Modifications de hauteur, d'échelle temporelle et de timbre

Bertrand David
bertrand.david@telecom-paristech.fr

Table des matières

Ce document s'inspire largement de passages de divers documents et principalement d'un ouvrage spécifiquement dédié au traitement du signal audio [9](Chap. 7). Il développe plus particulièrement les méthodes à base de vocodeur de phase et les méthodes temporelles.

Bertrand David bertrand.david@telecom-paristech.fr

Chapitre 1

Modifications de hauteur, d'échelle temporelle et de timbre

1.1 Introduction

L'objectif de ces modifications, qui correspondent à des besoins courants dans divers domaines du traitement du son et de la parole, est de contrôler de *manière indépendante* les évolutions temporelles, fréquentielle et éventuellement formantiques (variations lentes du spectre) du signal :

- dilatation temporelle : on doit pouvoir modifier les échelles de durée sans altérer le contenu spectral et particulièrement la hauteur dans le cas d'un signal harmonique,
- variation de hauteur : pour un signal harmonique, on doit pouvoir modifier la hauteur du son en conservant son évolution temporelle (par exemple le débit prosodique dans le cas de la parole) et en particulier sa durée,
- contrôle formantique : dans le cas d'une modification de hauteur, on peut choisir de modifier l'échelle spectrale dans son ensemble (et donc de déplacer les formants ou l'enveloppe spectrale) ou de garder l'enveloppe spectrale constante tout en effectuant une transposition du spectre de raies.

Les applications où ces types de modifications indépendantes sont nombreuses :

- synthèse par échantillonnage d'une table d'onde (sons musicaux ou segments de parole [1]),
- post-synchronisation : pour effectuer un "recalage" du son et de l'image,
- compression de données [13]
- lecture pour aveugle : notre lecture intérieure est beaucoup plus rapide que notre diction. Par rétrécissement des durées on peut permettre aux aveugles d'augmenter leur rapidité de parcours des documents,
- apprentissage des langues étrangères : un ralentissement du débit est appréciable,
- post-production musicale : pour raccorder plusieurs prises de son il peut être utile d'accélérer ou de réduire légèrement le tempo. Il peut être également intéressant de corriger localement la justesse d'une voix ou d'un instrument.

On peut classer en trois types les méthodes pour effectuer ces modifications :

- les méthodes inspirées de la tête de lecture circulaire ou radio cassette modifié (on ajoute/retranche des portions du signal [6]). Ces méthodes sont dites *temporelles*,
- méthodes à base de vocodeur de phase (méthodes *fréquentielles* utilisant la TFCT [18, 20]),
- méthodes à modèle de signaux (LPC [12], Sinus+Bruit [23], Grains audio [8],...).

Les méthodes temporelles ont donné lieu à de nombreux développements en numérique : les méthodes SOLA (Synchronized Overlapp Add) [19], PSOLA (Pitch Synchronized Overlapp Add) [16]. Cette dernière utilise une technique de recopie/suppression synchronisée sur les impulsions glottiques. On parvient ainsi à une très bonne qualité de modification d'échelle temporelle *sans rééchantillonnage du signal*.

La méthode PSOLA peut être adaptée pour effectuer des modifications formantiques en modifiant la durée des segments sans modifier la position des impulsions glottiques. De cette manière, on peut transposer l'enveloppe spectrale sans modification de hauteur ni de durée et modifier ainsi le timbre d'une voix (transformer une voix masculine en voix plus féminine par exemple). D'autres techniques de modifications formantiques utilisent des représentations cepstrales [3].

1.2 Modèles de signaux pour définir les distorsions temporelles et spectrales

Une simple relecture à 16 kHz d'un signal audio échantillonné à 8 kHz suffit à nous convaincre que les dilatations ou compression temporelles et fréquentielles sont interdépendantes. Cette dépendance peut être interprétée comme un simple résultat théorique sur la transformée de Fourier (TF) : la relation d'incertitude de Heisenberg traduit cette dépendance en terme de supports et la décroissance HF de la TF est reliée à la régularité du signal temporel. La définition des distorsions temporelles et fréquentielles de manière *indépendante* ne peut dès lors n'être obtenue que pour des *modèles* de signaux bien définis.

1.2.1 Modèle McAuley-Quatieri

Modèle de production vocale. Le modèle le plus courant et le plus utilisé de signal de parole est celui d'un filtre linéaire variant dans le temps, excité par une source harmonique (sons voisés) ou par un processus aléatoire stationnaire à spectre plat (sons non voisés). Dans le cas présent, nous considérons le cas voisé, pour lequel cette source est une somme de composantes sinusoïdales dont les fréquences instantanées sont multiples d'une fréquence fondamentale $f_0(t)$. Cette représentation est équivalente à écrire la source comme une peigne de Dirac dont la période dépend du temps.

Soit $g_t(\tau)$ la réponse impulsionnelle du système à l'instant t. Le signal s'écrit alors simplement en fonction de l'excitation $e(t)$:

$$x(t) = \int_{-\infty}^{+\infty} g_t(\tau)e(t-\tau)d\tau. \tag{1.1}$$

Ce système non invariant dans le temps peut être représenté par une réponse en fréquence dépendante du temps :

$$G(t, f) = M(t, f) \exp j\varphi(t, f).$$

Les variations temporelles de g_t sont liées au mouvements articulatoires et sont considérées comme lentes devant la période fondamentale du signal. D'autre part, ces variations sont supposées faibles sur la durée de la mémoire du filtre. Le système est *quasi stationnaire*.

Pour de la parole voisée, c'est à dire faisant intervenir la vibration périodique des cordes vocales, le signal d'excitation s'écrit :

$$e(t) = \sum_{k=-\infty}^{+\infty} \exp j\xi_k(t) \tag{1.2}$$

avec

$$\xi_k'(t) = 2\pi f_k(t).$$

Le caractère quasi-stationnaire de g_t entraîne une limitation pratique du support de cette fonction à une dimension de l'ordre de la mémoire du système. L'intégrale continue de l'expression 1.1 est donc définie en pratique. De même le support fréquentiel de la parole est limité en pratique et la somme discrète de l'expression 1.2 est de fait une somme finie de $L(t)$ termes exponentiels complexes. En tenant compte du fait que f_0 varie peu sur la durée de mémoire du filtre on peut développer

$$\xi_k(t-\tau) \approx \xi_k(t) - 2\pi\tau k f_0(t)$$

Bertrand David bertrand.david@telecom-paristech.fr

au voisinage de t (*i.e.* pour des τ inférieur à la mémoire du filtre). Il vient alors :

$$x(t) = \sum_{k=1}^{L(t)} M(t, f_k(t)) \exp j[\xi_k(t) + \varphi(t, f_k(t))] \tag{1.3}$$

Modèle de Mc-Auley et Quatieri. Ce modèle a été introduit par McAulay et Quatieri vers 1985 [15] principalement en vue du codage bas débit de la parole. Il fait donc référence à l'expression obtenue en 1.3. Il est cependant un peu plus général puisque ne supposant pas une relation nécessaire harmonique entre les fréquences instantanées. Le signal est représenté sous la forme d'une somme de sinusoïdes dont les fréquences, les amplitudes et les phases sont contrôlées au cours du temps.

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t) \qquad \text{avec} \qquad \Psi'_k(t) = \omega_k(t) = 2\pi f_k(t) \tag{1.4}$$

où $A_k(t)$ est l'amplitude à l'instant t de la sinusoïde k, $\Psi_k(t)$ est la *phase instantanée* de cette sinusoïde à l'instant t et $f_k(t)$ est sa *fréquence instantanée*. Cette décomposition n'est pas univoque et on admet généralement que les fonctions $A_k(t)$ et $\omega_k(t)$ sont à variation lente par rapport aux fonctions $\exp(j\Psi_k(t))$.

1.2.2 Modèle Serra-Smith

Ce modèle fut développé au début des années 90 [23] pour répondre au besoin d'un système d'analyse/synthèse tenant compte de la composante bruitée des signaux de musique ou de parole. Cette composante est très coûteuse à représenter sous forme de sinusoïdes. Le modèle proposé est donc une extension de celui de MacAuley-Quatieri :

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t) + b(t) \tag{1.5}$$

où $b(t)$ est un processus aléatoire stationnaire filtré par un filtre variant dans le temps, à l'instar du filtre g_t présenté plus haut. Soit h_t ce filtre, on écrira alors en tenant compte de la causalité des signaux :

$$b(t) = \int_0^t h_t(\tau) u(t - \tau) d\tau \tag{1.6}$$

où $u(t)$ est un processus aléatoire stationnaire blanc.

Le système complet d'analyse/modification/synthèse comprend

- une phase d'estimation des composantes déterministes,
- une phase d'interpolation linéaire des amplitudes et cubique des phases d'une trame à l'autre du signal pour ces composantes,
- la soustraction de cette partie déterministe pour obtenir $b(t)$ pour chaque trame,
- application d'un algorithme de transformation possiblement distinct pour chacune des deux composantes,
- resynthèse.

1.3 Définitions et équivalences

Toutes les définitions données ici sont relatives à un modèle de signal à composantes sinusoïdales. Elles s'appliquent donc au modèle McAuley-Quatieri ou à la partie déterministe du modèle Serra-Smith. Les phases à $t = 0$ seront supposées nulles par simplification (ce terme peut être incorporé dans la définition des amplitudes).

1.3.1 Distorsion temporelle

On définit la fonction de distorsion temporelle à l'aide de la nouvelle échelle de temps τ et de l'échelle d'origine t par :

$$\tau = T(t). \tag{1.7}$$

Cette fonction est continue et bijective de $\mathbb{R}^+$ dans $\mathbb{R}^+$. La modification d'échelle temporelle du signal $x(t)$ est alors définie par

$$y(\tau) = \sum_{k=1}^{L(T^{-1}(\tau))} A_k(T^{-1}(\tau)) \exp(j\phi_k(\tau)) \tag{1.8}$$

La conservation du contenu fréquentiel impose alors de maintenir les valeurs des fréquences instantanées, soit la relation :

$$\phi_k(\tau) = \int_0^\tau \omega_k(T^{-1}(u))du \tag{1.9}$$

1.3.2 Modification de hauteur

Pour modifier la hauteur du signal $x(t)$ on construit le signal :

$$y(t) = \sum_{k=1}^{L(t)} A_k(t) \exp(j\Phi_k(t)) \tag{1.10}$$

L'altération du contenu fréquentiel est défini à l'aide d'une fonction $\alpha(t)$ appelée taux de compression fréquentiel, selon l'expression :

$$\Phi_k(t) = \int_0^t \alpha(u)\omega_k(u)du \tag{1.11}$$

1.3.3 Réciprocité

Par un calcul rapide on montre que la séquence opératoire : $x \rightarrow x_1$ par distorsion temporelle ($\tau = T(t)$) suivie d'un simple réechelonnement temporel (*i.e.* sans maintenir les caractéristiques fréquentielles) $x_1(\tau) = y(v)$ avec $v = T^{-1}(\tau)$ est équivalente à une modification fréquentielle gouvernée par la fonction $\alpha(t) = T'(t)$, c'est à dire :

$$y(v) = \sum_{k=1}^{L(v)} A_k(v) \exp j\Phi_k(v) \tag{1.12}$$

avec

$$\Phi_k(t) = \int_0^v T'(u)\omega_k(u)du \tag{1.13}$$

Cette relation est particulièrement utile dans le cas où la distorsion temporelle correspondante est un facteur multiplicatif, comme par exemple $T(t) = 2t$. Alors $T'(t)$ est constante et l'opération de réechellonnement temporel est une simple relecture du signal obtenu à une cadence différente (par exemple, pour des signaux échantillonnés, $F'_e = 2Fe$ dans le cas précédent).

1.4 Transformée de Fourier à Court Terme

Les méthodes d'analyse/synthèse et de modification des sons reposant sur l'utilisation de la Transformée de Fourier à Court Terme (TFCT ou STFT, *Short Time Fourier Transform* en anglais) sont très courantes. L'outil correspondant est généralement appelé *Phase VoCoder* ou vocodeur de phase. Il désigne la représentation polaire (module & phase) de la TFCT.

Bertrand David bertrand.david@telecom-paristech.fr

1.4.1 Rappels théoriques

Le schéma de principe de la TFCT est représenté dans la figure 1.1, tel qu'elle est réalisée numériquement. Le principe est celui d'un transformée de Fourier glissante, effectuée sur des trames recouvrantes du signal. Chacune des trames est fenêtrée par une fenêtre d'analyse. On écrira la TFCT d'un signal numérique $x(n)$ sous la forme

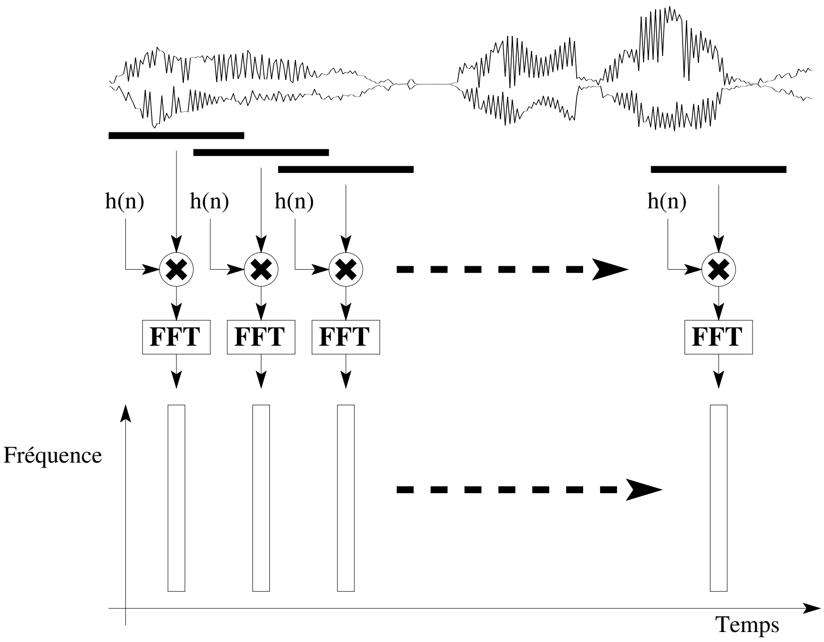

FIGURE 1.1 – Transformée de Fourier à Court Terme

$$\tilde{X}(t_a, \nu) \triangleq \sum_{n \in \mathbb{Z}} x(n + t_a) w_a(n) e^{-j2\pi\nu n}. \tag{1.14}$$

w_a désigne la fenêtre d'analyse, très couramment de longueur finie, réelle et symétrique. Les temps d'analyse sont implicitement indéxés par un entier naturel u, soit $t_a = t_a(u), \quad u \in \mathbb{N}$. On a ici préféré une notation fonction de ν alors qu'une notation fonction de $e^{j2\pi\nu}$ aurait été plus cohérente avec l'interprétation en termes de transformée de Fourier glissante, mais ce choix allège les écritures.

Interprétation. Un calcul rapide montre que, en posant $h(n) = w_a(-n) e^{j2\pi\nu_p n}$, l'écriture 1.14 peut se mettre sous la forme du produit de convolution :

$$\tilde{X}(t_a, \nu_p) = [x * h](t_a). \tag{1.15}$$

Si $w_a(n)$ est une fenêtre réelle et paire de longueur finie, sa TF $W_a(e^{j2\pi\nu})$ est réelle et paire. La TF de h est alors simplement $H(e^{j2\pi\nu}) = W_a(e^{j2\pi(\nu-\nu_p)})$ Un exemple de résultat typique est donnée figure 1.2 pour $\nu_p = 0.3$. Cet exemple montre que $\tilde{X}(t_a, \nu_p)$ réalise un filtrage RIF passe-bande autour de la fréquence ν_p. Les caractéristiques du filtre sont liées à celle de la fenêtre d'analyse choisie. Cette interprétation est à l'origine du qualificatif de *convention passe-bande* donnée à l'écriture 1.14. Il existe une autre convention, dite passe-bas, souvent utilisée pour sa facilité de manipulation calculatoire. Nous nous en tiendrons cependant à la convention passe-bande parce qu'elle correspond à la réalisation pratique.

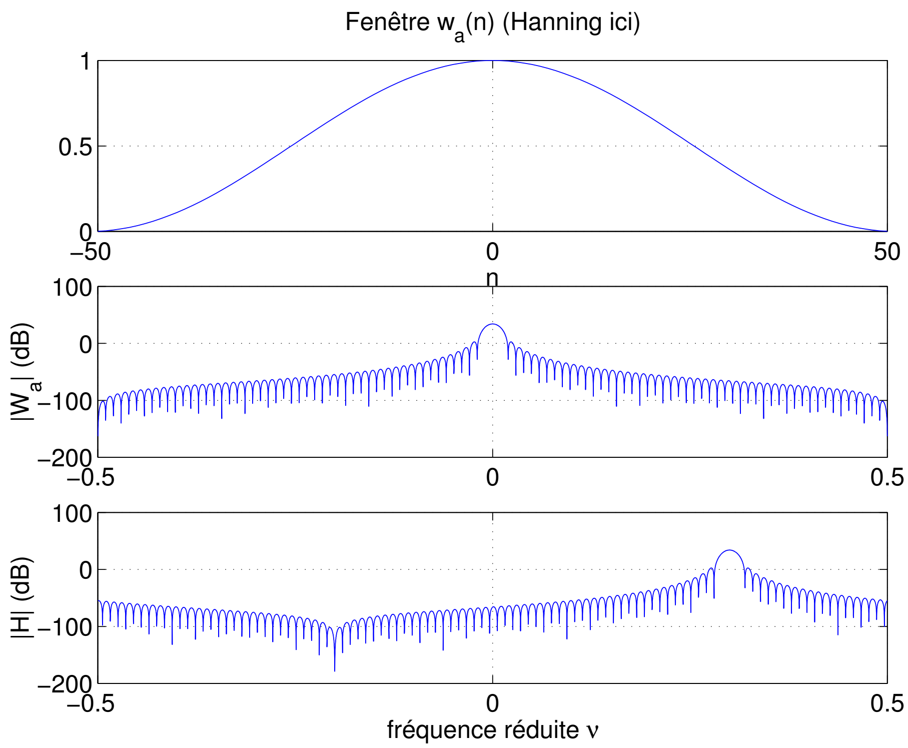

FIGURE 1.2 – Filtrage passe-bande équivalent à un canal de TFCT

Version discrète de la TFCT. En pratique, la transformée de Fourier est évaluée à l'aide de la TFD. Ceci revient à poser $\nu_p = p/N$ dans l'expression de $\tilde{X}(t_a, \nu_p)$. N est l'ordre de la TFD. On obtient ainsi une version discrète, c'est à dire échantillonnée en fréquence de la TFCT, soit

$$\tilde{X}(t_a, \nu_p) = \sum_{n=0}^{N-1} x(n + t_a) w_a(n) e^{-j2\pi \frac{pn}{N}}. \tag{1.16}$$

Pour éviter le repliement temporel, la longueur des fenêtres d'analyse sera inférieure ou égale à N.

Modifications et problèmes posés. Les modifications de son par phase vocoder impliquent l'obtention d'une TFCT modifiée à partir de $\tilde{X}(t_a, \nu_p), \quad k = 0, \ldots, N-1$ puis la resynthèse du signal. On note $t_s = t_s(u)$ les marques temporelles de synthèse. Soit la modification

$$\tilde{X}(t_a(u), \nu_p) \rightarrow Y(t_s(u), \nu_p).$$

La principale difficulté rencontrée vient de ce que Y doit satisfaire des conditions fortes [18] pour correspondre à une séquence originale bien définie. La solution de ce problème est trouvée au sens des moindres carrés [17]. Cependant, on peut toutefois écrire les conditions de reconstruction parfaite dans le cas où aucune modification n'est effectuée (*i.e.* $t_s = t_a$ et $Y = \tilde{X}$).

Condition de reconstruction parfaite. On effectue l'opération inverse de l'analyse pour la synthèse : à partir du flux de spectres discrets $\tilde{X}(t_a(u), \nu_p)$ on opère une TFD inverse et on reconstruit le signal par addition-recouvrement (overlapp-add ou OLA). Le résultat est donné par

$$y(n) = \sum_u w_s(n - t_s(u)) y_w(n - t_s(u), t_s(u)) \tag{1.17}$$

avec $t_s(u) = t_a(u)$ et

$$y_w(n, t_s(u)) = \frac{1}{N} \sum_{p=0}^{N-1} Y(t_s(u), \nu_p) e^{j2\pi\nu_p n}.$$

En tenant compte de $Y = \tilde{X}$ et en introduisant l'expression 1.14 dans 1.17, on montre que $x(n) = y(n)$ s'obtient à l'aide de la condition suffisante :

$$\sum_u w_a(n - t_a(u)) w_s(n - t_a(u)) = 1 \tag{1.18}$$

1.5 Modifications à l'aide du phase-vocoder

1.5.1 Fréquence instantanée

Les transformations indiquées dans la section 1.3 nécessitent le calcul des fréquences instantanées $\omega_k(t)$ de chacune des composantes de la somme 1.4. Ce calcul est effectué à partir de la donnée de deux spectres à court terme successifs $\tilde{X}(t_a(u), \nu_p)$ et $\tilde{X}(t_a(u+1), \nu_p)$ $p = 0, \ldots, N-1$, sous certaines conditions qui assurent l'existence d'une solution.

Condition de bande étroite. Cette première condition assure la présence d'*au plus* une composante par canal de TFCT. Le report de l'expression 1.4,

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t),$$

dans l'expression 1.14 de la TFCT donne :

$$\tilde{X}(t_a(u), \nu_p) = \sum_{n=0}^{N-1} \sum_{k=1}^{L(n+t_a)} A_k(n + t_a) \exp(j\Psi_k(n + t_a)) w_a(n) e^{-j2\pi\nu_p n}$$

On utilise alors la quasi-stationnarité du modèle à savoir :

$$\begin{aligned} A_k(n + t_a) &\approx A_k(t_a) \\ \Psi_k(n + t_a) &\approx \Psi_k(t_a) + n\omega_k(t_a) \end{aligned}$$

soit finalement, en posant $\omega_k(t) = 2\pi f_k(t)$:

$$\tilde{X}(t_a(u), \nu_p) = \sum_{k=1}^{L(n+t_a)} A_k(t_a) \exp(j\Psi_k(t_a)) W_a(e^{j2\pi(\nu_p - f_k(t_a))}) \tag{1.19}$$

La condition de bande étroite conduit à des valeurs non négligeables de $W_a(e^{j2\pi(\nu_p - f_k(t_a))})$ pour au plus une valeur k. Soit $k = l$ cette valeur si elle existe. Si on note f_c la fréquence de coupure du filtre passe-bas dont $w_a(n)$ est la réponse impulsionnelle, alors l'existence de l implique

$$|\nu_p - f_l(t_a)| \leq f_c \ ,$$

c'est à dire que la composante numéro l est dans la bande passante du filtre correspondant au p-ième canal de la TFCT. L'expression 1.19 se réduit alors à la contribution de la seule l-ième composante :

$$\tilde{X}(t_a(u), \nu_p) = A_l(t_a) \exp(j\Psi_l(t_a)) W_a(e^{j2\pi(\nu_p - f_l(t_a))}) \tag{1.20}$$

Si on suppose que w_a est réelle et paire, et donc W_a réelle et paire, cette écriture s'interprète comme suit : la phase de la TFCT donne accès aux phases instantanées des composantes de $x(t)$ à une indétermination de 2π près et le module de la TFCT donne accès aux amplitudes instantanées de $x(t)$ à un facteur de filtrage d'amplitude près. On peut donc déduire les fréquences instantanées de chaque composante à partir de la phase du flux de spectres à court terme, à condition de lever l'indétermination de 2π.

Exemple : pour une fenêtre d'analyse de Hann de longueur L, la condition de bande étroite appliquée à un spectre de raies harmoniques (segment de parole voisé par exemple) conduit à un espacement des pics spectraux au moins égal à la largeur de bande de la transformée de Fourier de la fenêtre, soit $4/L$. Cela revient à $f_0 < 4/L$, soit une longueur de fenêtre au moins égale à 4 fois la période fondamentale.

Condition de recouvrement. On va voir ici que la levée de l'indétermination conduit à une condition de recouvrement minimal des fenêtres d'analyse. En effet, la différence de phase entre deux instants d'analyse successifs, pour le p-ième canal de TFCT s'écrit, en posant $\Phi(t_a(u), \nu_p) = \arg \tilde{X}(t_a(u), \nu_p)$

$$\begin{aligned}
\Delta\Phi_p &= \Phi(t_a(u+1), \nu_p) - \Phi(t_a(u), \nu_p) = \Psi(t_a(u+1)) - \Psi(t_a(u)) \ [2\pi] \\
&= 2\pi f_l \Delta t_a(u) + 2n\pi \\
&= 2\pi(f_l - \nu_p)\Delta t_a(u) + 2\pi\nu_p \Delta t_a(u) + 2n\pi
\end{aligned}$$

où n est un entier relatif et $\Delta t_a(u) = t_a(u+1) - t_a(u)$. En tenant compte de $|\nu_p - f_l(t_a)| \leq f_c$, l'équation précédente conduit, si la condition 1.21 ci-dessous est vérifiée

$$f_c \Delta t_a(u) < 1/2 \tag{1.21}$$

à l'inégalité

$$|\Delta\Phi_p - 2\pi\nu_p \Delta t_a(u) - 2n\pi| < \pi \ ,$$

or il n'existe qu'une et une seule valeur de n qui vérifie cette propriété. On a ainsi levé l'indétermination. En résumé, on peut donc obtenir la valeur de fréquence instantanée *dans chaque canal de TFCT* par l'algorithme suivant :

1. Calcul de la TFCT à deux instants d'analyse successifs, qui donne $\Delta\Phi_p$ pour chaque canal ($p = 0, \ldots, N-1$),
2. Pour chaque canal, on cherche la valeur $Q(n_0)$ de $Q(n) = \Delta\Phi_p - 2\pi\nu_p\Delta t_a - 2n\pi$ telle que $|Q(n_0)| < \pi$,
3. on en déduit les fréquences instantanées par $f_l = \nu_p + \dfrac{Q(n_0)}{2\pi\Delta t_a}$.

Interprétation : l'inégalité 1.21 conduit à une condition de recouvrement minimal entre les fenêtres d'analyse. En effet, si on prend par exemple une fenêtre de Hann, pour laquelle on peut estimer $f_c = 2/L$ où L est la longueur de la fenêtre, elle devient

$$\Delta t_a < \frac{L}{4}$$

ce qui traduit un recouvrement minimal de 75% en analyse.

1.5.2 Distorsion temporelle

Une fois déduites les fréquences instantanées dans chaque canal [1], la distorsion temporelle du signal peut-être envisagée. Les phases instantanées en particulier, peuvent être "déroulées" de manière à synchroniser la TFCT modifiée sur les instants de synthèse. On obtient alors l'algorithme de modification qui suit, en supposant calculées pour l'index u la TFCT d'analyse $\tilde{X}(t_a(u), \nu_p)$ et la TFCT de synthèse $\tilde{Y}(t_s(u), \nu_p)$ et étant donnée la loi de distorsion temporelle $T(t)$:

1. calcul de la TFCT à l'instant $t_a(u+1)$ et déduction de la fréquence instantanée $f_k(t_a(u))$ dans chaque canal,
2. calcul du nouvel instant de synthèse $t_s(u) = T(t_a(u))$; dans la pratique on prend la partie entière de ce nouvel instant,

1. ce faisant, on suppose qu'il existe une et une seul composante par canal et par suite, on peut confondre les index p(canal de TFCT) et k(composantes).

3. itération de la phase instantanée de synthèse

$$\Phi_s(t_s(u+1), \nu_p) = \Phi_s(t_s(u), \nu_p) + 2\pi f_p(t_a(u))(t_s(u+1) - t_s(u))$$

4. calcul de la TFCT de synthèse pour l'index $u+1$ selon

$$\tilde{Y}(t_s(u+1), \nu_p) = A_p(t_a(u+1)) \exp j\Phi_s(t_s(u+1), \nu_p)$$

1.5.3 Modification de hauteur

La modification de hauteur, ou, plus généralement, d'échelle fréquentielle, s'obtient soit par rééchantillonnage temporel soit par rééchantillonnage spectral.

Rééchantillonnage temporel. Cette méthode s'appuie sur les propriétés de réciprocité telles que vues au paragraphe 1.3.3.

Dans le cas d'un taux de compression fréquentiel $\alpha(t) = \alpha_0$ constant, on obtient la modification souhaitée par

1. un étirement temporel d'un facteur α_0,
2. une relecture à la fréquence d'échantillonnage $\alpha_0 F_e$

où F_e est la fréquence d'échantillonnage d'origine. Cette technique est équivalente à effectuer un rééchantillonnage de facteur $\alpha_0 = F_e'/Fe$ et de lire à F_e. Dans ce dernier cas, il est toutefois à noter que le support temporel est divisé par α_0.

Une extension de cette technique par rééchantillonnage peut être appliquée pour obtenir des taux de compression $\alpha(t)$ variables dans le temps. On utilise la méthode canonique de rééchantillonnage des signaux numériques en approchant α par une fraction rationnelle à chaque instant d'analyse $\alpha(t_a) = L(t_a)/M(t_a)$ et en réalisant la chaîne de traitement de la figure 1.3 ou $H(z)$ est un filtre passe-bas de fréquence de coupure $\nu_c = \min(1/2L, 1/2M)$. On peut

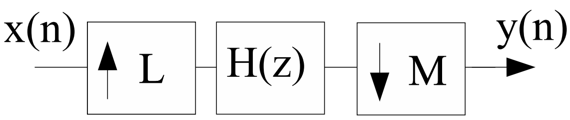

Figure 1.3 – Chaîne canonique de rééchantillonnage d'un facteur L/M

donc appliquer ce traitement à chaque trame du signal et utiliser des instants de synthèse et d'analyse synchronisés, $t_a(u) = t_s(u)$. Il est à noter que dans ce cas, il n'est pas fait usage du vocodeur de phase. Cette méthode peut s'avérer assez lourde car elle requiert le calcul d'un nouveau filtre d'interpolation H à chaque pas d'analyse.

Rééchantillonnage spectral. Le vocodeur de phase permet une solution moins gourmande que le rééchantillonnage dans les cas de taux de compression variable, en effectuant un rééchantillonnage dans le domaine fréquentiel. Ce rééchantillonnage fréquentiel est effectué par interpolation linéaire du spectre à court terme d'analyse, soit

$$\begin{aligned} q &= \lfloor p/\alpha(t_a(u)) \rfloor \\ \mu_p &= p/\alpha(t_a(u)) - q \\ \tilde{Y}(t_s(u), \nu_q) &= (1-\mu_p)\tilde{X}(t_a(u), \nu_p) + \mu_p \tilde{X}(t_a(u), \nu_{p+1}) \end{aligned}$$

p et q sont les entiers naturels qui indexent les canaux de la TFCT, ils varient donc de 0 à $N-1$. On note que cette interpolation, si elle ne présente pas de difficulté pour des taux de compression supérieur à l'unité (la hauteur est accrue) nécessite en revanche une complétion du spectre en haute fréquence pour des taux inférieur à l'unité (le son synthétisé est plus grave). Une manière de réaliser cette complétion fut suggérée dans [22] et consiste simplement à recopier la partie basse fréquence du spectre pour la partie manquante. Cette copie spectrale présente un bon rendu pour des fréquences d'échantillonnage d'au moins 16 kHz. En effet dans ce cas, la complétion intervient pour des fréquences élevées où le son a des caractéristiques principalement non voisées.

Enfin, pour réaliser la modification, il faut tenir compte de la modification locale de l'échelle temporelle occasionnée par la modification fréquentielle. En effet les phases de la TFCT de synthèse $\Phi_s(t_s(u), \nu_p) = \Phi_a(t_a(u), \nu_p)$ sont maintenant synchronisées sur des instants de synthèse différents des instants d'analyse :

$$\begin{aligned}\Phi_s(t_s(u+1), \nu_p) &= \Phi_s(t_s(u), \nu_p) + 2\pi f_p(t_a(u))\Delta t_a(u) \\ &= \Phi_s(t_s(u), \nu_p) + 2\pi\alpha(t_a(u)) f_p(t_a(u))\Delta t_s(u)\end{aligned}$$

On voit donc que la durée locale $\Delta t_a(u)$ d'analyse a été divisée par α en synthèse. Soit $\Delta t_s(u) = \Delta t_a(u)/\alpha(t_a(u))$. Ceci correspond à une distortion temporelle virtuelle

$$T(t) = \int_0^t \alpha(w)^{-1} dw$$

Pour réaliser la modification d'échelle temporelle, il faut donc appliquer *in fine* une distorsion temporelle compensatoire $D(t) = T^{-1}(t)$.

Remarque concernant le traitement de la voix parlée ou chantée. Dans le cas des modifications de hauteur de la voix parlée ou chantée, une transposition directe du signal conduit à l'effet "Donald Duck". En effet, la transposition du spectre global conduit à une transposition de son enveloppe et donc des formants. Le timbre est alors sévèrement modifié et la voix acquiert une caractéristique nasillarde évoquant le cancanement du canard. Cet effet est également produit par la modification de l'impédance caractéristique du milieu occasionnée par le mélange respiré par les plongeurs. Une solution pour palier ce défaut consiste à estimer l'enveloppe du spectre avant le traitement (par LPC ou modélisation directe [5]). Le traitement est ensuite appliqué au signal de source (résiduel LPC par exemple) puis le résultat obtenu est filtré pour retrouver l'enveloppe spectrale originelle, inchangée.

1.6 Méthode temporelle "pitch synchrone"

Cette méthode (dénommée TD-PSOLA, Time-Domain Pitch-Synchronous Overlap-Add) suppose que l'on traite un signal de parole dont on connaît la période.

L'idée [16] est fondée sur l'hypothèse que le signal de parole est constitué d'impulsions glottales filtrées par le conduit vocal. On observe ainsi une succession de réponses impulsionnelles, positionnées à des temps multiples de la période (hypothèse du peigne temporel convolué avec la réponse impulsionnelle du conduit vocal).

On définit alors des "marques d'analyses" synchrones de la fréquence fondamentale pour les parties voisée, positionnées sur la forme d'onde à chaque période. Les modifications d'échelles sont alors effectuées de la façon suivante :

1.6.1 Modification de l'échelle temporelle.

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

Les signaux à court-terme situés autour de chaque marque d'analyse sont alors extraits (par l'utilisation d'une fenêtre temporelle-par exemple de type hanning- de durée égale à deux périodes et centrée sur la marque d'analyse) et 'recopiés' autour des marques de synthèse correspondantes et le signal modifié est obtenu par une simple méthode d'OverLap-Add (addition-recouvrement). La figure 1.4 illustre le principe de cette méthode pour un taux d'étirement temporel local de 1.5.

On voit que deux périodes du signal original ont donné naissance à trois périodes dans le signal modifié, ce qui correspond bien à un étirement temporel mais la durée de la période n'est pas modifiée (l'écartement des marques de synthèse est le même que celui des marques d'analyse), la fréquence fondamentale du signal est conservée. La figure 1.6 donne un exemple d'application à la phrase 'il s'est' dont l'original est donné figure (1.5). On remarque la partie non-voisée au centre de la fenêtre (le son 's'), séparant les deux parties voisées /i/ et /e/.

Bertrand David bertrand.david@telecom-paristech.fr

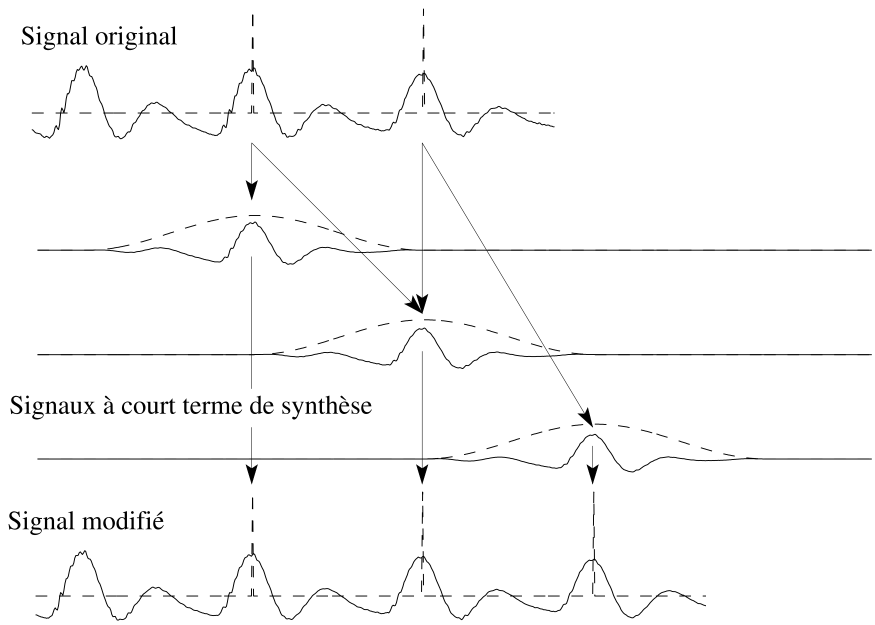

FIGURE 1.4 – Modification de la durée du signal par la méthode TD-PSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des deux signaux à court-terme centrés autour des deux premières marques d'analyse. En bas, signal modifié.

1.6.2 Modification de l'échelle fréquentielle.

Si l'on est capable de positionner dans le signal les marques d'analyse exactement sur le début de chaque onde glottale (réponse impulsionnelle du conduit vocal se produisant à chaque fermeture glottale), on conçoit que diminuer (resp. augmenter) l'intervalle de temps séparant deux marques d'analyse consécutives va permettre d'augmenter (resp. de diminuer) la fréquence du fondamental, sans que les formants soient modifiés (la réponse impulsionnelle n'est pas modifiée, en particulier sa décroissance temporelle et ses fréquences de résonance–les formants).
On est ainsi conduit à définir des marques de synthèse correspondant à la valeur modifiée du fondamental, et à les associer aux marques d'analyse comme précédemment. Puisque les marques de synthèse sont plus serrées (élévation du fondamental) ou écartées (abaissement du fondamental) que dans le signal original, il faut pour conserver la durée du signal dupliquer ou éliminer certaines marques. La figure 1.7 illustre le principe de cette méthode.

On constate que les marques de synthèse étant plus écartées que les marques d'analyse, la période du signal est allongée. Pour éviter une élongation du signal, il est nécessaire d'éliminer périodiquement certains signaux à court-terme.

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

La méthode décrite ci-dessus est appliquée principalement à la parole, et réalise des modifications de très bonne qualité. Par sa simplicité, elle peut faire l'objet d'une implémentation temps réel. En revanche, son application à des sons plus complexes, ou dénués de "pitch" (cas de la musique en général) pose de sérieux problèmes.
Les modifications de fondamental sont cependant très sensibles à la position des marques d'analyse. Pour rendre la méthode plus robuste, les modifications de l'échelle fréquentielle peuvent être réalisées dans le domaine des

Bertrand David `bertrand.david@telecom-paristech.fr`

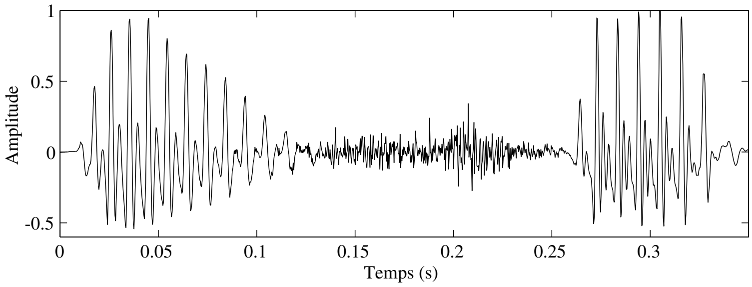

FIGURE 1.5 – Original : "il s'est".

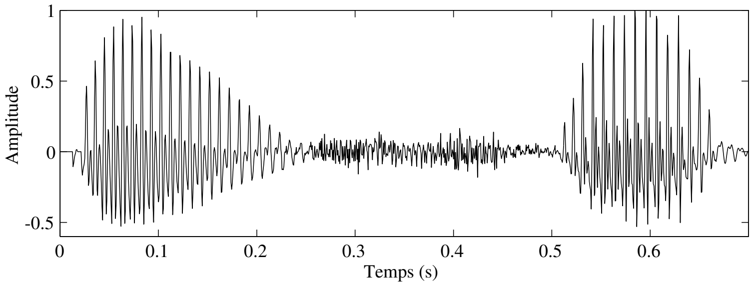

FIGURE 1.6 – Signal étiré d'un facteur 2.

fréquences (méthode FD-PSOLA) [16, 17].

Pour d'autres méthodes basées sur des idées très similaires, on pourra se référer à [7, 14, 21, 26]

1.6.3 La technique de la mémoire circulaire

La technique de la mémoire circulaire est la plus simple et la plus ancienne des techniques de modification de l'échelle temporelle et fréquentielle [2]. Il s'agit également d'une méthode fonctionnant dans le domaine temporel.

1.6.3.1 L'origine analogique

Cette technique dérive d'un système analogique proposée dans les années 50 [6]. Elle consiste à utiliser un magnétophone muni d'une tête rotative. La bande en boucle fermée s'enroule sur la moitié du cylindre (comme pour les magnétoscopes et les DAT) et défile à vitesse constante. Le cylindre est muni de deux têtes de lecture diamétralement opposées dont les signaux sont mélangés avec un gain identique. Il est possible de contrôler le sens de rotation et la vitesse du cylindre.

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

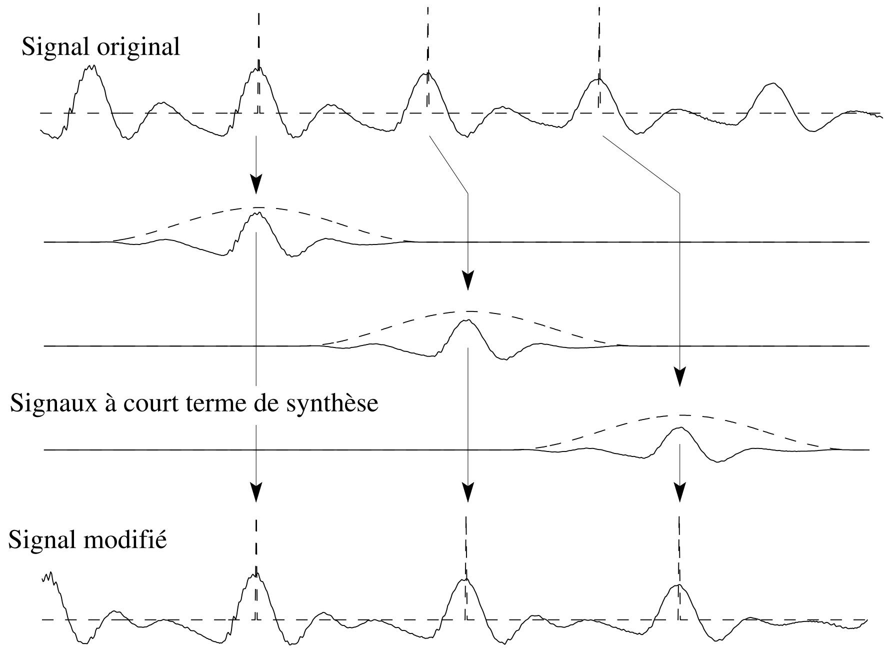

FIGURE 1.7 – Modification de la hauteur du signal par la méthode TD-PSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des trois premières marques d'analyse. En bas, signal modifié. L'écartement des marques de synthèse n'est pas identique à celui des marques d'analyse.

est contracté vers l'origine puisque la bande est lue à une vitesse moindre qu'elle n'est enregistrée. Dans ce cas, certaines portions du signal peuvent ne pas être lues du tout.
Le rapport de l'homothétie en fréquence s'exprime par :

$$\alpha = \frac{V_r}{V_a} = \frac{V_a + R\,\Omega_{cylindre}}{V_a}$$

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

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

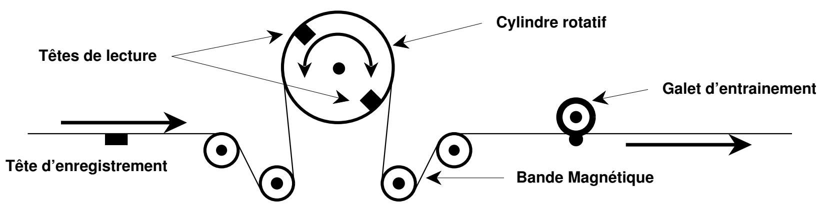

FIGURE 1.8 –

1.6.3.2 Implémentation numérique

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

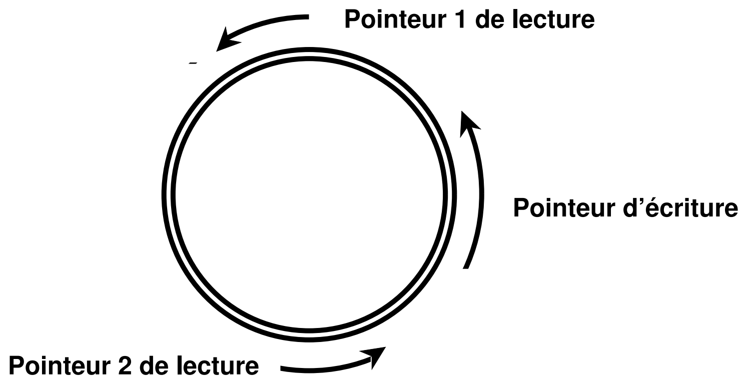

FIGURE 1.9 –

A chaque échantillon écrit dans la mémoire (toutes les ΔT secondes), on avance les pointeurs de lectures de $\alpha\Delta T$ secondes, où α est le taux de modification, puis on lit un échantillon dans la mémoire. En général (pour les valeurs de α non entières), on se retrouve entre deux échantillons, et comme dans le cas du "flanger", il est nécessaire de calculer la valeur du signal à cet instant. Ici aussi, une simple interpolation linéaire convient.
Ainsi, le signal est lu avec une fréquence d'échantillonnage différente de celle à laquelle il a été écrit, ce qui provoque une modification de l'échelle des fréquences de taux α. Un problème se pose lorsque le pointeur de lecture rattrape (lorsque $\alpha > 1$) ou est rattrapé (lorsque $\alpha < 1$) par le pointeur d'écriture. Comme dans l'équivalent analogique, la continuité est assurée par un mixage des deux pointeurs au moment où se produit la rencontre ("cross-fade") : l'échantillon lu par le pointeur de lecture courant (par exemple le pointeur 1) subit une pondération décroissante tandis que celui lu par l'autre pointeur de lecture (le pointeur 2) subit une pondération croissante. Finalement, le second pointeur devient le pointeur courant, et garde sa pondération maximale, jusqu'à ce que le pointeur d'écriture s'en rapproche.

Implémenté de cette façon, le pitch shifter a un comportement sensiblement équivalent à son homologue analogique (à ceci près qu'il est plus facilement paramétrable). Son implémentation en temps réel ne pose pas de problème particulier, puisqu'il ne réclame que très peu de calculs.
Il produit malheureusement un bruit artificiel qui provient du mixage périodique des deux pointeurs de lecture. Pour tenter d'améliorer la qualité obtenue, on cherche à mieux raccorder les signaux lus par les deux pointeurs de lecture, de façon un peu similaire à ce qui est fait dans les méthodes synchrones. On peut par exemple utiliser la fonction d'autocorrélation du signal pour déterminer l'endroit le plus adéquat pour le "cross-fading" [4, 10].

1.6.3.3 Modification de la durée par la technique de la mémoire circulaire

Comme son homologue analogique, la technique du buffer circulaire peut être également utilisée pour la modification de l'échelle temporelle : Si l'on dispose d'un pitch shifter à mémoire circulaire, et que l'on veut effectuer un "time scaling" de paramètre α, il suffit de changer la fréquence d'échantillonnage du signal d'un taux α, puis de le traiter par le pitch shifter. Ainsi, pour ralentir le signal 2 fois, il suffit de le sur-échantillonner 2 fois. Si l'on écoute le signal obtenu à la fréquence originale, il sera deux fois plus long, mais aussi une octave plus grave. Il suffit donc de l'écouter à la fréquence originale en intercalant un pitch shifter de taux $\alpha = 2$.

On se rend rapidement compte qu'il est plus simple de faire les deux opérations une seule fois : La technique consiste alors à répéter ou éliminer périodiquement des portions de signal pour en augmenter (ou diminuer) la durée. Vue sous cet angle, cette technique (qui prend alors le nom de 'splicing method') se rapproche d'une technique TD-PSOLA dans laquelle on ne connaîtrait pas la valeur de la fréquence fondamentale. Les artefacts inhérents à cette méthode, qui proviennent des ruptures de la périodicité du signal lors des répétitions ou des éliminations peuvent être considérablement réduits par l'utilisation de méthodes basées sur l'autocorrélation du signal pour optimiser la longueur et l'emplacement des portions de signal à dupliquer ou à détruire [4, 10, 11, 19, 24, 25].

Bertrand David bertrand.david@telecom-paristech.fr

Voir page 19

Bibliographie

[1] J. Allen. Overwiew of text-to-speech systems. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, chapter 23, pages 741–790. Marcel Dekker, 1991.

[2] J. Benson. *Audio Engineering Handbook*. mcGraw-Hill, New York, 1988.

[3] O. Cappé, J. Laroche, and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. *IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, 1995.

[4] J. Dattorro. Using digital signal processor chips in a stereo audio time compressor/expander. *Proc. 83rd AES Convention, New York*, Oct 1987. preprint 2500 (M-6).

[5] A. El-Jaroudi and J. Makhoul. Discrete all pole modeling. *IEEE Trans. Acoust., Speech, Signal Processing*, 39(2) :411–423, Fev 1991.

[6] G. Fairbanks, W.L. Everitt, and R.P. Jaeger. Method for time or frequency compression-expansion of speech. *IEEE Trans. Audio Electroacoust.*, AU-2 :7–12, Jan 1954.

[7] E. Hardam. High quality time scale modification of speech signals using fast synchronized overlap add algorithms. *Proc. IEEE ICASSP-90*, pages 409–412, 1990.

[8] D.L. Jones and T.W. Parks. On the generation and combination of grains for music synthesis. *Computer Music J.*, 12(2) :27–34, Summer 1988.

[9] M. Kahrs and K. Brandenbourg. *Applications of Digital Signal Processing to Audio and Acoustics*. Kluwer Academic Press, Dortrecht, Netherland, 1998.

[10] J. Laroche. Autocorrelation method for high quality time/pitch scaling. *IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, 1993.

[11] F. Lee. Time compression and expansion of speech by the sampling method. *J. Audio Eng. Soc.*, 20(9) :738–742, 1972.

[12] J. Makhoul. Linear prediction : A tutorial review. *Proc. IEEE*, 63(11) :1380–1418, Nov 1975.

[13] J. Makhoul and A. El-Jaroudi. Time scale modification in medium to low rate speech coding. *Proc. IEEE ICASSP-86*, pages 1705–1708, 1986.

[14] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(2) :121–133, 1979.

[15] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4) :744–754, Aug 1986.

[16] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6) :453–467, Dec 1990.

[17] E. Moulines and J. Laroche. Non parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16 :175–205, Feb 1995.

[18] M. R. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24(3) :243–248, Jun 1976.

[19] S. Roucos and A. M. Wilgus. High quality time-scale modification of speech. *Proc. IEEE ICASSP-85, Tampa*, pages 493–496, Apr 1985.

[20] M.R. Schroeder, J.L Flanagan, and E.A. Lundry. Bandwidth compression of speech by analytic-signal rooting. *Proc. IEEE*, 55 :396–401, Mar 1967.

[21] R. Scott and S. Gerber. Pitch-synchronous time-compression of speech. *Proceedings of the Conference for Speech Communication Processing*, pages 63–65, Apr 1972.

[22] S. Seneff. System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24 :358–365, 1982.

[23] X. Serra and J. Smith. Spectral modeling synthesis : A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music J.*, 14(4) :12–24, Winter 1990.

[24] B. Sylvestre and P. Kabal. Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation. *Proc. IEEE ICASSP-92*, pages 81–84, 1992.

[25] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. *Proc. IEEE ICASSP-93, Minneapolis*, pages 554–557, Apr 1993.

[26] J.L. Wayman and D.L. Wilson. Some improvements on the sychronized-overlap-add method of time scale modification for use in real-time speech compression and noise filtering. *IEEE Trans. Acoust., Speech, Signal Processing*, 36(1) :139–140, Jan 1988.

Bertrand David `bertrand.david@telecom-paristech.fr`

Bertrand David **bertrand.david@telecom-paristech.fr**