du degré de voisement du signal de parole pour doser la proportion de bruit et de signal sinusoïdal à utiliser pour la synthèse.

1 Rôle de l'estimation de voisement

Nous avons déjà vu que les parties voisées et les parties non voisées ne devaient pas subir le même traitement. En fait, cette séparation n'est pas binaire et on peut considérer que le signal est voisé jusqu'à une fréquence de coupure de voisement notée f_v et non voisé au-delà.

Plusieurs méthodes permettent d'estimer le degré de voisement et/ou de calculer f_v [5], basées sur des critères de sinusoïdalité (évaluées sur le spectre) ou plus simplement sur un critère d'harmonicité à respecter dans les zones dites voisées.

Ces méthodes ne font pas partie du travail de ce projet et l'on considère par la suite que l'on dispose d'une estimation de f_v et du degré de voisement (coefficient entre 0 et 1 ; 0 pour un signal totalement non voisé, 1 pour un signal totalement voisé).

Ainsi, à chaque trame d'analyse, on peut considérer 2 zones spectrales :

- En dessous de f_v : le signal est voisé. Dans ce cas, la génération du bruit est inhibée, et les algorithmes « shape invariant » sont utilisés (puisque les fréquences analysées sont des harmoniques du signal).
- Au-dessus de f_v : le signal devient non voisé. Dans ce cas, les algorithmes shape invariant ne sont plus utilisés et les fréquences estimées ne sont plus des harmoniques du signal : on ne les connecte plus ensemble d'une trame à l'autre. On introduit tout ou partie du bruit estimé.

Bien entendu, toute la difficulté réside dans la transition entre les 2 parties. Il faut, au-delà de f_v, progressivement réduire le niveau des sinus évalués, et augmenter la part de bruit, progressivement aussi.

2 Résultats obtenus

Plusieurs fonctions de transition ont été essayées et si l'amélioration de la qualité audio avec la prise en compte du voisement est tout-à-fait sensible, l'objectif de ce travail dans le cadre limité d'un projet n'est bien sûr pas d'arriver à une fonction optimisée donnant une qualité haute fidélité, utilisable directement sans modification aucune.

Les fonctions qui ont été utilisées pour modifier les paramètres issus de l'analyse sont les suivantes (où $A_{i,p}$ et $K_{i,p}$ sont respectivement les amplitudes sinusoïdales extraites lors de l'analyse additive et les coefficients de bruit associés).

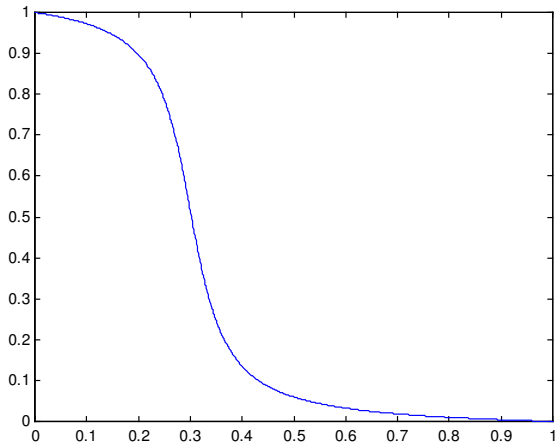
$$A_{i,p}' = \begin{cases} A_{i,p} \text{ si } f_{i,p} \le f_v \\ A_{i,p} \cdot \alpha \cdot \max\left(0 ; \dfrac{F_{max} - f_{i,p}}{F_{max} - f_v}\right) \text{ si } f_{i,p} > f_v \end{cases} \qquad (4.1)$$

avec $0 \le \alpha \le 1$, $F_{max} = v \cdot f_v \le \dfrac{f_s}{2}$.

Les valeurs α=0.5 et v = 15 donnent de bons résultats.

$$K_{i,p}' = \begin{cases} 0 \text{ si } f_{i,p} \le f_v \\ K_{i,p} \cdot \alpha' \cdot \min\left(1 ; \dfrac{f_{i,p} - f_v}{F'_{max} - f_v}\right) \text{ si } f_{i,p} > f_v \end{cases} \quad \text{puis} \quad K_{i,p}' = f\left(K_{i,p}'\right) \qquad (4.2)$$

avec $0 \le \alpha' \le 1$, $F'_{max} = v' \cdot f_v \le \dfrac{f_s}{2}$ et f est une fonction décroissante du degré de voisement estimé, entre 0 et 1. La fonction suivante basée (illustrée sur la figure ci-dessous) sur une arctan donne d'assez bons résultats, avec α=1 et $F'_{max} = \dfrac{f_s}{2}$.



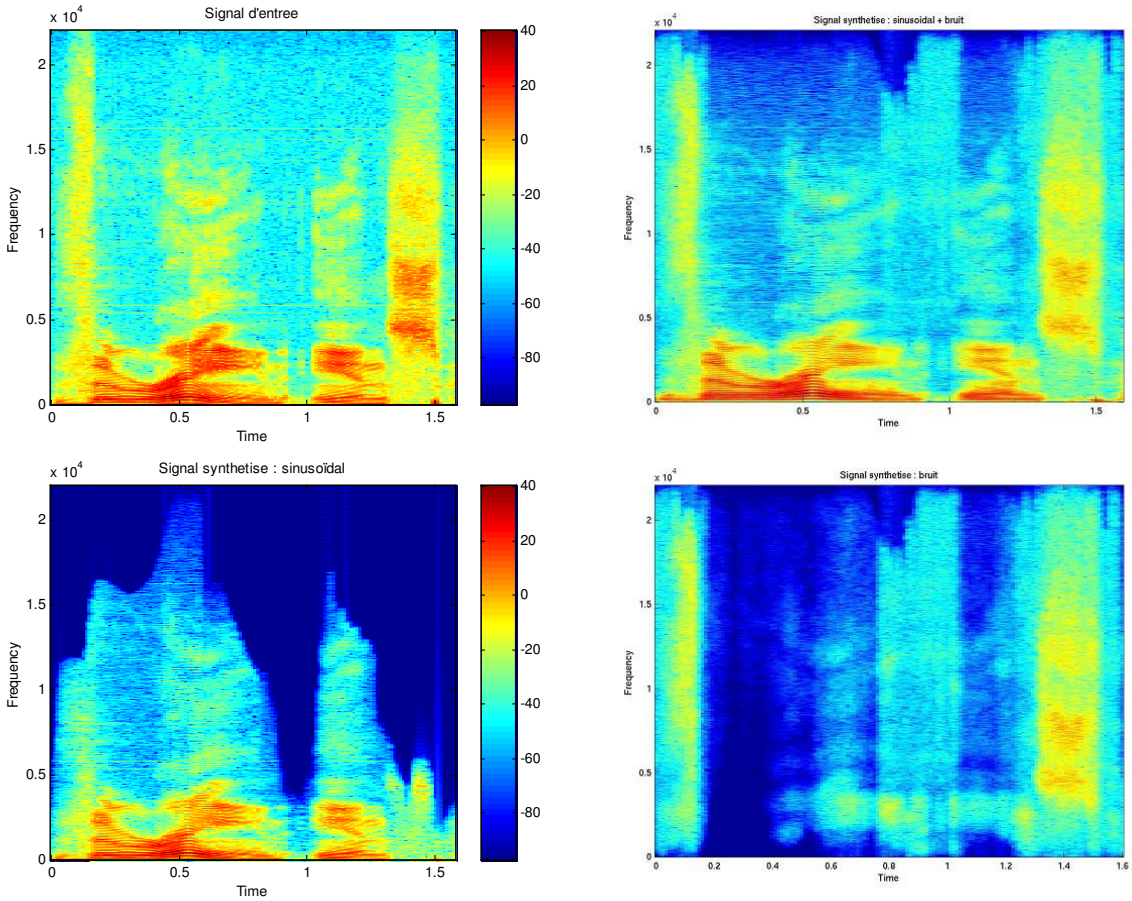
[Fig. 19] Fonction f donnant le gain à appliquer au facteur de bruit en fonction du degré de voisement

NB : *Cette partie, tout-à-fait préliminaire dans la durée limitée du projet, ne constitue en aucun cas un résultat définitif et optimisé. On n'a notamment pas pris en compte de critère de conservation de l'énergie. Le but, ici, n'est que de regarder l'impact sur la qualité audio d'un passage progressif des modèles « shape invariant » aux modèles de bruit, en fonction du voisement.*

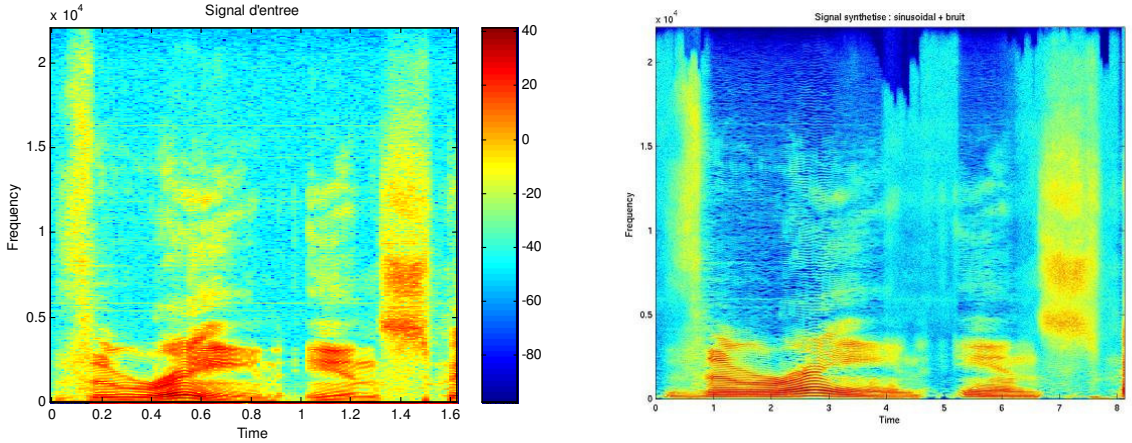
D'un point de vue perceptif, on obtient une qualité satisfaisante mais plusieurs points restent à améliorer :

- Malgré la prise en compte du voisement, il persiste un aspect métalique au son pour les parties traitées par le « shape invariant »
- Le règlage du modèle de bruit est sensible et il probable que les optimisations diffèrent suivant les fichiers à traiter. Le modèle reste certainement à affiner si l'on désire avoir une fusion plus importante entre les parties bruitées et sinusoïdales. On peut notamment songer à un algorithme adaptatif permettant d'optimiser ces paramètres en vue d'un critère de minimisation du résiduel.

Les figures suivantes illustrent les résultats obtenus sous forme de spectrogrammes sans transformation d'abord puis avec un facteur de dilatation de 5.



[Fig. 20] Spectrogrammes du signal de parole « the rain in S… » : original en haut à gauche, synthétisé en haut à droite, partie sinusoïdale seulement en bas à gauche, partie bruit seulement en bas à droite (sans transformation)



[Fig. 21] Spectrogrammes du signal de parole « the rain in S… » : original à gauche, synthétisé à droite avec une dilatation temporelle d'un facteur 5

Conclusion

La synthèse additive de la parole reste délicate si l'on veut maintenir un réalisme important. Si le modèle proposé demeure encourageant en terme de performance, la fusion entre le modèle sinusoïdal et le modèle de bruit reste difficile à ajuster.

Finalement, nous avons montré que la synthèse des parties voisées était considérablement améliorée par les techniques « shape invariant » mais qu'il persiste certains défauts de type son réverbérant ou métalisé

Pour les parties non voisées, le modèle de bruit proposé permet une synthèse correcte, même pour des transformations importantes puisque le bruit synthétisé est « attaché » aux partiels du signal et donc apte à suivre les mêmes transformations.

La spécificité du signal de parole nous oblige à prendre toujours en compte une partie voisée et une partie non voisée et une fois de plus cela limite les défauts qu'on pouvait avoir auparavant (on atténue notamment les défauts du « shape invariant » en ne l'appliquant plus sur tout le spectre).

Cependant, en vue d'obtenir une très haute fidélité, il serait nécessaire d'ajuster le passage des zones voisées aux zones non voisées par des techniques plus robustes. L'égalisation adaptative en vue de minimiser le résiduel étant sans doute une piste intéressante à creuser.

Bibliographie

[1] Robert J. McAulay, Thomas F. Quatieri, " Speech analysis/synthesis based on a sinusoidal representation ", IEEE transactions on Acoustics, Speech and Signal processing, vol. ASSP-34, No 4, August 1986.

[2] Robert J. McAulay, Thomas F. Quatieri, " Shape invariant time-scale and pitch modification of speech ", IEEE transactions on Signal processing, vol. 40, No 3, March 1992.

[3] Riccardo Di Federico, " Waveform preserving time stretching and pitch shifting fro sinusoidal models of sound ", Centro di Sonologia Computazionale, Universita degli Studi di Padova

[4] Thierry Dutoit, « Traitement Automatique de la Parole, Notes de cours / DEC2 », Faculté Polytechnique de Mons Première édition, 2000.

[5] Marine Campedel Oudot, « Etude du modèle 'sinusoïdes et bruits' pour le traitement des signaux de parole. Estimation robuste de l'enveloppe spectrale. », thèse du 13/11/98.

[6] Geoffroy Peeters, "Modèles et modélisation du signal sonore adaptés à ses caractéristiques locales", thèse de juillet 2001.

[7] Ph. Depalle, T. Hélie, " Extraction of spectral peak parameters using a STFT modeling and no sidelob windows. " IRCAM.

[8] Xavier Serra, " Musical sound modeling with sinusoids plus noise ", C. Roads, S. Pope, A. Picialli, G. De Poli, editors 1997. " Musical Signal Processing ". Sets & Zeitlinger Publishers.

[9] E. Bryan George, " Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model ", IEEE transactions on speech and audio processing, Vol. 5 No. 5, September 1997.

[10] Xavier Rodet, " Musical sound signal analysis/synthesis: sinusoidal+residual and elementary waveform models ", TFTS'97, IEEE Time-frequency and time-scale Workshop 97, Coventry, UK, Août 1997.

[11] Site web : http://www.math.ucdavis.edu/%7Estrohmer/ research/gabor/ gaborintro/node3.html

[12] Kelly Fitz, Lippold Haken, Paul Christensen, " A new algorithm for bandwidth-enhanced additive sound modeling ", CERL sound group, University of Illinois at Urbana-Champaign

[13] Kelly Fitz, Lippold Haken, " Bandwidth enhanced sinusoidal modeling in

Lemur ", ICMC proceedings 1995, p.154-157

[14] Kelly Fitz, " The reassigned bandwidth-enhanced method of additive synthesis ", Thesis, 1999.

[15] A. de Cheveigné, H. Kawahara, " YIN, a fundamental frequency estimator for speech and music ", J. Acoust. Soc. Am., 2002.