MINES Paris
http://moodle.psl.eu *2024-2025*

Ce mini-projet, à effectuer en binôme, fera l'objet d'un rapport incluant notamment équations et graphiques obtenus par des simulations sous Python. La forme de ce rapport est laissée libre (pdf, notebook, version papier...). Des fichiers python avec le prototype des différentes fonctions à implémenter sont disponibles pour le mini-projet. Le rendu s'effectuera sur Moodle. La date limite de rendu est le 25 Avril 2025, aucun rendu ne sera possible après cela.

Codage prédictif linéaire (LPC)

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 Encodage du signal

Les signaux de parole sont des signaux non stationnaires échantillonnés à haute fréquence. 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.

Les méthodes de compression couramment utilisées découpent le signal en segments temporels très courts à l'aide d'une opération de fenêtrage, puis utilisent un modèle de parole afin d'estimer pour chaque segment les coefficients d'un filtre permettant de reconstruire le signal de parole à partir d'un signal d'excitation. Ce sont les valeurs de ces coefficients et les caractéristiques du signal d'excitation qui sont transmises en pratique. Un algorithme de décodage s'attache ensuite à reconstruire le signal de parole original à partir de ces paramètres.

1.1 Fenêtrage du signal

En découpant un signal de parole à l'aide d'une fenêtre temporelle de petite taille, son étude peut se ramener à celle d'une séquence de signaux stationnaires (cf. figure 1). La première étape d'un algorithme de traitement de la parole consiste à effectuer un *fenêtrage* du signal. Pour ce faire, on multiplie le signal par une fenêtre de largeur fixe. Dans ce projet, on considérera pour effectuer le fenêtrage une fenêtre *de Hamming* de largeur T définie par :

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

Le signal de parole que nous cherchons à traiter ici est échantillonné avec une fréquence de 24 kHz. On peut tout d'abord ré-échantillonner ce signal à une fréquence

de 8 kHz et lui appliquer ensuite une opération de fenêtrage à partir d'une fenêtre de Hamming de largeur $T = 20$ms, 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.

Question 1: Implémenter l'opération de fenêtrage dans *block_ decomposition*. On pourra corriger les effets aux bords en ajoutant des 0 de chaque côté du signal d'origine.

Question 2: Montrer qu'il est possible de reconstruire le signal original à l'identique à partir des segments fenêtrés et implémenter l'opération de reconstruction dans la fonction *block_ reconstruction*.

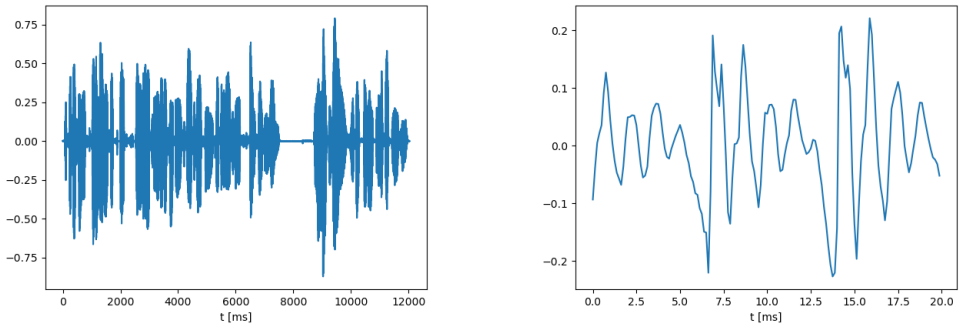

FIGURE 1 – Signal de parole non-stationnaire (gauche) et portion de signal sur une fenêtre temporelle de courte durée (droite). La portion de signal peut être localement considérée comme stationnaire : on voit en effet qu'il présente des motifs se répétant de manière périodique.

1.2 Coefficients du filtre

La formation d'un son de parole peut être modélisée par l'application d'un filtre linéaire H à un signal d'excitation $(f[n])_{n=0,\dots,N-1}$ produit par le larynx ou la gorge. Le filtre H est caractérisé par ses coefficients $(a_1, \dots a_p)$ et transforme le signal d'excitation f en un signal de sortie s via l'équation aux différences

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

où w correspond à la fenêtre utilisée pour découper le signal. Sur la fenêtre temporelle de durée réduite sélectionnée, les coefficients du filtre peuvent être considérés comme constants.

Une fois le fenêtrage du signal effectué, nous disposons pour chaque segment d'un signal fenêtré $s[n], n = 0, \dots, N-1$. 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

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

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

Question 3: Montrer que les coefficients $(\alpha_k)_{1\leq k\leq p}$ sont solutions 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, définie pour un signal de taille N par :

$$r_s[p] = \frac{1}{N} \sum_{i=1}^{N-p} s[k]s[k+p] \tag{1}$$

Question 4: 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. Déterminer la prédiction $(\tilde{s}[n], 0 \leq n \leq N-1)$ du signal obtenue avec le filtre et les résidus correspondants.

2 Décodage du signal

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 5: 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 des résidus $w[n]f[n] = s[n] - \tilde{s}[n]$ et des coefficients du filtre.

2.1 Encodage des résidus

En pratique, on ne fait pas transiter les résidus directement dans le canal de transmission, ce qui s'avèrerait trop coûteux, mais on re-synthétise un signal d'excitation à partir d'un modèle de parole. Les segments de paroles peuvent schématiquement être classifiés en deux types de signaux dits *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 lors du passage du son dans le larynx. Dans le cas d'un son voisé, les cordes vocales émettent un signal d'excitation qu'on peut modéliser par un train d'impulsions de Dirac émises avec une période T appelée *pitch* du signal :

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

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.

— Les sons non voisés ne font à l'inverse pas intervenir les cordes vocales, et sont produits à partir d'un bruit blanc transformé en son de parole au cours de son passage dans le larynx. 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

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

Question 6: Implémenter la reconstruction du signal $(s[n], 0 \leq n \leq N-1)$ à partir d'un signal d'excitation obtenu en générant un bruit blanc gaussien de variance σ^2 égale à la variance du résidu.

2.2 Estimation du pitch (en bonus)

Afin d'estimer le pitch pour un segment de parole, une méthode simple consiste à calculer la fonction de corrélation du segment avant multiplication par la fenêtre, puis à retenir pour le pitch le temps pour lequel la fonction de corrélation est maximale sur l'intervalle correspondant à la voix humaine, dont les fréquences fondamentales varient de 50 à 200 Hz environ. Dans le cas d'un signal voisé, le pic est facilement identifiable.

Question 7: (bonus) Implémenter l'estimation du pitch pour un segment dans la fonction, et déterminer si le segment est voisé ou non-voisé. Reconstruire le signal de parole en générant les signaux d'excitations appropriés selon la nature du segment.

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.