Chapitre 1

Généralités

I. Introduction

La parole est le principal moyen de communication dans toute société humaine. Son apparition peut être considérée comme concomitante à l'apparition des outils, l'homme ayant alors besoin de raisonner et de communiquer pour les façonner, contact physique avec la machine, libérant ainsi l'utilisateur qui peut alors effectuer d'autres tâches [1].

L'analyse numérique des signaux suppose la réalisation de certaines opérations telles que : la déconvolution, la corrélation, la transformée de Fourier, le filtrage, la décimation, l'interpolation des séquences, etc. Par exemple, la déconvolution s'avère un outil très pratique pour le calcul de la réponse d'un système numérique linéaire et invariant, défini par sa réponse impulsionnelle à une séquence quelconque appliquée à son entrée.

Ce premier chapitre est décomposé en deux partie, la première partie consiste la présentation les notions de base du signal de la parole pour comprendre le système de production de la parole et les mécanismes de production de parole et les caractéristique fondamentales de signal de parole.

Dans le deuxième partie, Nous présentons des généralités sur la déconvolution ont été introduite, ainsi son utilisation dans l'analyse de signal de parole et la déconvolution et Transformée de Fourier et homomorphique.

II. Généralité sur le signal de parole

II.1 Notions de phonème

Le phonème est la plus petite unité discrète ou distinctive que l'on puisse isoler par segmentation dans la chaine parlée. En d'autres mots c'est une entité abstraite qui peut

correspondre à plusieurs sons. Il convient de remarquer que la définition du phonème ne tient en compte que des caractéristiques qui est pertinentes pour les distinctions de signification.

Les réalisations physiques d'un phonème peuvent donc varier considérablement en fonction du contexte, de la cadence d'élocution, du dialecte, du style et du locuteur.

Les phonèmes d'une langue ne forment pas une liste amorphe, mais se regroupe en catégories naturelles dont les éléments se caractérisent par des traits distinctifs qui expriment une similarité au niveau articulatoire, acoustique ou perceptif.

II.2 Mécanisme de production la parole

La parole est le résultat de l'action volontaire et coordonnées d'un certain nombre d'organes. La production de la parole peut être assimilée à un système dynamique, dont le comportement à un moment donné dépend de ses états antérieurs. Le système est donc dépendant d'une variable paramétrable fonction du temps qui dans ce cas est un geste articulatoire [4].

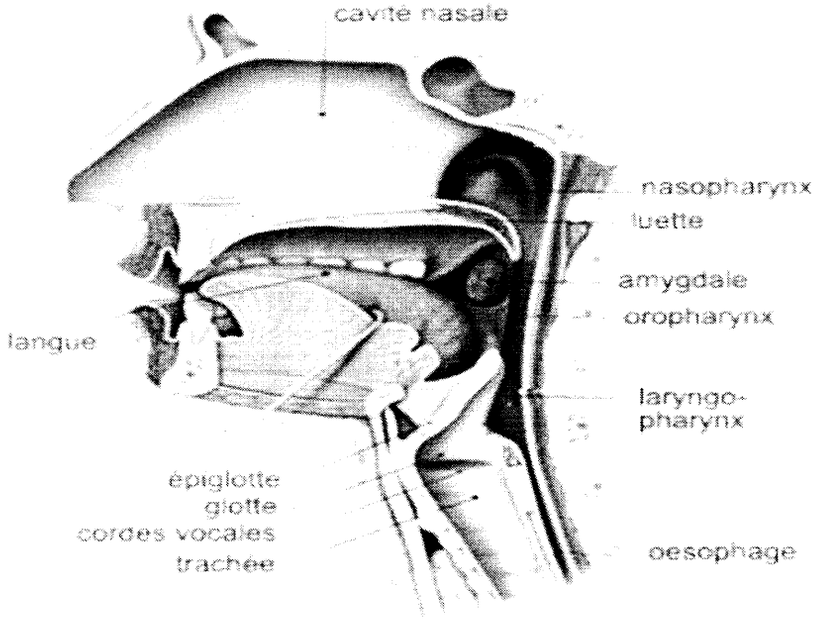

Figure 1. 1 : Anatomie du système de production de la parole.

Les poumons constituent la source du flux d'air nécessaire à la production des sons. L'air expulsé par les poumons et acheminé par la trachée passe à travers le larynx et le conduit vocal. Le signal acoustique généré suite à une modulation rapide du débit d'air est rayonné aux lèvres.

Le larynx a pour rôle de produire une excitation périodique dans le cas des sons voisés. Pour ce type de sons, la pression d'air augmente jusqu'à ce que les cordes vocales s'écartent entre-elles formant ainsi une fente appelée glotte. Le passage d'une bouffée d'air à travers la glotte met les cordes vocales en mouvement vibratoire et crée le son.

La structure anatomique du larynx est complexe. La figure 1.2 montre les différentes composantes constituant le larynx [5]. Le cartilage thyroïde est constitué de deux lames qui forment la plus grande partie de la paroi antérieure et latérale du larynx. Les aryténoïdes constituent une autre paire de cartilages auxquels les cordes vocales sont attachées dans la partie postérieure.

Les cartilages principaux du larynx sont reliés entre eux par des ligaments et muscles qui contrôlent leurs postures. Les cordes vocales sont constituées par une paire de muscles et de la membrane muqueuse qui s'étendent du cartilage thyroïde jusqu'aux aryténoïdes.

L'air expulsé par les poumons et acheminé par la trachée passe à travers le larynx et le conduit vocal. Le signal acoustique généré suite à une modulation rapide du débit d'air est rayonné aux lèvres.

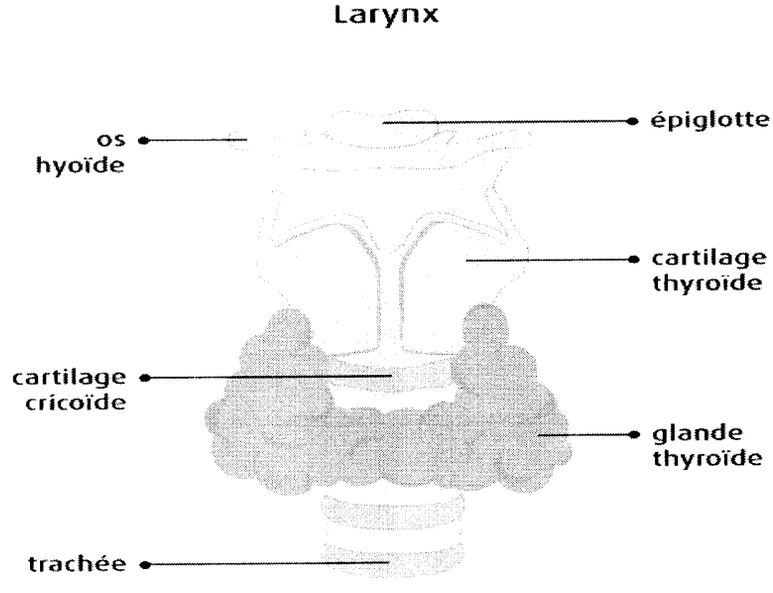

Figure 1.2 : Anatomie du larynx.

Le conduit vocal comprend le pharynx, la cavité orale et la cavité nasale. Sa longueur varie de 17 cm à 24 cm pour les hommes et de 13 cm à 17 cm pour les femmes.

Le positionnement des articulateurs (langue, dents, mâchoire, voile de palais, lèvres) permet de varier l'aire de la section droite du conduit vocal de 0 à plus de 20 cm^2.

La longueur du conduit nasal est typiquement de 12 cm. La taille de l'ouverture du voile de palais détermine le degré de couplage entre la cavité orale et la cavité nasale.

Le couplage de la cavité nasale peut influencer considérablement les caractéristiques spectrales des sons rayonnés par les lèvres. Si le voile de palais est orienté vers le bas, la cavité nasale est acoustiquement couplée à la cavité orale.

Le timbre produit par ce couplage est nasal. Pour produire un son non nasal, le voile de palais doit être positionné de sorte à fermer l'entrée de la cavité nasale en la découplant ainsi de la cavité orale. La figure 1.3 donne une représentation simplifiée des différentes parties du conduit vocale.

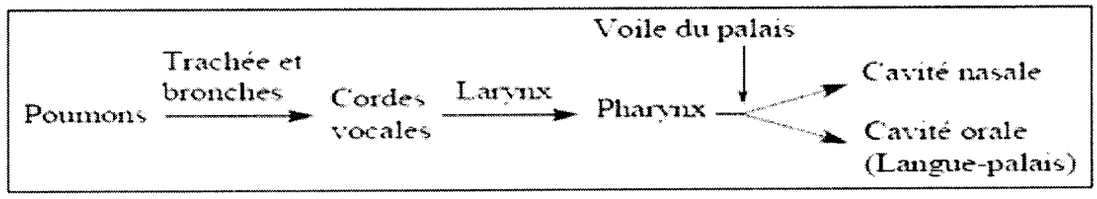

Figure 1. 3 : Différentes parties du conduit vocal.

II.3 Production de la parole par model source – filtre (parole synthétique)

Le mécanisme de production de la parole peut être considéré comme une opération de filtrage linéaire. La figure 1.4 montre le modèle discret source-filtre utilisé pour modéliser le système de production de la parole Lors de la production de la parole, la forme du conduit vocal change au cours du temps [6] .

Le filtre $H(z)$ associé au conduit vocal doit changer aussi pour simuler les effets du changement de la forme du conduit vocal. Dans les applications de la parole, les paramètres du filtre peuvent être considérés, avec une bonne approximation.

La fonction de transfert du conduit vocal est souvent modélisée au moyen d'un modèle tout pôle d'ordre p exprimé par

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad (1.1)$$

Où p_k, $k = 1, \ldots, p$, sont les pôles de la fonction de transfert.

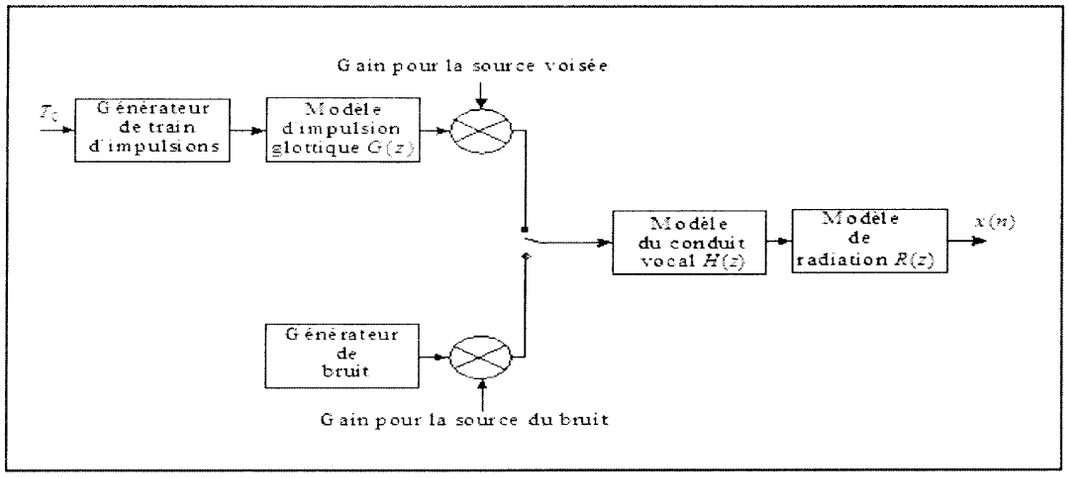

Figure 1.4: Modèle source-filtre du système de production de la parole

II.4 Classification des sons

Selon l'état des cordes vocales, les sons sont classés en deux catégories :

a- Sons voisés

Lorsque les cordes vocales sont tendues, le flux d'air les fait vibrer, c'est la phonation.

Les sons voisés contiennent l'ensemble des voyelles qui sont produites lorsque le conduit vocal est ouvert, les cordes vocales vibrent (sons voisés) et la forme des cavités (essentiellement la bouche) modifie le timbre. Les voyelles sont orale sou nasales selon que la cavité nasale n'est pas ou est mise en parallèle avec la cavité buccale plus quelques consonnes et sont issues principalement des vibrations des cordes vocales.

Le flux d'air est découpé en un train d'impulsion quasi périodique qui « résonnent » dans les différentes cavités: pharynx, bouche et optionnellement nez.

Physiquement, le train d'impulsion quasi périodique subit une modulation en fréquence fondamentale de la voix communément appelée pitch, en passant par les différentes cavités.

Différents muscles et mécanismes (mâchoire, langue, luette, lèvres, bouche) modifient la configuration des cavités pour produire les différents types de sons voisés [3] [7].

La figure 1.5 montre l'évolution temporelle du signal acoustique d'un son voisé

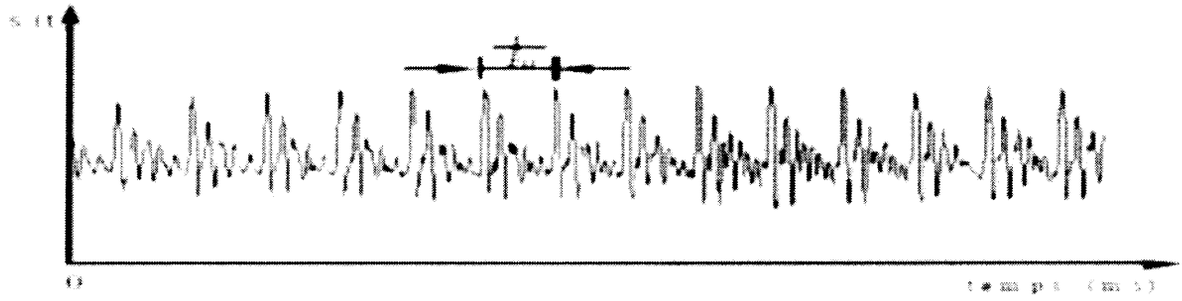

Figure 1.5 : Forme d'onde d'un son voisé [8].

Caractérisation acoustique et articulatoire des voyelles :

- les voyelles sont caractérisées acoustiquement par leurs formants (renforcement spectraux qui correspondent aux fréquences de résonance du conduit vocal).
- Les voyelles sont caractérisées articulatoire ment par le lieu de la plus forte constriction du conduit vocal.
- Parole continue (coarticulation : variabilité de la réalisation d'un phonème en fonction de ces condition de production.

b- Sons non voisés

Lorsque les cordes vocales sont relâchées, l'air passe librement au niveau du larynx. De façon similaire aux sons voisés, différents muscles et mécanismes (mâchoire, langue, luette, lèvres, bouche) modifient la configuration des cavités pour produire les différents types de sons non-voisés comme les différents consonnes qui sont produites lorsqu'un rétrécissement apparaît dans l'appareil phonatoire, les cordes vocales peuvent vibrer ou laisser passer librement l'air (sons voisés et non voisés).

Les fricatives sont produites lorsqu' un air turbulent passe à travers une constriction étroite en un point du conduit vocal. Si cette constriction est localisée au niveau de la glotte, les sons produits sont dits aspirés.

Les plosives (occlusives) sont produites par le relâchement brusque de la forte pression créée derrière une occlusion en un certain point du conduit vocal. Les nasales sont produites lorsque le voile de palais est a baissé de sorte que la cavité nasale soit couplée à la cavité orale. La

figure 1.6 montre l'évolution temporelle du signal acoustique correspondant à un son non voisé

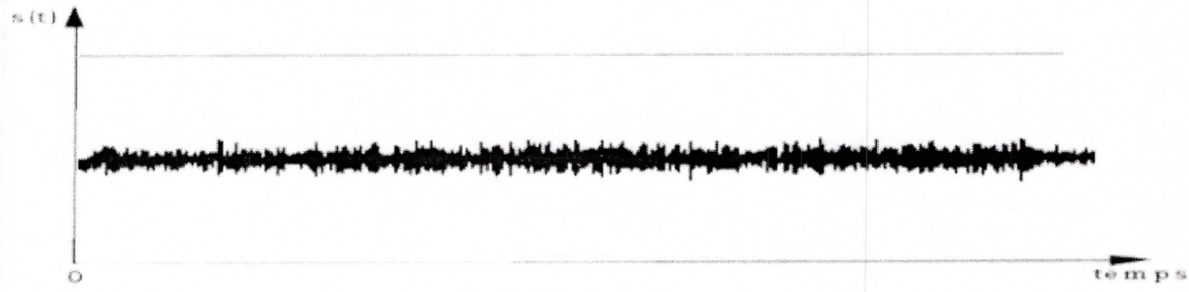

Figure 1.6 : Forme d'un son non voisé [8].

II.5 Caractéristique du signal de parole

On peut spécifier trois caractéristique essentiale :

II.5. 1 La fréquence fondamentale ou pitch

Elle représente la fréquence du cycle d'ouverture/fermeture des cordes vocales. Cette fréquence caractérise seulement les sons voisés, elle peut varier [9] :

De 80Hz à 200Hz pour une voix masculine,
De 200Hz à 300Hz pour une voix Femina,
De 300Hz à 600Hz pour une voix d'enfant.

II.5.2 les formants

Le flux d'air en prévenance des poumons est modulé par les cordes vocales, en créant des ondes de pression qui se propagent à travers le conduit vocal, ce dernier constitue des cavités orales et nasales, se comporte comme un filtre caractérisé par des fréquences de résonances appelées formants. Par conséquent, les formants représentent les maxima spectraux caractérisant la structure acoustique des voyelles et des consonnes, c.-à-d. les zones de fréquences où les harmoniques sont particulièrement intenses. Pour un conduit vocal dans la longueur est de l'ordre de 17 cm, on peut observer 3 ou 4 formants entre 100 Hz et 5000 Hz. Les trois premiers formants sont indispensables pour caractériser le spectre vocal, les deux suivants sont utiles pour une synthèse de qualité.

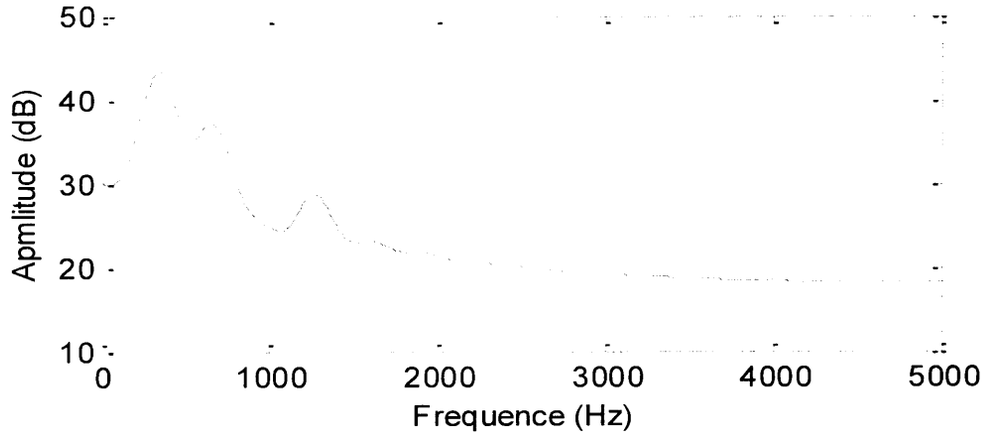

Figure 1.7: Illustration des formants

II.5.3 la stationnarité

La parole est généralement un signal non stationnaire. Il présente une évolution lente dans le temps, son spectre change de façon continue. Entre autre la parole est un signal intermittent, c.-à-d. qu'il y a des silences entre les mots. Dans une conversation typique plus de 50% du temps présente des silences, les algorithmes de rehaussement de la parole doivent prendre en considération ces caractéristiques. Le but du fenêtrage est de découper le signal de la parole en petites tranches (chacun de durée 20 mn environ) où il peut être considéré localement comme quasi-stationnaire. Le fenêtrage permet le traitement en temps réel et facilite aussi l'analyse des signaux sur la machine.

II.6 Caractéristiques de la voix

Tous les sons simples peuvent être décrits de manière exhaustive par trois paramètres : l'intensité, la hauteur et le timbre.

II.6.1 L'intensité

L'intensité d'un son, appelée aussi volume, permet de distinguer un son fort d'un son faible. Elle correspond à l'amplitude de l'onde. L'amplitude est donnée par l'écart maximal de la grandeur qui caractérise l'onde. Pour le son, onde de compression, cette grandeur est la pression. L'amplitude sera donc donnée par l'écart entre la pression la plus forte et la plus faible exercée par l'onde acoustique. Lorsque l'amplitude de l'onde est grande, l'intensité est grande et donc le son est plus fort. L'intensité du son se mesure en décibels (dB). On distingue différentes façons de mesurer l'amplitude d'un son :

II.6.2 La hauteur

La hauteur d'un son correspond à la sensation auditive aiguë ou grave liée à la fréquence :

- plus la fréquence est élevée, plus le son est aigu.
- Un son grave a une fréquence fondamentale basse.
- Un son aigua une fréquence fondamentale élevée.
- La perception se situe entre 16 et 16 000 Hz.

II.6.3 Le timbre

Le timbre est l'ensemble des caractéristiques qui permettent de différencier une voix.

Il provient en particulier de la résonance dans la poitrine, la gorge la cavité buccale et le nez, ce sont les amplitudes relatives des harmoniques du fondamental qui déterminent le timbre du son. Les éléments physiques du timbre comprennent :

- la répartition des fréquences dans le spectre sonore.
- les relations entre les parties du spectre, harmoniques ou non.
- les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont l'énergie est limitée à une ou plusieurs bandes de fréquence).
- l'évolution dynamique globale du son.
- l'évolution dynamique de chacun des éléments les uns par rapport aux autres.

III. Généralité sur la déconvolution

III.1 La déconvolution

La déconvolution consiste à retrouver le signal d'entrée x (t) connaissant le signal de sortie du système de mesure y (t) et sa réponse impulsionnelle h (t). Le but est donc de résoudre l'équation de convolution.

$$y(t) = h(t) * x(t) \tag{1.2}$$

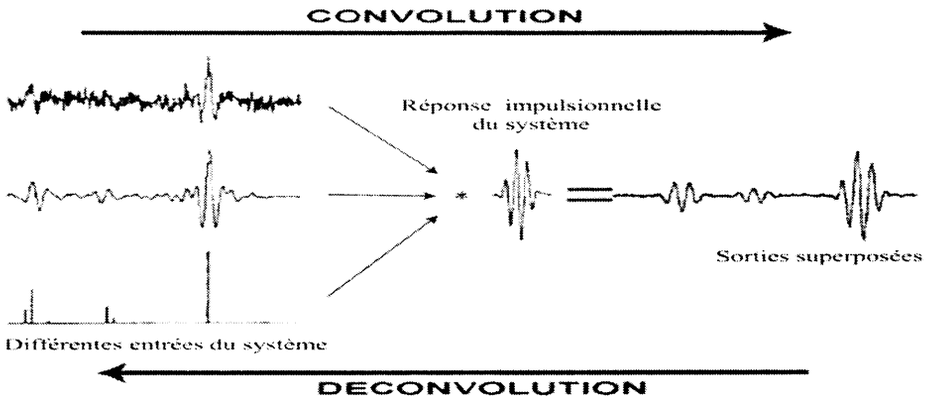

Figure 1.8: Illustration de la convolution et la déconvolution[10].

C'est donc la manière la plus simple et la plus directe de déconvoluer un signal. En effet $H(f)$ doit être différente de zéro pour toutes les valeurs de la fréquence f. D'autre part, $H(f)$ ne doit pas être à décroissance trop rapide en fonction de f.

On notera que $H(f)$ décroît en fonction de la fréquence et donc $H(f)$ tend vers zéro assez rapidement[8].

III. 1.1 La déconvolution dans le modéle source-filtre

La production des sons de parole peut être modélisée selon la théorie source-filtre. Cette théorie consiste à dire que la source des sons de la parole cavités supra-glottiques qui jouent un rôle d'amplificateur provient du souffle pulmonaire qui fait entrer en vibrations les cordes vocales. Une trame du signal de parole peut être modélisée comme étant la convolution de la source glottique avec la réponse impulsionnelle du conduit vocal, comme illustrée par la figure 1.10.

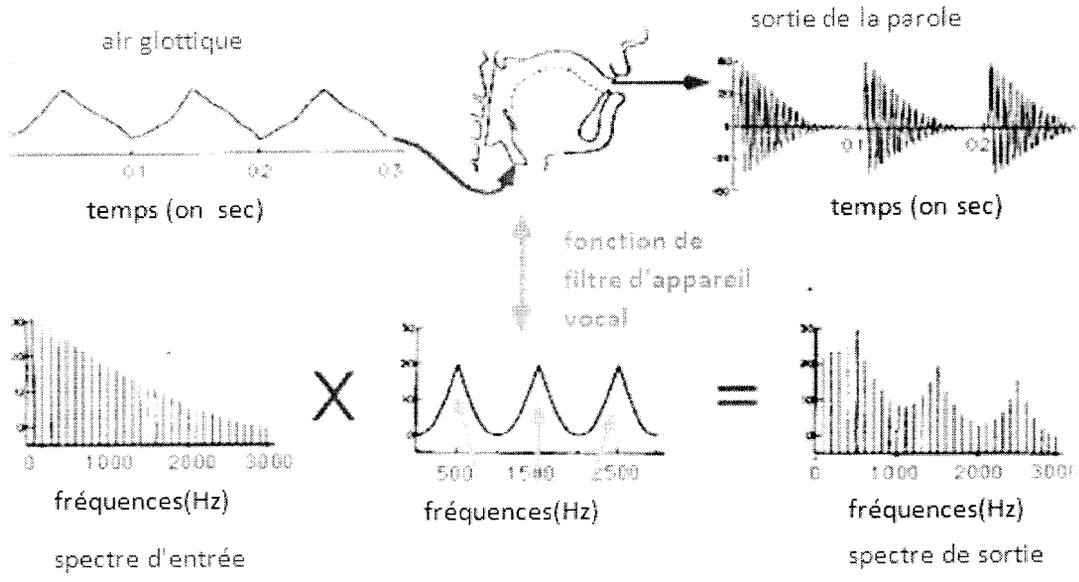

Figure 1.9 : production de la parole par le modèle source –filtre.

d'après l'équation (1.2) . Où y(t) est le signal de parole, x(t) est la réponse impulsionnelle du système modélisant le conduit vocal et h(t) est le signal d'excitation ayant pour origine les cordes vocales, et ∗désigne le produit de convolution. La transformée de Fourier du signal transforme la convolution en produit décrit par l'équation suivante

$$Y_w(f)=H_w(f).\,X_w(f) \tag{1.3}$$

Ou Les transformations de Fourier directe définie par l'équation suivante:

$$Y(f) = TF\{y(\,t)\} = \int_{-\infty}^{+\infty} y(t)e^{-2.\pi.f.t}df \tag{1.4}$$

Où f dénote la fréquence,$Y_w(f)$ et$H_w(f)$sont respectivement les spectres d'amplitude des trames d'analyse du signal et de l'excitation glottique pondérées par la fenêtre $w(t)$ et$X_w(f)$est la réponse fréquentielle du conduit vocal [10].

III.1.2 Analyse homomorphique

Cette méthode a pour but de séparer les contributions du conduit vocal et des sources. D'après le modèle source-filtre, la parole est le résultat d'une convolution de la source par le filtre constitué du conduit vocal et de la radiation aux lèvres : Le but est de déconvoluer ce produit de convolution de manière à séparer les contributions de la source et du filtre (conduit),

Selon le modèle source-filtre de production de la parole, le signal de parole peut être considéré comme le résultat de la convolution de l'excitation de l'appareil vocal (excitation glottique) et de sa réponse impulsionnelle.

$$s(t) = e(t) * v(t) \quad (1.5)$$

Où s(t) est le signal de parole, v(t) est la réponse impulsionnelle du système modélisant le conduit vocal, et e(t) le signal d'excitation ayant pour origine les cordes vocales, et * désigne le produit de convolution. La transformée de Fourier du signal transforme la Convolution en produit décrit par l'équation suivante :

$$S_{w(f)} = E_w(f).V(f) \quad (1.6)$$

La Convolution de deux fonctions dans le domaine temporel se traduit par un produit de ces deux fonctions dans le domaine des fréquences et not par (.).

Où $\boldsymbol{f}$est la fréquence,$\boldsymbol{S_w(f)}$ et$\boldsymbol{E_w(f)}$son respectivement, les spectres d'amplitude des trames d'analyse du signal et de l'excitation glottique pondérées par une fenêtre et $V(f)$est la réponse fréquentielle du conduit vocal.

Le spectre d'amplitude d'une trame pondérée du signal de parole peut être écrit comme la suit :

$$|S_w(f)| = |E_w(f) \times V(f)| \quad (1.7)$$

On applique le logarithme sur l'équation (1.7), on obtint :

$$\log|S_w(f)| = \log|E_w(f)| + \log|V(f)|$$

La propriété de la transformée de Fourier : la transformée de Fourier d'un produit de convolution est égale au produit des transformées de Fourier des facteurs. Puis on utilise la propriété du logarithme : le logarithme d'un produit est égal à la somme des logarithmes facteurs. On retourne dans un domaine pseudo-temporel par une transformée de Fourier inverse, la transformée de Fourier inverse d'une somme étant égale à la somme des transformées de Fourier inverse des termes, il résulte alors de ces transformations une contribution additive de la source et du conduit [11].

Par transformation de Fourier inverse de (1.2), on obtient le cepstre.l'expression du cepstre réel est donc :

$$c = TF^{-1}(\log(TFD(y(t)) \quad (1.8)$$

Ou $TF^{-1}(.)$dénote-la transformée de Fourier inverse.

L'espace fréquentiel de représentation du cepstre est équivalent à un espace temporel. Donc pour ça, on définit que la harmonique est l'inverse de l'harmonique et quéfrence est

l'inverse de la distance entre rais successives dans la transformée de Fourier du signal temporel étudié.

A partir du cepstre, il est possible de définir la fréquence fondamentale de la source e en détectant les pics périodique (harmonique) au-delà d'un certain nombre N de coefficients. En effet, les N premiers points du cepstre contiennent l'information l'harmonique due à la contribution de source.

Cependant, déterminer la fréquence fondamentale d'un signal de parole reste encore un problème difficile .en effet, les algorithmes classique manquent de robustesse quand le bruit est présent, quand la fréquence fondamentale change rapidement ou quand la valeur de celle-ci n'est pas assez élevée.

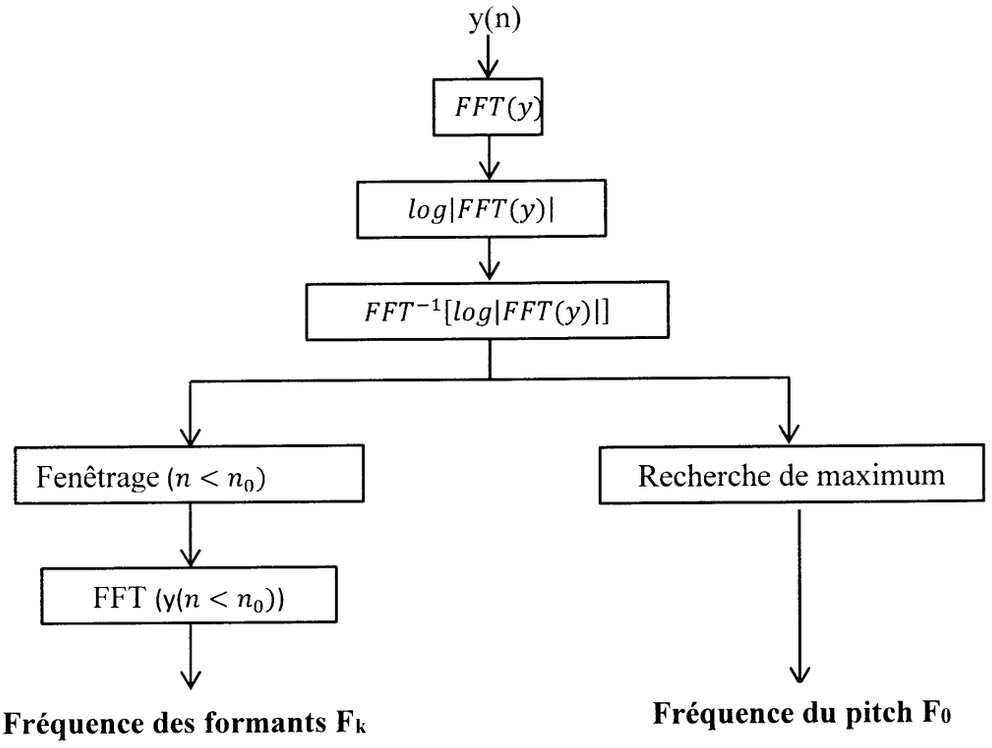

Figure 1.10:Organigramme de décomposition par le cepstre

La transformée homomorphique est une grandeur très utilisée dans le traitement de la parole pour des applications de filtrage, de détermination de hauteur totale, ou de reconnaissance vocal. Elle permet dans une certaine mesure, de séparation l'influence de canal de transmission. Le cepstre de voix donne l'image de ses variations, c'est d'ailleurs par ce biais que la déréverbération est envisagée. Le signal vocal varie plus lentement que ses multiples réflexions. En pratique ces méthodes sont bien plus difficiles à mettre en œuvre lorsque l'on considéré le

signal audio dans sa continuité.il faut alors fragmenter le signal en vue d'une décomposition en série de Fourier, ce qui génère des artefacts (paradoxe temp-fréquence et effet de fenêtrage).

VI. Conclusion

Dans ce chapitre, nous avons présenté les différentes représentations du signal de parole ainsi que le modèle de l'appareil phonatoire, ce qui permet une bonne compréhension du mécanisme de production de la parole. De même, nous avons présenté le modèle discret source-filtre qui permet de modéliser le mécanisme de production de la parole.

On a présenté aussi sur la méthode de déconvolution et son rôle important d'application et la déconvolution dons le modéle source-filtre, d'autre part on parle sur l'analyse homomorphique et le problème inverse de la déconvolution.