ESSAI D'UN VOCODER

0. INTRODUCTION

Le Vocoder décrit ici permet de "faire parler" un son quelconque e(t) (cascade etc.) comme s'il était fournit par la glotte.

- Pour ce faire, nous allons réaliser la **représentation temps-fréquence** du signal de parole original où pour chaque tranche de temps n°m centrée en t = m . T, on calcule le spectre court-terme (dit aussi instantané) X(t, f) du signal.

- On va ensuite estimer la réponse fréquentielle court-terme H(t, f) du conduit vocal au cours du temps.

- La représentation temps-fréquence du son quelconque va être déterminée.

- La spectre court-terme de la voix artificielle est obtenu par le produit de la réponse fréquentielle instantanée du conduit vocal et du spectre court-terme.

- Enfin, on en déduit la voix artificielle.

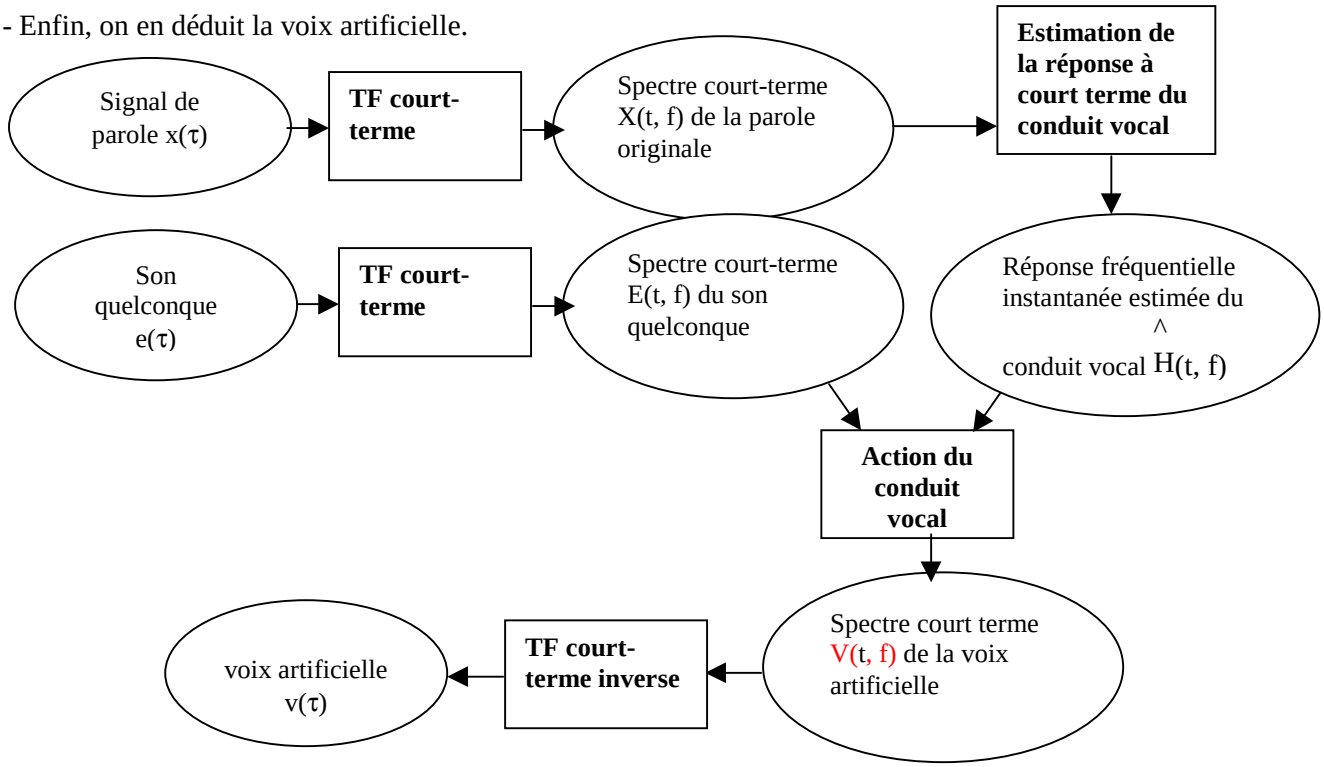

Figure 1: Diagramme de Flot de Données (DFD) du traitement complet
Les traitements sont encadrés par un rectangle, les signaux sont dans des ovales

Notation:

pour un signal x, la valeur à l'instant t est noté : x(t)

l'échantillon n° n de x est noté : x[n]

par exemple, pour une période d'échantillonnage T_e on a : x[n] = x(n . T_e)

Les opérateurs comme TF s'écrivent ici : TF{x(t)} avec des accolades.

Les formules numériques utilisables sont encadrées en double

1. GÉNÉRATION DE LA REPRÉSENTATION TEMPS-FRÉQUENCE PAR TF COURT-TERME

1.1 Découpage du signal de parole

a/ Introduction

On a un signal de parole $x(\tau)$ où τ est la variable de temps, d'une durée de quelques secondes (quelques mots).

On va prendre un morceau $y(t, \tau)$ de durée T de ce signal autour de l'instant t, pendant lequel le **conduit vocal** ne change pas trop: on pourra donc considérer qu'en gros le conduit vocal est **stationnaire** durant ce temps là, et donc appliquer le formalisme classique (TF classique etc.) sur ce morceau de signal y.

b/ Durée de la découpe de signal

• La prononciation de "photographe" dure environ 1 s, et comprend environ 8 phonèmes f-o-t-o-g-r-a-f pour chacun desquels le conduit vocal a une forme précise.
En gros un phonème dure donc environ[1] 1/8 s = 125 ms

On retient que $\boxed{T < 125 \text{ ms}}$

• Par ailleurs, il faut que T comprenne plusieurs périodes du son de parole lorsqu'elle est voisée (i.e. contient un signal périodique venant de la glotte), sinon la TF du signal découpé ne pourra pas représenter correctement le signal de glotte. La hauteur la plus basse est obtenue avec la voix d'un homme "basse" est d'environ 50 Hz.

D'où $\boxed{T > 1/50 \text{ s} = 20 \text{ ms.}}$

Lorsque T augmente, la représentation de l'évolution du conduit vocal au cours du temps sera plus grossière mais on aura un spectre plus précis .

Un compromis serait a priori: $T \approx 40$ ou 50 ms mais ceci est très approximatif, et la valeur doit être choisie expérimentalement selon les cas (articulation rapide/lente, voix masculine/féminine etc.).

On va bien sûr choisir T multiple entier N de la période d'échantillonnage T_e du signal $x(\tau)$:

$\boxed{T = N \,.\, T_e = N / f_e}$ avec f_e = fréquence d'échantillonnage = $1/T_e$

On déduit $N = T \,.\, f_e$

Lors de l'échantillonnage, pour chaque morceau, *en excluant le dernier point qui est commun avec le morceau suivant*, nous avons bien N échantillons uniques par morceau, et donc N échantillons à traiter. Nous verrons que ce nombre **doit être une puissance de 2** afin de pouvoir utiliser la FFT ("Fast Fourier Transform") pour calculer rapidement la DFT ("Discrete Fourier Transform") c'est à dire la TF discrète (où le temps a des valeurs discrètes comme $n \,.T_e$ avec n entier au lieu d'être continu comme t pour la TF).

Exemple: Supposons que le signal soit échantillonné à la fréquence d'échantillonnage d'un CD-A: $f_e = 44100$ Hz
Alors pour $T_{voulu} = 50$ ms , on a : $N_{voulu} = T_{voulu} \,.\, f_e = 50 \times 10^{-3} \times 44100 = 2205$ échantillons uniques dans chaque découpe. Rappelons les puissances de 2 à partir de 2^6: 64, 128, 256, 512, 1024, 2048, 4096, 8192 ...

Ici la puissance de 2 la plus proche est donc : $N = 2048 = 2^{11}$

Ceci change un peu la valeur de $T = N / f_e = 2048 / 44100 = 46{,}4399$ ms mais le choix T_{voulu} était approximatif.

[1] Mais ceci comprend aussi les transitions entre phonèmes ...
Pour en tenir compte, la durée de stationnarité est donc d'environ la moitié soit 1/8 / 2 = 1/16 s, presque 1/20 s = 50 ms.
L'oreille n'entend les sons qu'à partir de 20 Hz, ce n'est pas un hasard ...)

c/ Découpage

Le morceau est centré sur l'instant t et dans l'intervalle [-T/2 ; T/2[.

On note que prendre une morceau de durée T centré sur $\tau = t$ revient à multiplier le signal $x(\tau)$ par une fonction "porte" brutale de durée T et placée en t :

$y(t, \tau) = x(\tau) . \prod_T(\tau - t)$

Pour éviter les problèmes spectraux (ondulations "ripples" anormaux dans le spectre) dus à un découpage brutal, on va le multiplier par une fenêtre $w(\tau)$ plus progressive, en forme de cloche, de durée T et centrée sur l'instant t :

$y(t, \tau) = x(\tau) . w(\tau - t)$ avec $\tau - t \in [-T/2 ; T/2[$

Finalement, on va ramener le morceau découpé $y(t, \tau)$ depuis l'instant t à l'instant zéro pour obtenir une découpe :

$u(t, s) = y(t, s + t)$ avec la position relative au centre de la fenêtre $s = \tau - t \in [-T/2 ; T/2[$
(attention, s n'a rien à voir avec la variable de Laplace)
D'où :

$u(t, s) = x(s + t) . w(s)$

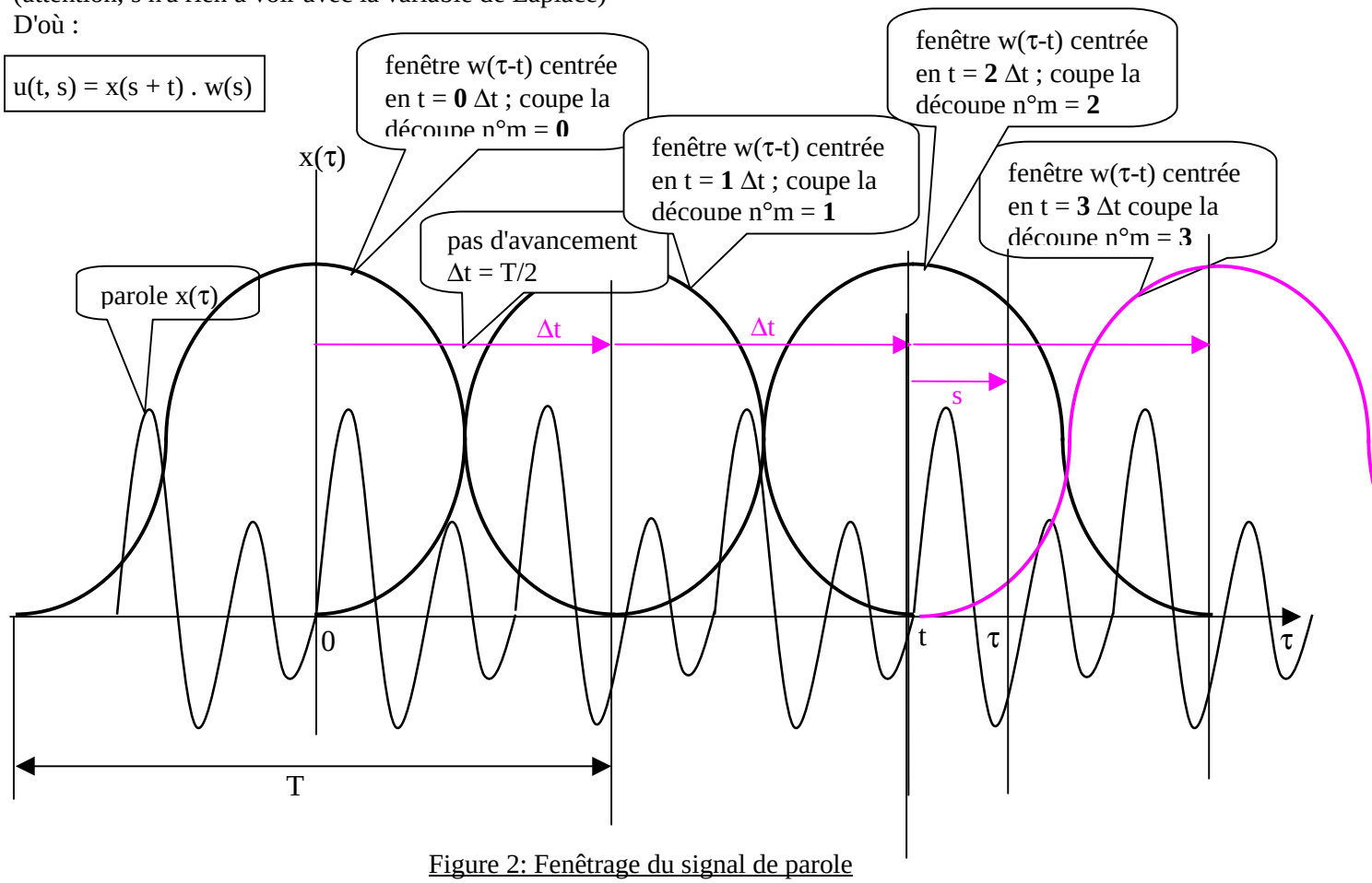

Figure 2: Fenêtrage du signal de parole

d/ Choix de la fenêtre

Cette fenêtre $w(\tau)$ est ici une **fonction de Hann**, une cloche de hauteur 1 et de largeur de base T, nulle en dehors de [-T/2 ; T/2[; elle est obtenue en relevant de 1 une période de fonction cosinus ("**raised cosine**": cosinus relevé) de période T et en divisant le tout par 2 :

$w(s) = [1 + \cos(2 \pi . s / T)]/2$ pour $s \in [-T/2 ; T/2[$

$w(s) = 0$ pour $s = < -T/2$ ou $\tau \geq T/2$

Numériquement:

en appelant $s = q \cdot T_e$

$w[q] = [1 + \cos(2\pi \cdot q / N)] / 2$ pour $q \in [-N/2 .. N/2-1]$

(On note que $w[-N/2] = 0$ ainsi que $w[N/2] = 0$, ce qui justifie de supprimer le point $q = N/2$).

e/ Avancement

Pour traiter tout le signal, on déplace la fenêtre d'un pas Δt , en incrémentant t de Δt, recommence le calcul de la découpe u(t, s) , et ainsi de suite (cf. figure 2).

Pour ne rien perdre du signal, il faut bien sûr que $\Delta t < T$.

La valeur idéale du pas d'avancement est $\Delta t = T/2$.

Numériquement:

En appelant $t = n \cdot T_e$, $\Delta n = N/2$

Chaque découpe n° m du signal $x(\tau)$ aura été prise autour de l'instant $t = m \cdot \Delta t$
c'est à dire au n° d'échantillon $n = m \cdot \Delta n$

En discrétisant le temps avec (cf. figure 3):

$s = q \cdot T_e$
$t = m \cdot \Delta t$

on peut échantillonner :

$u(t, s) = x(s + t) \cdot w(s)$

$u(m. \Delta t, q.T_e) = x(q.T_e + m. \Delta t) \cdot w(q.T_e)$

$m. \Delta t = m \cdot \Delta n \cdot T_e$

On peut donc écrire pour chaque découpe n° m de signal ramenée en t=0 :

$u[m, q] = x[m \cdot \Delta n + q] \cdot w[q]$ pour $q \in [-N/2 .. N/2-1]$ et $m \in [0 .. M_{max}]$ (i.e. jusqu'à la fin du signal x[n])

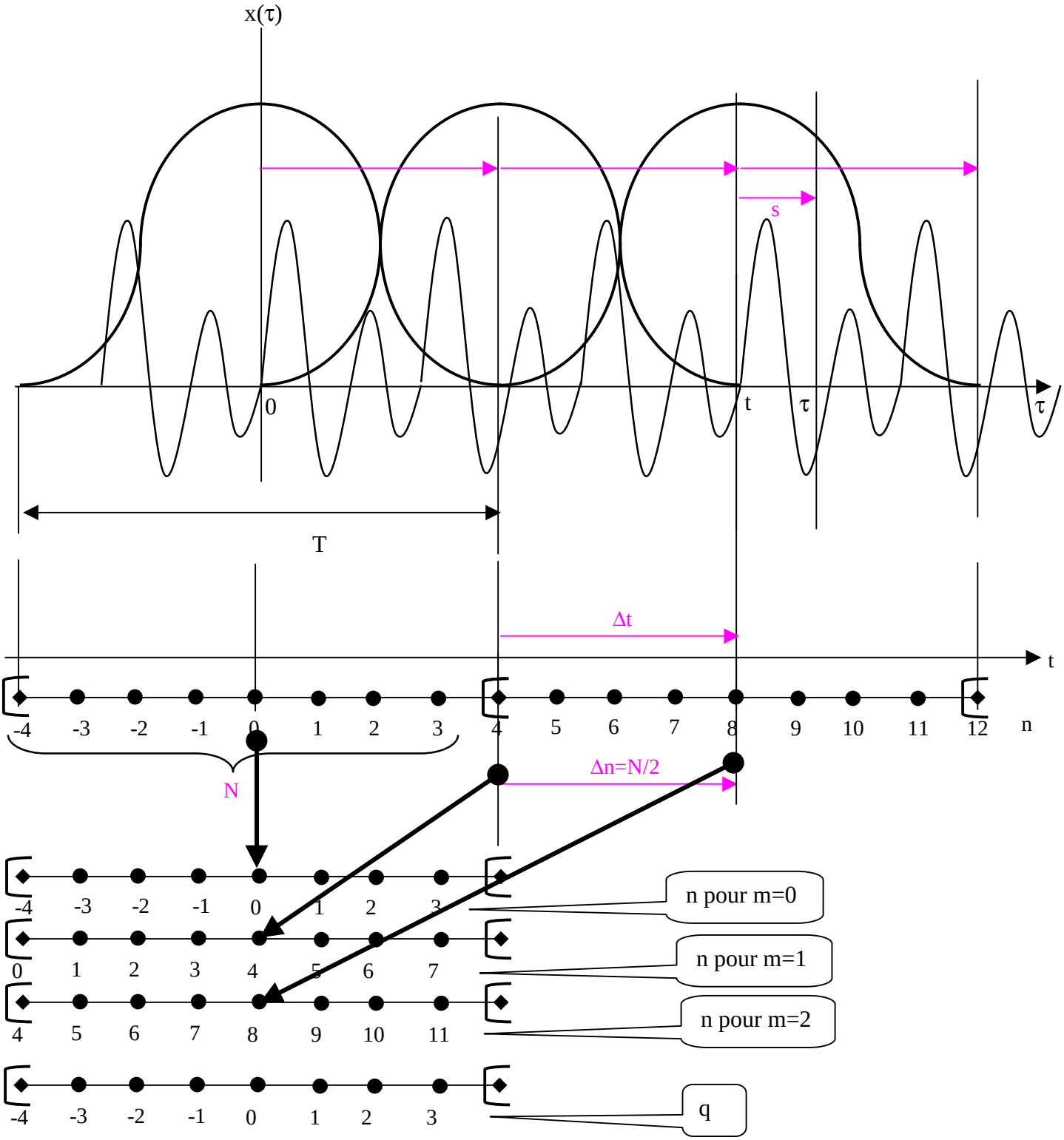

Figure 3: Fenêtrage numérique du signal de parole

f/ Reconstruction

On note qu'on peut reproduire une valeur x(τ) avec τ placé entre les instants du milieu de deux découpes successives n°m et m+1 centrées respectivement en m.Δt et (m+1).Δt ;

en appelant s la position relativement au centre de la fenêtre t=m.Δt :

$s = \tau - m.\Delta t$

découpe n° m ramenée en 0: u(m.Δt , s)

découpe n° m+1 ramenée en 0: u((m+1).Δt , s)

Du fait de la symétrie de w(τ) on peut facilement montrer que :

x(τ) = u(m.Δt , s) + u((m+1).Δt , s-Δt)

Ceci confirme le choix de la fenêtre symétrique et du pas d'avancement Δt = T/2.

Numériquement:

Supposons qu'on veuille reproduire l'échantillon x[n] n°n du signal x(τ) avec $\tau = n . T_e$.

En appelant m et m+1 les n° des découpes successives auxquelles appartient l'échantillon n°n :

n = m . Δn + q avec q ∈ [0 .. Δn - 1]

On peut donc trouver m et q par une division entière du dividende n par le diviseur Δn ; m est le quotient et q le reste :

m = Ent[n / Δn]

q = n - m . Δn

On a alors:

x[n] = u[m, q] + u[m+1, q-Δn]

Note: Il est facile de reproduire tout échantillon x[n] à partir des deux découpes successives n° m et m+1 auxquelles il appartient à cause de la symétrie de la fenêtre, et parce qu'on ne doit considérer que deux découpes pour reconstituer un échantillon, car la largeur de la fenêtre est aussi de deux pas d'avancement.
C'est pourquoi ce type de Vocoder est choisi ici.
En revanche, ceci constitue un inconvénient quant à la qualité: d'autres solutions utilisent des fenêtres compliquées et beaucoup plus grandes que le pas d'avancement pour améliorer ce point, au détriment de la simplicité et de la quantité de calculs nécessaires.

1.2. Transformation de Fourier à court terme

Par définition, la transformée de Fourier à court terme est la transformée de Fourier du signal découpé y :

$X(t, f) = TF_\tau\{ y(t, \tau) \}(f) = TF_\tau\{ x(\tau) . w(\tau - t) \}(f)$

la TF étant appliquée sur τ et le spectre obtenu est fonction de f mais aussi toujours de t.

Mais si on prend la TF de la découpe du signal ramenée en t=0 :

$TF_s\{ u(t, s) \} = TF_s\{ y(t, s + t) \} = TF_s\{ y(t, s) \} . \exp(2.i.\pi.f . t) = X(t, f) . \exp(2.i.\pi.f . t)$

D'où :

$X(t, f) = TF_s\{ u(t, s) \} . \exp(2.i.\pi.f . t)$ avec $s \in [-T/2 ; T/2]$

Toutefois, si on ne s'intéresse qu'au module $| X(t, f) |$ du spectre X(t, f), on peut ignorer le facteur de phase $\exp(2.i.\pi.f . t)$, car:

$| X(t, f) | = | TF\{ u(t, \tau) \} | . | \exp(2.i.\pi. f . t) | = | TF\{ u(t, \tau) \} | . 1 = | TF\{ u(t, \tau) \} |$

<u>Numériquement:</u>

On peut montrer qu'en ignorant facteur d'échelle et correction de phase (inutile pour l'effet recherché ici) l'échantillonnage du spectre temps-fréquence est :

$X[m, k] = DFT_N\{ u[m , q-N/2] \}[k]$ la DFT agit sur q dans [0 .. N-1] , k est dans [0 .. N-1]

Ici u[m, q] est un tableau avec q dans [-N/2 .. N/2-1] et m

<u>Note:</u> Ces formules doivent être adaptées lors de l'implémentation avec par exemple q' dans [0 .. N-1] pour le langage C ou dans [1 .. N] pour Matlab ou Fortran.

Par exemple, pour C on aurait :

$X[m, k] = DFT_N\{ u[m , q'] \}$ avec q' dans [0 .. N-1]

A ce stade, on peut facilement reconstituer u[m, q] à partir de X[m, k] avec la DFT inverse IDFT et de là x[n] pour vérifier que le Vocoder fonctionne bien.

<u>Note : Place mémoire</u>
Il y a un spectre instantané de N valeurs complexes pour N/2 échantillons d'entrée; on a donc N / (N/2) = 2 valeurs complexes par échantillon. Chaque valeur complexe comprend deux valeurs réelles (une pour la partie réelle et une pour la partie imaginaire).
(On peut diviser par 2 en notant que le spectre est symétrique: X[m, k] = X[m, N-k]
On peut se contenter de conserver le spectre pour k dans [0 .. N/2-1] et compléter par symétrie.)

Pour un signal échantillonné à f_e , on a donc 2 f_e complexes/s soit 44100 x 2 = 88200 complexes/s en CD-A.
10 s de voix nécessitent donc 888200 complexes soit (en utilisant un codage flottant double-précision à 64 bits)
2 x 8 x 888200 ≈ 14 Mo.

Les ordinateurs ayant tous aujourd'hui au moins 1 Go de mémoire, le tableau devrait y tenir largement !

2. ESTIMATION DE LA RÉPONSE FRÉQUENTIELLE À COURT TERME DU CONDUIT VOCAL

2.1 Théorie

a/ Formation du spectre de voix

Prenons un conduit vocal stationnaire pour simplifier, de réponse impulsionnelle h(t).

En supposant que la glotte fournisse un signal d'excitation b(t), la voix obtenue est :

$x(t) = b(t) * h(t)$

Le spectre de ce signal est :

$X(f) = TF\{ x(t) \}= TF\{ b(t) * h(t) \} = B(f) . H(f)$

B(f) est le spectre du signal d'excitation original de la glotte, et H(f) la réponse fréquentielle du conduit vocal.

H(f) varie lentement avec la fréquence f: elle est constituée de quelques formants en cloche entre 300 Hz et 3 kHz.

En revanche, quand le signal vocal est voisé, b(t) est quasi-périodique et donc B(f) est constitué de pics équidistants de f_p , la fréquence fondamentale (hauteur, "pitch") de la voix. Le signal de glotte étant constitué d'impulsions périodique très fines, les pics de son spectre B(f) sont d'amplitude à peu près égale.

Chaque pic n° n_h correspond à une harmonique de fréquence $n_h . f_p$

En multipliant les deux fonctions, on obtient donc un spectre constitué de nombreux pics du à B(f) "modulés" par une douce enveloppe H(f).

Pour un homme : $f_p \approx 100$ Hz

largeur d'un formant : $w_f > 500$ Hz

On voit donc que les formants sont beaucoup plus larges que la distance entre les harmoniques: il semble donc possible de récupérer une estimation de |H(f)| en lissant |X(f)|, c'est à dire en appliquant un filtre passe-bas sur ce spectre.

b/ Filtrage homomorphique

Prenons le logarithme de X(f) :

$C(f) = Ln(X(f)) = Ln(B(f) . H(f)) = Ln(B(f)) + Ln(H(f))$

Dans C(f) on veut garder Ln(H(f)) qui change lentement avec f, et supprimer Ln(B(f)) qui varie vite avec f.

Si on calcule le (super-)spectre de C(f) en prenant sa transformée de Fourier, on obtient une fonction c(θ) appelée **cepstre** de x(t) (où θ est équivalent à une super-fréquence: la quéfrence).

Dans ce super-spectre les composantes additives de C(f) variant rapidement avec f comme Ln(B(f)) seront localisées en θ élevé alors que celles variant lentement comme H(f) seront localisées en θ faible. En supprimant les parties de quéfrence θ élevée de c(θ) on obtient alors à peu près le spectre de Ln(H(f)).

Par convention, on utilise la TF inverse pour calculer le cepstre, mais ça ne change rien au raisonnement ci-dessus : rappelons qu'entre TF et TF^{-1}, la seule différence est que i change de signe.

$$c(\theta) = TF^{-1}\{ C(f) \} = TF^{-1}\{ Ln(X(f)) \}$$

Note: Les cepstre est réel puisqu'il est la TF inverse d'un spectre (complexe de symétrie hermitienne).
On peut donc facilement l'afficher.

On calcule le nouveau cepstre $c'(\theta) = c(\theta)$ pour $\theta \in [-\theta_{seuil} .. \theta_{seuil}]$
sinon : $c'(\theta) = 0$
Le cepstre étant symétrique, il faut en effet aussi supprimer les quéfrences fortes mais négatives.
(le prime ' de $c'(\theta)$ n'a ici rien à voir avec la dérivée).
On peut écrire ceci avec une expression, en appelant $\prod_{2\theta seuil}(t)$ la fonction porte de largeur $2\theta_{seuil}$ centrée en $\theta=0$

$$c'(\theta) = \prod_{2\theta seuil}(\theta) \cdot c(\theta)$$

Ceci revient à appliquer un passe-bas sur C(f), c'est à dire à la lisser afin d'obtenir C'(f); on lisse donc le logarithme de X(f) (et pas seulement son module).

C'(f) et donc $c'(\theta)$ ne dépendent plus que de H(f) . On obtient alors une estimation $\hat{H}(f)$ de H(f) par :

$$C'(f) = Ln(\hat{H}(f)) = TF\{ c'(\theta) \}$$

$$\hat{H}(f) = \exp (TF\{ c'(\theta) \})$$

En théorie, on considérant le logarithme complexe, ce procédé permet de d'estimer non seulement le module |H(f)| de H(f) mais aussi sa phase $\angle H(f)$!

En effet, $H(f) = |H(f)| \cdot \exp(i.\angle H(f))$

Donc :

$Ln(H(f)) = Ln[|H(f)| \cdot \exp(i\angle H(f))] = Ln[|H(f)|] + Ln[\exp(i.\angle H(f))] = Ln[|H(f)|] + i.\angle H(f)$

$Ln[|H(f)|]$ étant réel, la partie imaginaire de $Ln(H(f))$ est donc $\angle H(f)$.

Toutefois, ce point est techniquement délicat, et de toute façon la phase de H(t, f) varie assez peu avec le temps ; or seule la variation temporelle de phase est audible, car elle équivaut à un décalage en fréquence;
la phase de H(f) n'est donc pas nécessaire pour l'application recherchée.

2.2. Calcul du cepstre à court terme

Nous n'allons pas chercher à gérer la phase comme expliqué ci-dessus. Par conséquent, nous n'allons considérer que le module |H(f)| de la réponse fréquentielle H(f) et |X(f)| du spectre X(f) de x(t).

Le cepstre à court terme est alors :

$$c(t, \theta) = TF^{-1}\{ Ln[| X(t, f) |] \}$$

avec TF^{-1} s'appliquant sur f

θ est la quéfrence en secondes (inverse pour la fréquence en Hz).

Numériquement:

$$c[m, j] = IDFT_N\{ Ln(| X[m, k] |) \}[j]$$ l'IDFT agit sur k dans [0 .. N-1] ; j = numéro de quéfrence dans [0 .. N-1]

Note: dans la IDFT c[m, -j] pour j > 0 est c[m, N-j] car le spectre numérique est N-périodique.
Normalement on doit avoir c[m, j] réel.

2.3. Fenêtrage du cepstre

Il faut visualiser le cepstre en fonction de la quéfrence θ à divers instants t = m . Δt.

Le cepstre pour les quéfrences θ faibles < θ_{seuil} contient l'information sur la réponse fréquentielle du conduit vocal.

En revanche, pour les quéfrences θ élevées > θ_{seuil}, il s'agit du signal d'excitation de la glotte (cordes vocales).

Nous allons donc garder le cepstre original pour $\theta < \theta_{seuil}$ et mettre c(θ) à 0 pour $\theta > \theta_{seuil}$.

Numériquement:

Nouveau cepstre $c'[m, j] = 0$ pour $j \in [j_{seuil} .. N-j_{seuil}]$

$c'[m, j] = c[m, j]$ sinon

à vérifier

La valeur j_{seuil} sera déterminée expérimentalement.

2.4 Estimation de la réponse fréquentielle à court terme du conduit vocal

$C'[m, k] = DFT_N\{ c'[m, j] \}[k]$ la DFT agit sur j dans [0 .. N-1] et k est dans [0 .. N-1]

Note: Normalement, C'[m, k] a sa partie imaginaire nulle à de faibles valeurs près dues à l'imperfection du calcul.

Nous avons :

$\ln(\hat{H}[m, k]) = C'[m, k]$

Donc la réponse fréquentielle numérique instantanée estimée du conduit vocal :

$\hat{H}[m, k] = \exp(C'[m, k])$ pour k dans [0 .. N-1]

Note: normalement, $\hat{H}[m, k]$ devrait $\in R^+$ (à de faibles valeurs près dues à l'imperfection du calcul).

3. CALCUL DU SPECTRE COURT-TERME DU SON QUELCONQUE

Le son e(τ) qui va servir d'excitation en lieu et place de celle produite par la glotte va être mis sous la forme d'une représentation temps-fréquence numérique de la même manière que pour x(τ); on obtient : E[m, k]

3.1. Calcul des découpes ramenées en 0:

$p[m, q] = e[m . \Delta n + q] . w[q]$ pour $q \in [-N/2 .. N/2-1]$ et $m \in [0 .. M_{max}]$ (i.e. jusqu'à la fin du signal x[n])

3.2 Calcul du spectre à court terme:

$E[m, k] = DFT_N \{ p[m , q-N/2] \}[k]$ la DFT agit sur q dans [0 .. N-1] , k est dans [0 .. N-1]

4. ACTION DU CONDUIT VOCAL SUR LA NOUVELLE EXCITATION

La nouvelle voix aura pour représentation temps-fréquence :

$$V[m, k] = E[m, k] . \hat{H}[m, k]$$ pour k dans [0 .. N-1] et m dans [0 .. Mmax] (sur tout le signal)

En effet, pour chaque découpe n°m, on applique la relation classique qui est que le spectre du signal de sortie est le produit du spectre du signal d'entrée par la réponse fréquentielle du conduit vocal.
Il y a ici un facteur d'échelle à considérer, mais nous ne le calculerons pas ici afin de simplifier l'exposé.

Il faudra néanmoins multiplier le produit par un facteur d'échelle empirique afin que le signal obtenu soit d'une valeur raisonnable.

5. RECONSTRUCTION DU SIGNAL DE LA VOIX ARTIFICIELLE

On va reconstituer v[n] à partir de V[m, k]

5.1 Calcul des découpes

Tout d'abord, on va calculer les découpes a[m, q] temporelles ramenées en 0 du signal v à partir de la représentation temps-fréquence :

$a[m, q] = IDFT_N\{ V[m, k] \}[q+N/2]$ l'IDFT agit sur k dans [0 .. N-1] et fournit q+N/2 dans [0 .. N-1]

5.2. Reconstruction du signal

Reconstruction de v[n] à partir du tableau des découpes (comme vu précédemment) :

pour n dans $[0 .. n_{max}]$: (pour tout le signal x[n])

$m = Ent[n / \Delta n]$

$q = n - m . \Delta n$

$v[n] = a[m, q] + a[m+1, q-\Delta n]$

6. RÉSUMÉ DU TRAITEMENT NUMÉRIQUE :

TF court terme du signal de parole:

Nombre d'échantillons dans la fenêtre :
N = 2048
Pas d'avancement: $\Delta n = N/2$
Calcul de la fenêtre :
$w[q] = [1 + \cos(2\pi . q / N)] / 2$ pour $q \in [-N/2 .. N/2-1]$

- N est une puissance de 2 à régler expérimentalement selon la fréquence d'échantillonnage;
- j_{seuil} à régler expérimentalement
- prévoir des facteurs d'échelle pour certaines étapes afin d'avoir des valeurs raisonnables.

Calcul des découpes de x ramenées en 0 :
$u[m, q] = x[m . \Delta n + q] . w[q]$ pour $q \in [-N/2 .. N/2-1]$ et $m \in [0 .. M_{max}]$ (i.e. jusqu'à la fin du signal x[n])

Calcul du spectre court terme du signal de parole :
$X[m, k] = DFT_N \{ u[m , q-N/2] \}[k]$ la DFT agit sur q dans [0 .. N-1] , k est dans [0 .. N-1]

Estimation de la réponse fréquentielle du conduit vocal par filtrage homomorphique :

Calcul du cepstre court terme du signal de parole :
$c[m, j] = IDFT_N\{ Ln(| X[m, k] |) \}[j]$ l'IDFT agit sur k dans [0 .. N-1] ; j est dans [0 .. N-1]

Fenêtrage du cesptre (extraction des composantes décrivant H dans le cepstre) :
$c'[m, j] = 0$ pour $j \in [j_{seuil} .. N-j_{seuil}]$ sinon c(m, j)

à vérifier

Calcul du logarithme de la réponse fréquentielle
$C'[m, k] = DFT_N\{ c'[m, j] \}[k]$ la DFT agit sur j dans [0 .. N-1] et k est dans [0 .. N-1]

Calcul de l'estimée de la réponse fréquentielle
$\hat{H}[m, k] = \exp(DFT_N\{ c'[m, j] \}[k])$ la DFT agit sur j dans [0 .. N-1] k dans [0 .. N-1]

TF court terme** du son quelconque**:

Calcul des découpes de x ramenées en 0 :
$p[m, q] = e[m . \Delta n + q] . w[q]$ pour $q \in [-N/2 .. N/2-1]$ et $m \in [0 .. M_{max}]$ (i.e. jusqu'à la fin du signal x[n])

Calcul du spectre court terme du son quelconque:
$E[m, k] = DFT_N \{ p[m , q-N/2] \}[k]$ la DFT agit sur q dans [0 .. N-1] , k est dans [0 .. N-1]

***Action du conduit vocal** sur le spectre court-terme du son quelconque pour obtenir celui de la voix artificielle:*

$V[m, k] = E[m, k] . \hat{H}[m, k]$ pour k dans [0 .. N-1] et m dans [0 .. Mmax] (sur tout le signal)

TF court terme inverse:

Calcul des découpes de x ramenées en 0 :
$a[m, q] = IDFT_N\{ V[m, k] \}[q+N/2]$ l'IDFT agit sur k dans [0 .. N-1] et fournit q+N/2 dans [0 .. N-1]

Calcul du n° m de la découpe :
$m = Ent[n / \Delta n]$ pour n dans $[0 .. n_{max}]$: (pour tout le signal x[n])

calcul de l'index dans la découpe :
$q = n - m . \Delta n$

calcul de l'échantillon de voix artificielle:
v[n] = $a[m, q] + a[m+1, q-\Delta n]$