CHAPITRE 1

LE REHAUSEMENT DE LA PAROLE

Le rehaussement de la parole vise essentiellement l'intégrité et la préservation optimales de l'identité et des spécificités de la source sonore recherchée malgré ses altérations subies et provoquées par des éléments sonores perturbateurs qui lui sont étrangers mais présents dans son environnement.

Les applications où l'on retrouve la nécessité du rehaussement de la parole sont multiples et variées. À titre d'exemple, mentionnons les domaines des communications où la parole en général est mise en jeu dans sa reconnaissance ou son identification notamment dans les applications du multimédia, des radios, des communications commerciales, industrielles et militaires.

On peut également déduire que les solutions proposées pour le rehaussement de la parole peuvent s'exporter vers des domaines similaires mettant en jeu des signaux de nature sonore, comme dans le domaine médical, plus spécifiquement, l'échographie entraînant une plage fréquentielle du sonore plus étendue embrassant les ultrasons comme les infrasons.

Avant de présenter succinctement la nature de la parole, on doit définir la classification des signaux sonores perturbateurs. Ces sont généralement les signaux exclus du ou des signaux recherchés. Ils portent alors l'appellation de bruit qui se classifie selon l'effet perturbateur qu'il produit sur les ou le signal recherché, en l'occurrence ici, la parole.

L'environnement entourant la parole d'un locuteur peut générer des bruits non corrélés et additifs, ils ont donc un effet de superposition sur le signal de la parole comme c'est souvent le cas du bruit ambiant.

Le bruit convolutif est celui qui affecte le spectre du signal de la parole lorsqu'elle subit un filtrage linéaire.

Les causes sont principalement dues à la réverbération: son effet peut être modélisé par un filtre linéaire qui dépend de l'architecture, du contenu de la salle dont les surfaces réfléchissantes retransmettent le signal altéré dans ses composantes fréquentielles, et de la position du locuteur par rapport au microphone car plusieurs chemins se créent introduisant sur le signal recherché différents délais ou retards de ses copies introduisant ainsi des interférences.

Ensuite viennent les caractéristiques spectrales spécifiques d'un locuteur : la moyenne à long terme de la réponse fréquentielle du conduit vocal varie d'un locuteur à un autre à cause des différences physiologiques des conduits vocaux.

Finalement s'ajoute l'équipement d'enregistrement: si on utilise différents microphones, on peut constater des modifications tout au long du spectre. Ce changement dû au microphone peut être raisonnablement modélisé par un filtre linéaire.

Le bruit convolutif est une cause majeure de la dégradation des performances. Les mesures de distorsions spectrales utilisées pour la reconnaissance de la parole sont fortement affectées par ce type de bruit.

Si le modèle du canal de transmission peut être représenté par un filtre linéaire alors les signaux enregistrés peuvent être raisonnablement modélisés et l'implication des bruits additifs et convolutifs sera fondée sur ce modèle. Ainsi la représentation d'un signal observé y(t) affecté de ces bruits donne la relation suivante :

$$y(t) = s(t) * h(t) + n(t) \tag{1.1}$$

où s(t) est le signal de la parole d'origine sans distorsion, n(t) est le bruit additif et h(t) le bruit convolutif.

1.1 Les caractéristiques de la parole

La parole peut être considérée comme un signal sonore engendré par le système vocal humain faisant participer un ensemble d'organes constituant le conduit vocal. Il s'agit des muscles pulmonaires, des poumons, des bronches avec la trachée qui aboutissent aux cordes vocales, éléments vibratoires essentiels pour la génération des sons, ensuite le conduit se termine par un ensemble de cavités à volume variable dont certaines se caractérisent par une grande souplesse ou par une rigidité importante. Elles sont constituées du larynx, du pharynx, par la suite, le conduit se sépare au voile du palais pour aboutir sur les deux dernières cavités nasale et buccale. Cette dernière d'une grande souplesse est sans contredit la plus versatile. Elle joue aussi un rôle prépondérant dans la formation des sons par une mise en œuvre complexe de la coordination simultanée des muscles linguaux, maxillaires et buccaux. La figure 1 illustre les organes de la phonation.

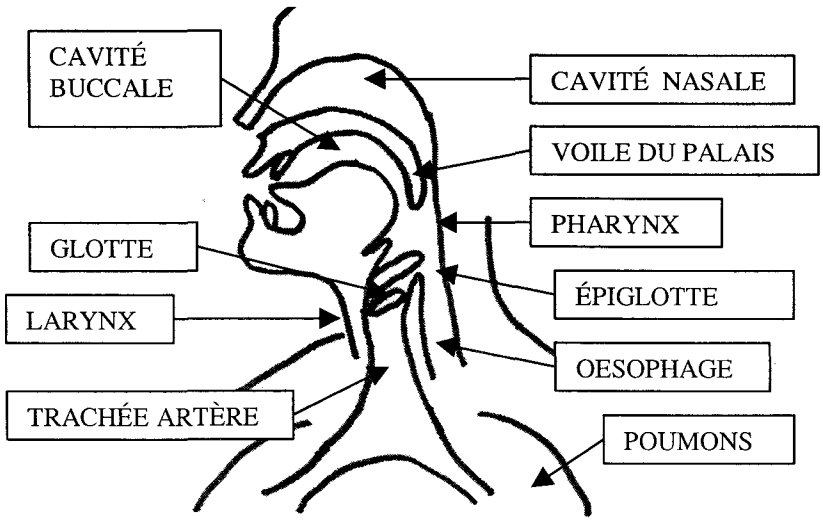

Figure 1 Les organes de la phonation selon [1]

Les sons représentant la parole sont considérés selon le point de vue du traitement du signal comme étant globalement un processus aléatoire. Mais si on fait une analyse segmentée sur le signal de la parole, on retrouve généralement un aspect pseudo-stationnaire ou quasi-périodique qui est caractéristique lorsqu'elle n'est pas parsemée d'éléments purement aléatoires.

Ce sont ces dernières considérations qui ont amené à identifier deux classes de sons issus de la parole. Il s'agit des sons voisés et non voisés.

La première classe à forte périodicité, englobe l'ensemble des voyelles plus quelques consonnes et sont issues principalement des vibrations des cordes vocales, donnant une fréquence fondamentale de la voix communément appelée pitch, plus les harmoniques de celle-ci engendrées par les cavités en aval des cordes vocales par effet de résonance, ces dernières composantes périodiques sont appelées les formants de la voix.

La seconde classe est remarquable par son absence totale de périodicité et s'apparente aisément à la nature aléatoire du bruit blanc, et se situe sur la partie haute du spectre vocal.

La voix présente une portée fréquentielle pouvant aller autour de 6 octaves en étendue. Cependant parmi les signaux vocaux, le ton de la voix est situé dans une bande restreinte du spectre vocal humain où ses limites sont fixées sur différentes valeurs selon les chercheurs rencontrés. À titre d'exemple, selon [2] le ton de la voix ou sa fréquence fondamentale f0 se situe dans les fréquences basses du spectre de la voix parlée, c'est-à-dire généralement entre 100 et 150 Hz pour une voix masculine, entre 200 et 300 Hz pour une voix féminine, et pour une voix enfantine, elle se situe entre 300 et 450 Hz.

Après avoir décrit brièvement la nature des signaux de la parole, nous allons passer en revue les techniques usuelles employées pour rehausser celle-ci en présence du bruit.

1.2 Les techniques de rehaussement de signal

En général, la majorité des méthodes de rehaussement traite principalement le bruit additif et plus rarement celui de convolution dont l'élimination ou la réduction s'avère être une tâche plus élaborée. Mais avant d'aborder le bruit convolutif, nous allons résumer les méthodes qui éliminent le bruit additif puisque celui-ci est toujours présent en pratique. La référence [3] présente les principales méthodes de rehaussement de la parole que nous allons résumer ici dans leurs grandes lignes.

La première méthode opère dans le domaine spectral. Le signal temporel observé est projeté dans le domaine fréquentiel par transformée de Fourier (en général avec la FFT) pour déterminer la puissance à laquelle on soustrait la densité spectrale de puissance du bruit mesurée ou connue lors des moments où la parole est absente. Le résultat est remis dans le domaine temporel par la transformée de Fourier inverse qui nous donne le signal rehaussé et nettoyé du bruit additif.

La seconde méthode opère tant dans le domaine temporel que fréquentiel, et emploie intensivement le filtrage optimal adaptatif qui ajuste les coefficients de sa réponse impulsionnelle de telle sorte que l'on obtienne en sortie le signal rehaussé à partir du signal observé en entrée du filtre. Pour se faire, le procédé nécessite la connaissance statistique de l'estimation du signal observé comme celle du bruit et s'appuie généralement sur le critère de l'erreur moyenne quadratique minimale de l'estimation.

La troisième méthode est l'annulation adaptative du bruit. Elle est également basée principalement sur le filtrage adaptatif qui nécessite souvent à son entrée un signal de référence qui dans notre cas est le bruit environnant non corrélé au signal recherché, nous obtenons alors à sa sortie une estimation du bruit qui est soustraite ou comparée au signal observé prélevé sur le deuxième canal. Le résultat donne le signal rehaussé. Cette méthode se base également souvent sur le critère de l'erreur moyenne quadratique

minimale de l'estimation et emploie un grand nombre d'algorithmes qui y sont associés. Elle s'opère le plus souvent dans le domaine temporel.

La quatrième et dernière des principales méthodes s'appuie sur les caractéristiques fréquentielles de la voix et à sa nature périodique pour éliminer sinon réduire le bruit perturbateur. Cette méthode contrairement aux autres mentionnées précédemment, s'affranchit de la nécessité d'une connaissance à priori de l'environnement bruité. Elle dépend essentiellement de la détermination de la fréquence fondamentale ou pitch de la voix étudiée. Dès que cette information est acquise, on procède différemment selon le choix du domaine emprunté, temporel ou fréquentiel.

Dans le domaine temporel, on introduit, après un délai égal à la période fondamentale, le signal observé dans un filtre adaptatif qui ajuste ses coefficients pour obtenir le signal estimé qui est comparé à une période près, le résultat est le bruit rejeté et constitue l'erreur estimée qui vient modifier les coefficients selon le critère de l'erreur moyenne quadratique minimale de l'estimation.

Dans le domaine fréquentiel, on emploie un filtre adaptatif en peigne qui va rechercher la fréquence fondamentale et ses harmoniques tout en éliminant le restant qui constitue essentiellement la partie bruitée.

Dans le travail présenté ici, nous utilisons certains éléments issus des méthodes qui viennent d'être mentionnées.

Nous avons retenu intentionnellement le bruit convolutif pour la fin, car les méthodes élaborées pour le contrer n'existent que depuis une décennie pour les signaux à large bande et depuis environ deux décennies pour les signaux à bande étroite.

Les références [4] et [5] traitent du sujet en abordant les mélanges convolutifs. Les méthodes décrites sont complexes et exigent souvent un calcul intensif et laborieux souvent associés aux statistiques d'ordre élevé.

Ce que nous retenons ici, c'est qu'il est admis [5] qu'un mélange convolutif peut s'approcher d'un mélange instantané lorsque les signaux considérés sont à bande étroite. C'est cette dernière caractéristique qui est mise à profit dans la solution retenue que nous verrons plus loin en détail.

1.3 La détermination de la fréquence fondamentale

La fréquence fondamentale f0 est une caractéristique essentielle dans notre application choisie, aussi les approches les plus classiques de la détermination de la fréquence fondamentale sont maintenant rappelées avant de présenter la méthode employant la transformée en ondelettes qui sera d'ailleurs expliquée plus en détail au chapitre 3.

Ces approches sont définies en deux classes; la première classe opère dans le domaine temporel, dans ce cas on cherche à estimer l'intervalle entre deux instants consécutifs ou survient la fermeture glottique de la parole. La seconde classe exploite la stationnarité segmentée du signal de la parole pour relever sa périodicité et emploie alors l'analyse spectrale fenêtrée, la recherche de la fréquence fondamentale dans ce cas, s'opère dans le domaine fréquentiel.

Mentionnons les quelques principales méthodes utilisées dont fait mention la référence [6]:

- Méthode basée sur l'analyse spectrale
- Méthode basée sur l'autocorrélation
- Méthode du Cepstrale
- Méthode du LPC (linear predictive coding)

- Méthode du SIFT (simplified inverse filtering technique)
- Méthode du AMDF (average magnitude difference function)

Parmi cette dernière liste nous montrerons quelques exemples à savoir les méthodes de la spectrale , de la corrélation ainsi que de la cepstrale et de l'AMDF.

La détermination de la fréquence fondamentale f0 peut se faire au moyen d'une analyse spectrale segmentée par transformée de Fourier sur le signal de la parole à analyser. La partie spectrale qui nous intéresse se situe en bas de 600 Hz. La f0 est la plus petite fréquence avec un maximum local parmi le contenu de ses harmoniques ayant des maxima locaux.

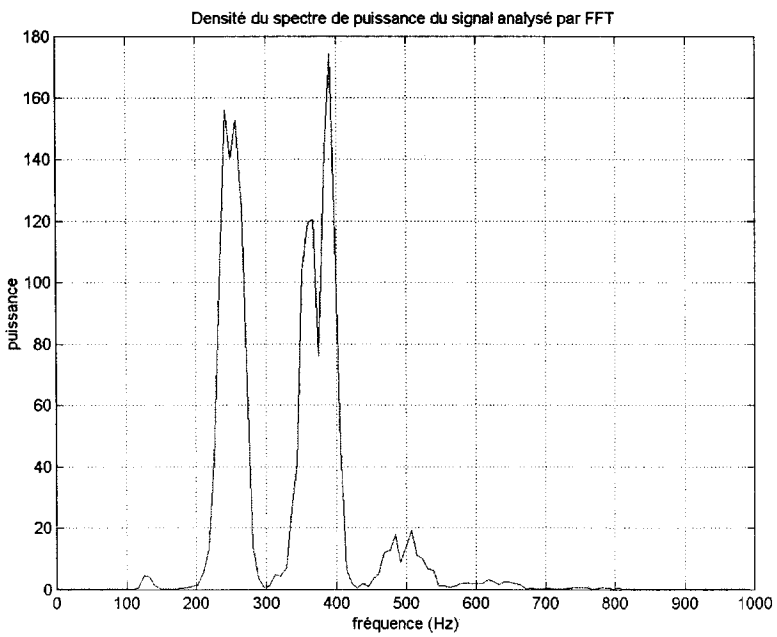

Figure 2 Détermination de la f0 par FFT

La figure 2 illustre un exemple avec le mot anglophone « we » où on peut voir le premier maximum (la f0) aux environs de 130 Hz avec d'autres maxima plus importants

de ses harmoniques qui sont définis comme les formants de la voix et se situent aux environs de 260 Hz (2f0), 390 Hz (3f0) et 520 Hz (4f0).

Il est important de remarquer l'importance de la résolution fréquentielle. Dans notre exemple elle est fixée à (8000/1024) Hz, soit le rapport de la fréquence d'échantillonnage sur le nombre d'échantillons traités par la fenêtre ou le segment analysé. Si cette résolution est trop grande, la f0 peut ne pas apparaître dans certains cas comme ici dont le maximum est relativement petit et on peut alors la confondre avec une de ses harmoniques aux maxima plus importants.

La seconde méthode repose sur le principe de la corrélation du signal à analyser avec une version de lui-même qui est plus ou moins retardée, il s'agit plus précisément de l'autocorrélation. Elle possède une propriété qui est ici mise à profit ; tout signal à caractère périodique va donner une autocorrélation également périodique. Le temps séparant les maxima consécutifs de l'autocorrélation donne la période du signal analysé.

La figure 3 montre un exemple d'autocorrélation avec le même signal analysé précédemment et illustré sur la figure 3a. La figure 3b donne le résultat de l'autocorrélation, la figure 3c est sa copie avec les premiers échantillons retirés qui ne tient pas compte du premier maximum autour de zéro de l'abscisse. La figure 3d donne le calcul de la f0 sur les deux premiers maxima consécutifs situés autour du $60^{ième}$ et $125^{ième}$ échantillons donnant respectivement comme f0, 130 Hz et 125 Hz.

On remarque un troisième maximum donnant une fréquence supérieure à 1 KHz en dehors du domaine spectrale de la fréquence fondamentale humaine. Plus on va dans la partie rétardée du résultat de l'autocorrélation de la figure 3c, plus les maxima diminuent et se confondent aux maxima des contributions bruités du signal. Il est donc important de fixer judicieusement un seuil au dessus duquel on considère les maxima retenu pour fin du calcul de la f0. En général le premier maximum fixe la valeur de la f0.

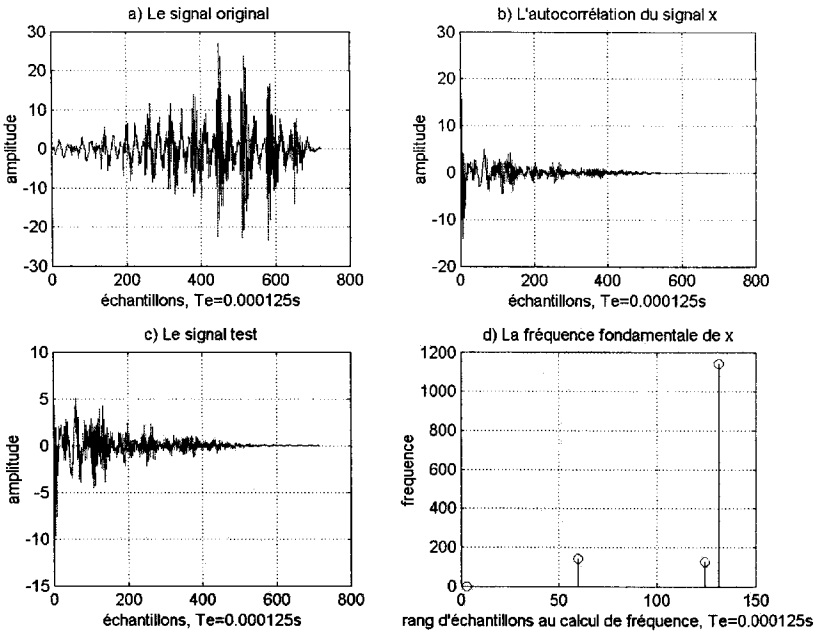

Figure 3 Détermination de la f0 par autocorrélation

La méthode de l'AMDF qui est le sigle anglais pour Average Magnitude difference function est une méthode dérivée de l'autocorrélation, elle est aussi définie comme l'autocorrélation par différence. Contrairement à l'autocorrélation, on va chercher les positions et les écarts entre les minima. Comme l'autocorrélation, c'est principalement le premier minimum qui fixe la période de la f0. La figure 4 donne les résultats de cette méthode.

La figure 4a représente le signal analysé. La figure 4b donne en graphique le résultat de l'AMDF le premier minimum après l'origine en abscisse se trouve autour du 65^{ième} échantillon, son calcul est montré sur la figure 4c où la f0 est déduite à 123 Hz. Ensuite les autres minima donnent respectivement des f0 de 133 Hz, 100 Hz, et 195 Hz.

Le dernier minimum sur la figure 4c est sujet aux mêmes contraintes vues pour l'autocorrélation, dès qu'on s'éloigne de l'origine de l'abscisse les minima ou les

maxima pour l'autocorrélation sont difficiles à bien déceler ou isoler des contributions bruités du signal.

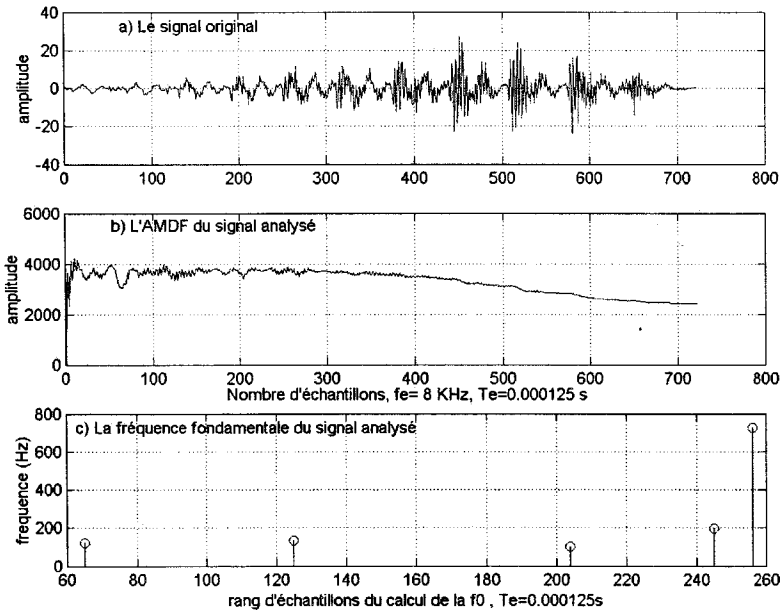

Figure 4 Détermination de la f0 par l'AMDF

La méthode cepstrale selon [7] est dérivée de la méthode d'analyse spectrale par transformée de Fourier segmentée. Il s'agit de faire la transformée de Fourier sur un signal segmenté. On calcule ensuite le logarithme de ce dernier résultat puis on lui applique la transformée de Fourier inverse pour revenir dans le domaine temporel. La figure 5 illustre l'analyse du signal par la méthode cepstrale pour déterminer la f0.

La figure 5a représente le signal analysé dans le domaine temporel. La figure 5b est la transformée de Fourier du signal dans le domaine fréquentiel. La figure 5c est l'application du logarithme sur la figure 5b dans le domaine fréquentiel. Finalement la

figure 5d est le résultat de la transformée inverse de Fourier sur le graphique de la figure 5c. Cette dernière opération nous a ramené dans le domaine temporel.

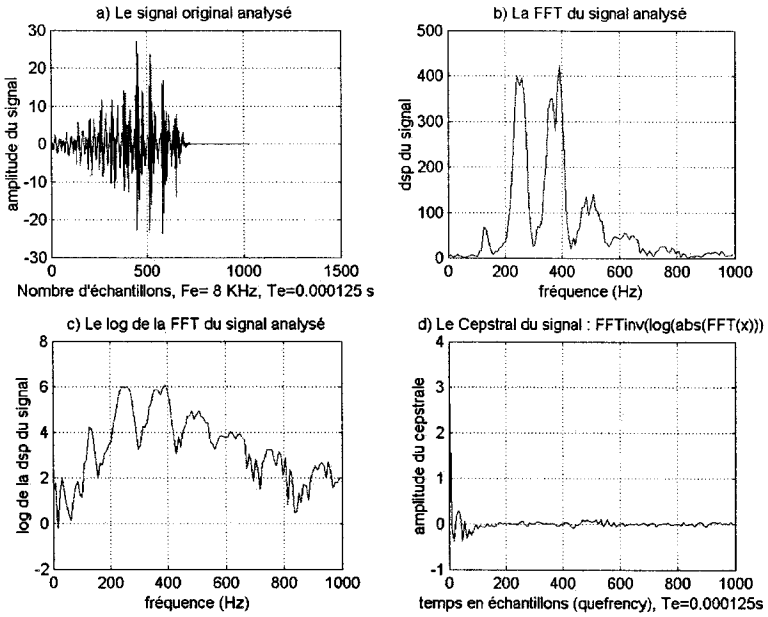

Figure 5 Détermination de la f0 par cepstrale

L'abscisse (en ms) appelée quefrence de la figure 5d est divisée en deux sous-domaine temporel où la frontière se situe autour des 2 ms. Cette frontière se traduit en terme d'échantillon autour du $16^{ième}$. La zone inférieure à cette limite constitue pour la voix la contribution du conduit vocal autrement dit les formants ou la partie fréquentielle au delà de 500 Hz (1/2 ms). La zone supérieure à cette limite constitue la contribution de l'onde glottique, autrement dit les fréquences en bas de 500 Hz, où la fréquence fondamentale de f0 se retrouve.

Le premier maximum que l'on observe dans la zone supérieure nous donne la f0 étant l'inverse de la valeur temporelle où se trouve ce maximum. Dans notre cas nous

devrions observer un maximum autour du 65^{ième} échantillon. On constate dans le cas présent que cette méthode a de la difficulté à détecter la f0.

Nous verrons plus loin, au chapitre 3 la transformée en ondelettes utilisée pour la détection de la f0 de notre application qui traitera également le signal étudié ici. De plus d'autres comparaisons avec les méthodes classiques seront étudiées et présentées par d'autres auteurs.