T.P. 7

Modélisation AR et prédiction

1 Introduction au traitement de la parole

1.1 Généralités

Un premier point concerne le choix de la fréquence d'échantillonnage. Dans le domaine de la téléphonie cela se traduit généralement par deux conditions : assurer l'intelligibilité du message et permettre l'identification du locuteur. Ces conditions conduisent à se contenter de la bande de fréquence $[0-4]$ kHz dite *bande téléphonique* et donc d'une fréquence d'échantillonnage de 8 kHz. Figure 7.1 nous avons reporté 2 000 valeurs, soit 0,25 s, d'un signal de parole échantillonné à une telle fréquence.

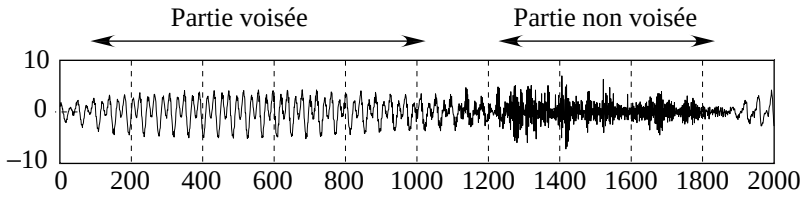

Figure 7.1: *Signal de parole échantillonné à* 8000 *Hz. Les abscisses donnent les numéros des échantillons.*

1.2 Typologie des sons

Le signal représenté figure 7.1 présente deux sections bien distinctes correspondant à deux types de sons :

- Les sons qui ont l'aspect d'une vibration harmonique et que l'on dit *voisés*. Les voyelles illustrent parfaitement ce type de son. Un exemple est fourni figure 7.1 dans la fenêtre allant des indices 0 à 1200.
- Les sons qui nous font davantage penser à un bruit et que l'on dit *non voisés*. Un exemple est fourni figure 7.1 dans la fenêtre allant des indices 1200 à 1800.

Les voyelles sont généralement de durée bien supérieure à celle des consonnes. Elles sont aisément reconnaissables par l'aspect très harmonique qu'elles présentent. Pour les consonnes, on distingue plus précisement :

- Les *nasales* /m/, /n/... pour lesquelles l'ensemble "cavité orale + pharynx" forme une cavité résonante fermée, l'air passant par les narines ;

– Les *fricatives non voisées* /f/, /s/, /ch/... qui sont produites par des turbulences dans la cavité buccale à partir d'un flux continu d'air. La cavité est divisée en deux sous cavités, celle du fond étant à l'origine de "zéros" dans la fonction de transfert ;

– Les *fricatives voisées* /v/, /z/... qui peuvent s'expliquer comme dans le cas précédent mais avec les cordes vocales qui vibrent ;

– Les *plosives voisées* /b/, /d/... qui sont des transitions provoquées par l'ouverture brusque de la cavité buccale préalablement mise en pression. Elles dépendent fortement des voyelles qui les accompagnent ;

– Les *plosives non voisées* /p/, /t/, etc.

La figure 7.2 illustre les cas des sons "ch" et "i".

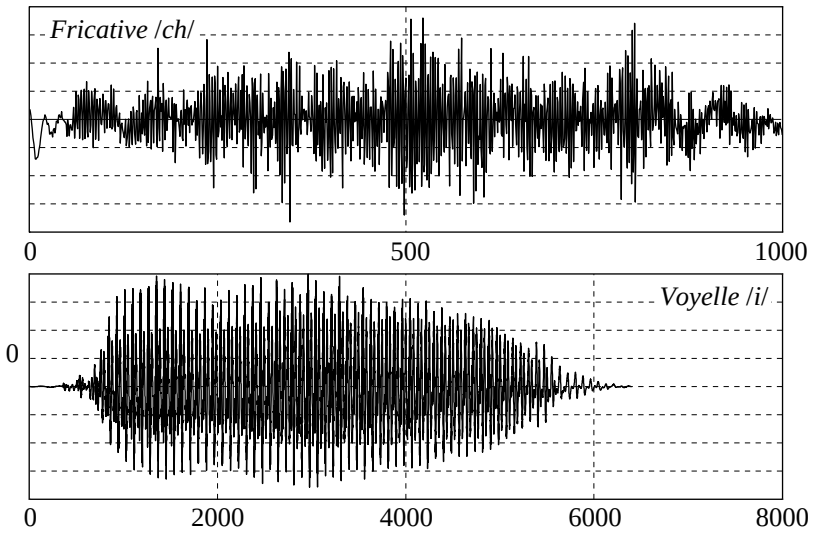

Figure 7.2: *Formes temporelles de signal de parole échantillonné à* 8000 *Hz : graphique du haut : son non voisé ; graphique du bas : son voisé.*

1.3 Modélisation AR de la production de la parole

La production des sons est un phénomène très complexe et difficile à modéliser. Elle est conditionnée par l'anatomie de l'appareil vocal représenté figure 7.3.

Un schéma fonctionnel est donné figure 7.4. Il représente de façon simplifiée le conduit vocal par une suite de cavités dont les formes évoluent au cours du temps et qui sont traversées par le flux d'air provenant des poumons.

Les études menées sur le conduit vocal montrent que les deux types de sons, voisés ou non voisés, peuvent être modélisés comme sorties d'un filtre linéaire tout-pôle de la forme $1/A(z)$, dont l'ordre est compris entre 10 et 20 et dont l'entrée est :

– un bruit blanc pour les sons non voisés,

– ou un train d'impulsions périodiques pour les sons voisés.

Le train d'impulsions associé aux sons voisés correspond à la suite des ouvertures et des fermetures de la glotte. On constate que les phases d'ouverture sont beaucoup plus longues que celles de fermeture. Au moment des phases de fermeture, la réduction brutale du flux d'air

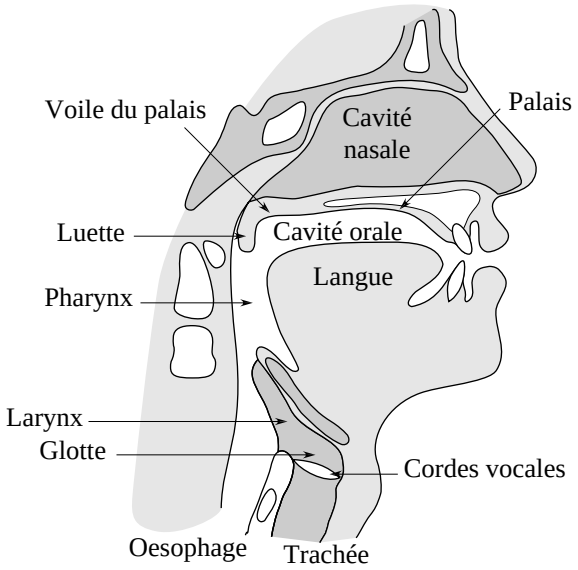

Figure 7.3: *Anatomie du conduit vocal.*

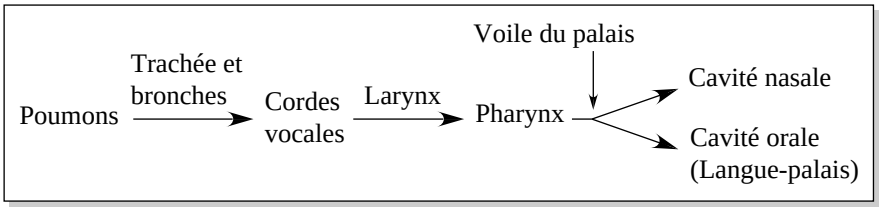

Figure 7.4: *Composantes du conduit vocal.*

provoque une impulsion brève. La fréquence fondamentale de cette suite périodique est appelée *pitch*. Cette fréquence va d'environ 70 Hz pour une voix très grave à environ 450 Hz pour une voix très aiguë. Pour un homme sa valeur est sensiblement comprise entre 70 et 200 Hz, pour une femme entre 140 et 350 Hz et pour un enfant entre 180 et 450 Hz. Dans tous les cas, pour une personne donnée, le *pitch* varie au cours de la conversation. Lors des sons voisés, le train périodique d'impulsions produit par les cordes vocales, agit comme un analyseur fréquentiel et provoque l'apparition de fréquences de résonance dans le conduit vocal. Ces fréquences sont appelées *formants*. On peut voir sur la figure 7.5, en suivant l'enveloppe du spectre représenté à droite, la présence de 4 formants.

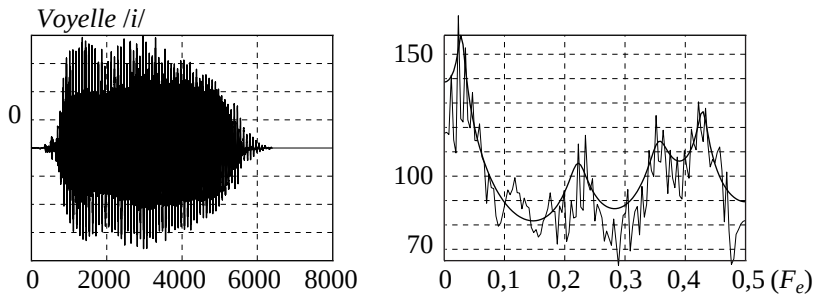

Figure 7.5: *Forme temporelle et spectre de signal de parole : on distingue sur le graphique de droite la présence de quatre formants (fréquence d'échantillonnage $F_e = 8000$ Hz.*

Notons $1/A(z)$ la fonction de transfert du filtre qui modélise le conduit vocal. Dans le cas d'un *son non voisé*, l'excitation peut être vue comme un bruit blanc et, par conséquent, le signal de parole est un processus autorégressif. On peut alors estimer les coefficients de $A(z)$, à partir d'une fenêtre de son non voisé, en utilisant les résultats obtenus pour les processus AR. Lorsqu'on applique ensuite le filtre RIF de fonction de transfert $A(z)$ au signal non voisé on obtient une estimation du bruit blanc d'excitation. Dans le cas d'un *son voisé*, on admettra que le conduit vocal peut encore être modélisé par un filtre tout-pôle $1/A(z)$. Le signal de glotte est alors un train d'impulsions périodique.

2 Etude théorique

2.1 Processus AR-P

On appelle processus autorégressif d'ordre P, en abrégé AR-P, l'unique solution stationnaire de l'équation :

$$X(n) + a_1 X(n-1) + \cdots + a_P X(n-P) = W(n) \tag{7.1}$$

où $W(n)$ est un processus aléatoire centré, stationnaire au second ordre, blanc, de variance σ_W^2, et où le polynôme :

$$A(z) = z^P + a_1 z^{P-1} + \cdots + a_P$$

est différent de zéro $\forall z$ tel que $|z| \geq 1$. Cette solution a pour expression :

$$X(n) = W(n) + h_1 W(n-1) + \cdots + h_k W(n-k) + \cdots$$

où la suite h_k est la suite des coefficients du développement en série de Fourier de la fonction $H(f) = H_z(e^{2j\pi f}) = 1/A(e^{2j\pi f})$ (périodique de période 1).

Notons que la solution stationnaire de l'équation (7.1) est causale et que son expression est la même que celle de la solution stable et causale que l'on obtient dans le cas des signaux déterministes. Si $W(n)$ est gaussien, alors $X(n)$ est lui-même gaussien, puisque le caractère gaussien se conserve par transformation linéaire.

Exercice 7.1 1. Donner l'expression de $\mathbb{E}\{W(n)X(n-k)\}$ pour $k > 0$ et $k = 0$. En déduire que, pour un processus AR-P causal, la relation entre les paramètres a_i du modèle et la fonction d'autocovariance $R(k) = \mathbb{E}\{X(n)X(n-k)\}$ est donnée par (*équations normales* ou *de Yule-Walker*) :

$$\begin{bmatrix} R(0) & R(-1) & \cdots & R(-P) \\ R(1) & R(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & R(-1) \\ R(P) & \cdots & R(1) & R(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \sigma_W^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{7.2}$$

2. En pratique, on dispose de N données $x(0) \ldots x(N-1)$ et on désire souvent calculer les coefficients du polynome $A(z)$ permettant d'interpréter les données comme une réalisation (de durée finie) du p.a. $X(n)$. Expliquez pourquoi la méthode consistant à remplacer dans (7.2) les covariances exactes par

$$\widehat{R}(k) = \frac{1}{N} \sum_{n=k}^{N-1} x(n)x(n-k)$$

est une méthode "raisonnable".

2.2 Prédiction

Considérons un processus aléatoire $X(n)$, stationnaire au second ordre, *centré*. L'expression générale d'un *prédicteur linéaire* $\widehat{X}(n)$ de $X(n)$ à l'instant n, construit à partir de ses P dernières valeurs passées, est :

$$\widehat{X}(n) = -\alpha_1 X(n-1) - \cdots - \alpha_P X(n-P) = -\sum_{i=1}^{P} \alpha_i X(n-i).$$

Exercice 7.2 1. Montrer que la minimisation de l'erreur quadratique $\mathbb{E}\left\{|X(n) - \widehat{X}(n)|^2\right\}$ entre la vraie valeur $X(n)$ et la valeur prédite $\widehat{X}(n)$ entraîne :

$$R(k) + \sum_{i=1}^{P} \alpha_i R(k-i) = 0 \text{ pour } k > 0 \tag{7.3}$$

2. Donner l'expression de l'erreur quadratique minimale.

3. En empilant la relation précédente et les P équations (7.3) sous forme matricielle, montrer que l'on obtient un système d'équations identique aux équations (7.2).

4. Conclusion ?

3 Modélisation d'un signal de parole

On décomposera l'étude en deux parties :

- On cherchera d'abord à analyser le comportement du *filtre d'analyse* $A(z)$ (ses capacités de blanchiment, ses limitations) sur du signal de parole en bande téléphonique ;
- On réalisera ensuite une approximation de l'entrée du *filtre de synthèse* $1/A(z)$ de façon à n'avoir à transmettre dans un canal de transmission qu'une information représentable sur peu de bits. On vise une compression de 64 kbit/s à 2,4 kbit/s.

  Un codeur à ce débit existe. On l'appelle le *codeur LPC10*. Il a été développé au début des années 70 pour le gouvernement américain, mais il ne présente plus aucun intérêt pratique actuellement étant donnée la qualité très médiocre du signal reconstruit. On dispose actuellement de codeurs beaucoup plus performants !

Exercice 7.3 1. Analyser le programme `encodeur.m`. Justifier le calcul des spectres.

2. Ecrire une fonction matlab donnant les coefficients du filtre $A(z)$ et la puissance de l'erreur de prédiction pour chaque fenêtre de signal.

3. En prenant une fenêtre d'analyse d'une durée de 30 ms, soit $N = 240$ échantillons, et un ordre de modélisation égal à $P = 16$, analyser le comportement du filtre blanchissant sur des signaux de parole variés (fichiers `voix_femme.wav, voix_homme.wav, voix_enfant.wav`).

Il s'agit maintenant de déterminer l'entrée du filtre de synthèse. Le *codeur LPC10* réalise une détection du voisement. Si le signal, dans la fenêtre d'analyse courante, est non-voisé, l'entrée est modélisée par un bruit blanc de variance σ^2. Si le signal est considéré comme voisé, l'entrée

est modélisée par un train d'impulsions périodique. Il faut donc réaliser dans ce cas une mesure du *pitch* et calculer l'amplitude des impulsions.

On rappelle que, lorsqu'un signal est périodique, sa fonction d'autocovariance présente des maxima séparés par la période du fondamental. La mesure du *pitch* peut ainsi se faire par recherche du maximum d'une estimation de la fonction d'autocovariance normalisée :

$$J(k) = \frac{\sum x(n)x(n-k)}{(\sum x^2(n))^{1/2}(\sum x^2(n-k))^{1/2}}. \tag{7.4}$$

On notera que cette fonction peut aussi être exploitée pour réaliser la détection de voisement.

Exercice 7.4 1. Dans le programme `encodeur.m`, imposer `voise = 0`. Analyser puis exécuter le programme `decodeur.m`. Ecouter le signal reconstruit. Conclusion ?

2. Ecrire une fonction matlab qui mesure la période du fondamental lorsque le son est considéré comme voisé. Justifier dans le programme `decodeur.m` la façon de calculer l'entrée du filtre de synthèse. Ecouter le signal reconstruit.

3. Complèter le programme pour réaliser une détection d'activité de voisement.

4. Donner un ordre de grandeur du débit nécessaire pour coder de la parole suivant ce principe. On supposera que tout scalaire pourra être représenté sur 5 bits. L'ordre $P = 16$ choisi précédemment est-il justifié ?