Mini-projet: Traitement des signaux de parole.

L'objectif de ce mini-projet est d'implémenter, de manière simplifiée, des modèles utilisés pour traiter les signaux de parole en téléphonie cellulaire [1].

1 Modèle de parole

Les signaux de parole sont des signaux aléatoires non stationnaires échantillonnés à haute fréquence. En découpant ces signaux à l'aide d'une fenêtre temporelle de petite taille, leur étude peut cependant se ramener à celle d'une séquence de signaux stationnaires. Parmi ces signaux stationnaires, on peut de façon schématique distinguer deux types de signaux. La voix émet en effet deux types de sons, dit *voisés* et *non-voisés* :

- Les sons voisés sont obtenus à partir de vibrations générées par les cordes vocales, qui sont ensuite transformées en un son de parole en passant dans le larynx.
- Les sons non voisés ne font à l'inverse pas intervenir les cordes vocales et sont produits à partir d'un bruit blanc transformé au cours de son passage dans le larynx.

Pour chaque type de son, des modèles aléatoires paramétriques peuvent être établis afin d'encoder et de resynthétiser les signaux stationnaires observés pour chaque fenêtre du signal.

1.1 Son voisés

Dans le cas d'un son voisé, les cordes vocales émettent un signal d'excitation qu'on peut modéliser par un train d'ondes émises avec une période T appelée *pitch* du signal :

$$f(t) = \sum_{m\in\mathbb{Z}} g(t - mT) = g * e(t),$$

où la fonction e correspond à un train d'impulsions de Dirac :

$$e(t) = \sum_{m\in\mathbb{Z}} \delta(t - nT).$$

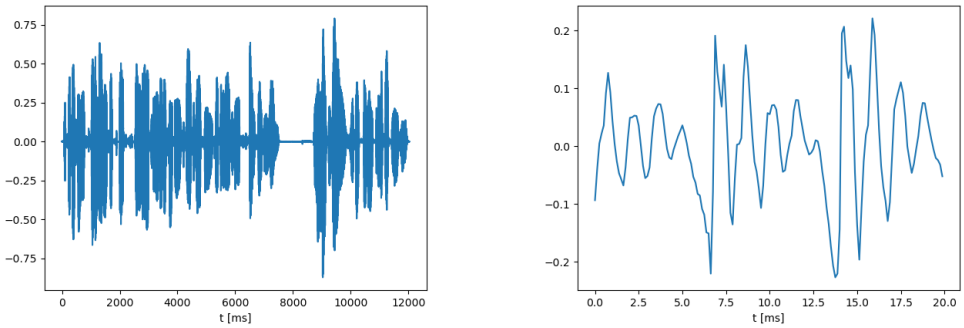

FIGURE 1 – Signal de parole non-stationnaire (gauche) et portion de signal sur une fenêtre temporelle de courte durée (droite). La portion de signal peut être localement considérée comme stationnaire.

Le support de la fonction g est en général de taille très inférieure au pitch T. Si on se concentre sur une petite portion temporelle du signal qu'on isole par une opération de fenêtrage en multipliant ce dernier par une fenêtre w, le pitch T peut être considéré comme constant.

La formation du son de parole peut être modélisée par l'application d'un filtre linéaire H à un signal d'excitation f donné par

$$f[n] = Ge[n],$$

G étant un terme de gain. Ce filtre agrège les effets du fenêtrage, de la forme de l'onde élémentaire g et du passage dans le larynx, et est caractérisé par ses coefficients $(a_1, \dots a_p)$. Il transforme le signal d'excitation f en un signal de sortie s via l'équation aux différences

$$s[n] = w[n]f[n] + \sum_{k=1}^{p} a_k s[n-k].$$

Sur la fenêtre temporelle de durée réduite sélectionnée, les coefficients du filtre peuvent être considérés comme constants.

1.2 Son non voisés

Dans le cas d'un son non voisé, le signal d'excitation f peut être décrit par un bruit blanc gaussien de variance σ^2. Comme pour les sons voisés, le signal de parole est construit par le passage de ce son dans le larynx, modélisé par le filtre linéaire

$$s[n] = w[n]f[n] + \sum_{k=1}^{p} a_k s[n-k].$$

2 Encodage des signaux de parole

En pratique, il est coûteux de faire transiter tel quel un signal de parole sur un canal de transmission. On cherchera donc généralement à compresser en temps réel le signal de parole pour réduire la taille des données à transmettre. Une méthode de compression couramment utilisée consiste à découper le signal en segments temporels très courts à l'aide d'une opération de fenêtrage, puis à utiliser le modèle de parole décrit dans la section précédente afin d'estimer pour chacun des segments les coefficients $(a_k)_{1\leq k\leq p}$ du filtre. C'est la valeur de ces quantités qui est transmise, un algorithme de décodage s'attachant ensuite à reconstruire le signal de parole original à partir de ces paramètres et du modèle de parole. Nous nous concentrons dans cette section sur l'implémentation de la méthode d'encodage du signal.

2.1 Fenêtrage du signal

Afin de fenêtrer le signal, on multiplie ce dernier par une fenêtre de largeur fixe. Dans ce mini-projet, on considérera pour effectuer le fenêtrage une fenêtre de Hamming de largeur T définie par :

$$w(t) = \begin{cases} 0.54 - 0.46\cos\frac{2\pi t}{T} & \text{si } 0 \leq t \leq T \\ 0 & \text{sinon.} \end{cases}$$

Question 2.1.1. *Le signal de parole que nous cherchons à traiter dans ce mini-projet est échantillonné avec une fréquence de* 24 *kHz. On peut dans un premier temps ré-échantillonner ce signal à une fréquence de* 8 *kHz. On lui applique ensuite une opération de fenêtrage à partir d'une fenêtre de Hamming de largeur* $T = 20ms$*, en faisant en sorte que deux fenêtres successives se recouvrent sur un intervalle de temps égal à la moitié de la largeur de la fenêtre, soit* 10 *ms dans le cas présent. Implémenter l'opération de fenêtrage dans la fonction* ***block_decomposition****. Montrer qu'il est possible de reconstruire le signal original à partir des segments fenêtrés et implémenter l'opération de reconstruction dans la fonction* ***block_reconstruction****. On pourra corriger les effets aux bords en ajoutant des* 0 *de chaque côté du signal d'origine.*

2.2 Encodage des segments

Une fois le fenêtrage du signal effectué, nous disposons pour chaque segment d'un signal fenêtré $s[n], n = 0, \ldots, N-1$. Afin de déterminer les coefficients du filtre, nous utilisons une approche, appelée *prédiction linéaire* dans la littérature, qui consiste à prédire la valeur du signal à l'instant n à partir des observations passées du signal. La prédiction $\tilde{s}[n]$ de la valeur du signal à un instant n est donnée en fonction des valeurs observées du signal avant l'instant considéré par :

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k].$$

L'erreur d'estimation à un instant n est par conséquent

$$\epsilon[n] = \|s[n] - \tilde{s}[n]\|^2 = \|s[n] - \sum_{k=1}^{p} \alpha_k s[n-k]\|^2$$

En pratique, on fixe les coefficients $(\alpha_k)_{1\leq k\leq p}$ de manière à minimiser l'erreur d'estimation moyenne $\sum_{i=1}^{N-1} \epsilon[i]$. On peut alors montrer que les coefficients $(\alpha_k)_{1\leq k\leq p}$ sont solution de l'équation matricielle

$$\begin{pmatrix} r_s[0] & r_s[1] & \cdots & r_s[p-1] \\ r_s[1] & r_s[0] & \cdots & r_s[p-2] \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ r_s[p-1] & \cdot & \cdots & r_s[0] \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} r_s[1] \\ r_s[2] \\ \vdots \\ \vdots \\ r_s[p] \end{pmatrix},$$

où r_s est l'autocovariance empirique du signal s.

Question 2.2.1. *Implémenter l'estimation des coefficients du filtre pour un segment de parole dans la fonction* ***lpc_encode****. On notera que la matrice qui intervient est une matrice de Toeplitz, et on pourra donc utiliser la méthode* ***solve_toeplitz*** *de la librairie scipy pour résoudre le système. Calculer enfin la prédiction* $(\tilde{s}[n], 0 \leq n \leq N-1)$ *du signal obtenue avec le filtre.*

Supposons que les coefficients $(\alpha_k)_{1\leq k\leq p}$ estimés lors de l'encodage vérifient $\alpha_k \simeq a_k$ pour tout $k = 1, \ldots, p$. Alors, en utilisant le modèle de parole, on vérifie que :

$$s[n] \simeq w[n]f[n] + \sum_{k=1}^{p} \alpha_k s[n-k]$$

de sorte que

$$s[n] - \tilde{s}[n] \simeq w[n]f[n].$$

Question 2.2.2. *En utilisant cette dernière formule, implémenter dans la fonction* ***lpc_decode*** *la reconstruction du signal* $(s[n], 0 \leq n \leq N-1)$ *à partir du résidu* $w[n]f[n] = s[n] - \tilde{s}[n]$ *et des coefficients du filtre.*

En pratique, pour compresser le signal de parole, on essaiera d'approximer de manière très simple le résidu $\{w[n]f[n]\}_{n=0,\ldots,N-1}$. D'après le modèle de parole, ce résidu correspond à un train d'impulsions de période égale au pitch T dans le cas d'un segment de son voisé, ou à un bruit blanc gaussien dans le cas d'un son non voisé. Pour chaque segment du signal, il nous faut donc déterminer si ce dernier est voisé, et la valeur du pitch le cas échéant.

2.3 Estimation du pitch

La détermination du pitch est un problème complexe, pour lequel des méthodes relativement élaborées ont été développées. Nous étudions dans ce mini-projet une approche basée sur le *cepstre* du signal, une notion détaillée dans le paragraphe qui suit.

2.3.1 Notion de cepstre

Soit s un signal discret d'énergie finie. Rappelons que la transformée de Fourier en temps discret de s est définie pour tout $\omega \in \mathbb{R}$ par

$$\widehat{s}(\omega) = \sum_{k \in \mathbb{Z}} s[n] e^{-i\omega n}.$$

On définit le *cepstre* S de s comme la TFTD inverse de la quantité $\log |\widehat{s}|(\omega)$:

$$\forall t \in \mathbb{R}, S(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |\widehat{s}(\omega)| e^{i\omega t}.$$

Du fait de l'application du logarithme, la TFTD inverse ne transforme pas réellement une représentation fréquentielle en une représentation temporelle. Plutôt que de parler de temps, on utilise donc en pratique le terme de *quéfrence* pour désigner l'espace de représentation du cepstre.

2.3.2 Application à la détermination du pitch

Nous avons vu que sur une courte fenêtre temporelle, un signal de parole voisé s peut être modélisé comme l'application d'un filtre linéaire de réponse impulsionnelle h à un signal d'excitation e de la forme :

$$e[n] = \sum_{k=0}^{M-1} \delta[n - kT],$$

T étant le pitch du signal. La TFTD du signal de parole est donc donnée par le produit :

$$\widehat{s}(\omega) = \widehat{a}(\omega)\widehat{e(\omega)}.$$

En règle général, la fonction $\omega \mapsto \widehat{a}(\omega)$ varie lentement et constitue ce qu'on appelle l'enveloppe du signal. A l'inverse, la fonction $\omega \mapsto \widehat{e}(\omega)$ varie beaucoup plus rapidement et vient donc en quelque sorte moduler l'enveloppe $\omega \mapsto \widehat{a}(\omega)$ du signal.

Les variations de $\omega \mapsto \log |\widehat{h}(\omega)|$ étant relativement lente, le support du cepstre de h va essentiellement être concentré sur les basses quéfrences. A l'inverse, le cepstre du signal d'excitation va se retrouver sur des quéfrences plus élevées : le cepstre permet de séparer les support quéfrenciels du signal d'excitation et de l'enveloppe.

On peut enfin montrer que le cepstre d'un train d'impulsion de période T est donné par la relation

$$E(t) = \sum_{q \in \mathbb{Z}} \frac{1}{q} \delta[t - qT].$$

En pratique, pour un segment correspondant à un son voisé, on sait que le pitch va se situer dans un intervalle de quéfrences bien précis. On sait en effet que la fréquence de pitch $1/T$ est comprise entre 50 Hz et 250 Hz pour une voix humaine, soit une fenêtre de quéfrences comprises entre 4 et 20ms. Ce pic n'apparaît pas dans le cas de segments non voisés. Pour déterminer si un segment de signal est voisé ou non, on commencera par extraire le maximum du cepstre sur l'intervalle > 4 ms. On comparera ensuite la valeur de ce maximum à la valeur moyenne du cepstre sur ce même intervalle. Si le ratio entre ces deux quantités est supérieur à un seuil fixé, de l'ordre de $2 - 5$, on considérera alors que le segment est voisé et l'estimation du pitch sera la quéfrence associée au pic. A l'inverse, si le ratio est inférieur au seuil, le segment sera considéré comme non voisé.

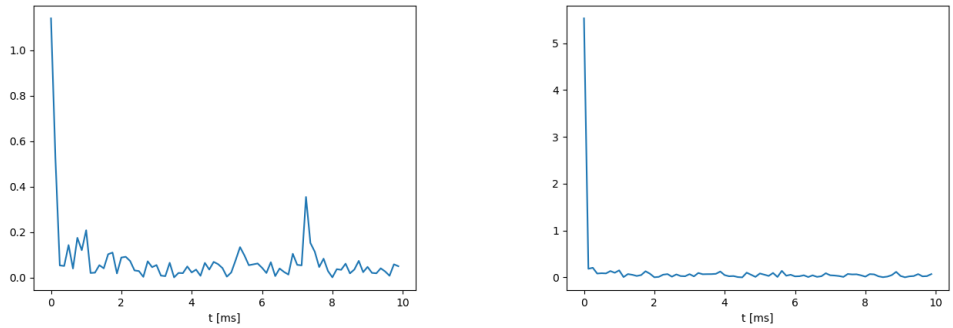

FIGURE 2 – Cepstre d'un segment voisé (gauche) et d'un segment non voisé (droite). On distingue nettement un pic à une quéfrence $\simeq 7$ ms dans le cas du signal voisé. A l'inverse, aucun pic significatif n'est identifiable dans le cas du segment non voisé.

Question 2.3.1. *Implémenter l'estimation du pitch par le calcul du cepstre en complétant les fonctions* ***compute_cepstrum*** *et* ***cepstrum_pitch_detection****.*

3 Décodage

Afin de resynthétiser le segment de parole, le principe est de générer un signal d'excitation $\tilde{f}$ qui approxime le signal d'excitation original f, avant de filtrer ce signal par le filtre caractérisé lors de l'encodage du signal. Deux cas se présentent :

— Si lors de la détection du pitch, le segment a été classifié comme non voisé, alors on génère un signal d'excitation $\tilde{f}$ en considérant un bruit blanc d'écart-type σ égal à l'écart-type du résidu.

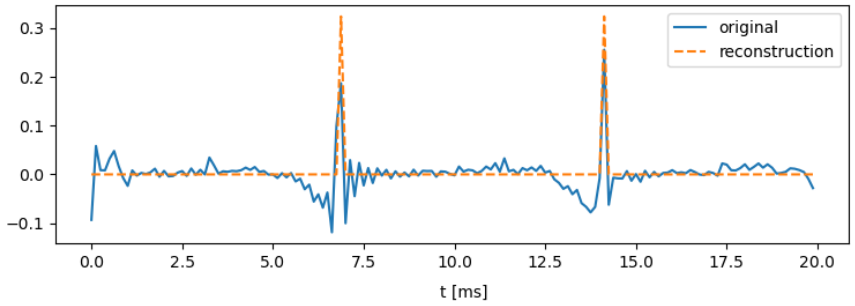

FIGURE 3 – Approximation d'un signal d'excitation par un train d'impulsions.

— Si au contraire le segment a été classifié comme voisé, on génère un signal d'excitation $\tilde{f}$ en fabriquant un train d'impulsion dont la période correspond au pitch estimé, qu'on normalise de sorte que l'écart-type de $\tilde{f}$ soit égale à σ.

Question 3.0.1. *Reconstruire le signal de parole en resynthétisant un signal d'excitation pour chaque segment et en utilisant la fonction* ***lpc_decode*** *et en additionnant les segments re-synthétisés. Tester différentes valeurs de p et de taille de fenêtre et commenter les résultats obtenus.*

Références

[1] Lawrence R Rabiner, Ronald W Schafer, et al. Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, 1(1–2) :1–194, 2007.